Capacity of binary perceptrons with binary inputs

El Mahdi CHAYTI
EPFL, IC

Supervised by
Emmanuel Abbé (MDS LAB) & Lenka Zdeborová (SPOC LAB)

February 16, 2021

Abstract

We extend the algorithm in (Kim & Roche, 1998) that deals with the problem of linearly separating binary patterns using a binary perceptron. Although the problem seems fairly simple at first sight, it is still open from a mathematical point of view with conjectures made more than half a century still resisting the test of time. Statistical physics with all its interesting approximations did solve the problem up to the rigor of the approximations that were used (Like the replica trick), but there is still a lack of algorithms that perform as well as what is conjectured. In this work, the algorithm that works for a "stability" parameter equal to zero is extended to work for a "stability" that can take any positive value. We also show the limitations of the class of local rules that contains the majority rule and discuss how "enhancing" those local rules in a procedure similar to what was done in (Kim & Roche, 1998) will lead to sub-optimal rule.

Contents

1 Introduction and state of the art

We are interested in the following problem : given k binary patterns $\{X_1, \dots, X_k\}$ drawn uniformly at random from $\{\pm 1\}^n$, is it possible to find a vector of binary weights $w \in \{\pm 1\}^n$ (a binary perceptron) such that

$$\forall i \in \{1, \cdots, k\} : \langle w, X_i \rangle \geq \kappa\sqrt{n} \tag{1}$$

Where $\langle , \rangle$ is the euclidean scalar product in $\mathbb{R}^n$ and κ is a stability parameter.
(1) can be written in a more compact form : $Xw \geq \kappa\sqrt{n}$, where $X \in \{\pm 1\}^{k \times n}$ is the matrix containing the k binary patterns $\{X_i\}$ in its rows and $\geq$ is a point-wise comparison.
Precisely, we will be interested in the probability of finding such w which will be denoted as $P_\kappa(n, k)$.
This is a special case of a more general problem where the patterns are sampled from a subset S_X of some n-dimensional euclidean space $\mathbb{E}^n$ and the weights are looked for in some other subset S_w of $\mathbb{E}^n$.
These problems were motivated by covering problems ((Cover, 1965), (Kim & Roche, 1998)) in geometry at first and then by the problem of memory storage in dynamical neural networks ((Gardner, 1988),(Gardner & Derrida, 1988),(Krauth & Mézard, 1989a)). Also, the problem is a special case of random constraints satisfaction problems (CSP, (Coja-Oghlan, 2009)), in addition to being a machine learning binary classification.
While almost all other instances of this problem (In general when at least one of the two sets S_X and S_w is not discreet) were fairly solved, the binary case, to which our interest is turned, has resisted the test of time with conjectures that are still unsettled to the time of this writing and the best of our knowledge.

1.1 Covering sets

Let $n \in \mathbb{N}$ (a dimension) and S be a set in $\mathbb{E}^n$ (the workspace, here, a Euclidean space of dimension equal to n). We say that S is covered by $H_1, \dots, H_k$ (k subsets of $\mathbb{E}^n$) if :

$$S \subseteq \cup_{i=1}^{k} H_i$$

In the case where each H_i is determined by one and only one vector $X_i \in \mathbb{E}^n$, we will say, abusively, $X_1, \dots, X_k$ cover S to mean, that the subsets generated by these vectors cover S.
In particular, when $H_{X_i} = \{w \in S_w \subseteq \mathbb{E}^n : \langle w, X_i \rangle < \kappa\sqrt{n}\}$ then not covering S by $H_{X_1}, \dots, H_{X_k}$ is equivalent to (1).

1.2 Associative memory

Given a dynamical system consisting of $n+1$ neurons $\{\sigma_i\}_{i=0}^n$ interacting between each other through the "synapses" J_{ij} in the following manner :

$$\sigma_i^{t+1} = sign(\sum_j J_{ij}\sigma_j^t) \tag{2}$$

A fixed state of this dynamic is a configuration of n+1 neurons such that

$$\sigma_i = sign(\sum_j J_{ij}\sigma_j)$$

In other terms, we ask that : $\sum_j J_{ij}\sigma_j\sigma_i \geq \kappa\sqrt{n}$
Where the term $\kappa\sqrt{n}$ is used instead of zero to guarantee more stability. Hence why κ is sometimes called stability (it has been shown that the bigger κ is the more stable the fixed points will be).
The fixed states of the dynamic (2) can be used as an associative memory to store k patterns, the question here is whether there are interactions that can store a given set of k patterns or not.
In the special case of systems with no self-interaction ($J_{ii} = 0$), when we isolate one neuron (say 0), write the stability for the chosen neuron can be written in the following form:

$$\sum_j w_j X_{lj} \geq \kappa\sqrt{n}$$

Where $w_j = J_{0j}$, $X_j = \sigma_j^{(l)}\sigma_0^{(l)}$ and $\{\sigma_j^{(l)}\}_{j=0}^n$ is the l-th pattern out of the k patterns we want to store.
We find again the original problem.

1.3 Binary classification

Finding $w \in \{\pm 1\}$ that verifies (1) is equivalent to learning how to separate binary patterns on the following training set $\{(X_i, +1)\}_{i=1}^k$ by using a perceptron with binary weights, bias equal to $-\kappa\sqrt{n}$ and non-linearity or activation function equal to the $sign$ function.

1.4 Random constraint satisfaction

The k inequalities in (1) are constraints imposed on the binary weights w. The question in random constraint satisfaction problems is if there is a valuation of w that makes all the constraints k to be $TRUE$.

1.5 The capacity

In random constraint satisfaction, we have in general k constraints imposed on n variables that need to be verified. Loosely speaking, for k small, we expect that the problem will always have a solution as it is under-constrained. When k gets bigger, finding a solution becomes harder and harder up to a point where there will be no solution unless a very fortuitous rare event happens.
This loose reasoning and simulations suggest that for this kind of problems there is a phase transition i.e. there is a $k_s(n)$ that separates the region of k's for which a solution will be guaranteed with high probability and the region of k's with no solutions.
In general, the threshold can be either sharp or coarse, sharp means that the threshold occurs in one specific value of k, coarse means that the threshold is not unique (interval, a specific order on n with different possible constants).

For the problem (1) and most of the problems in random constraint satisfaction, it is conjectured that there is a threshold and that it is sharp.

For our purposes, the capacity will be the quantity $c(n) = \frac{k_s(n)}{n}$ when it exists. In general, we want c to be a constant. c is the critical number of constraints to the number of variables.

Asking for the capacity is, generally speaking, asking for too much as it is not always true that it exists. For this reason, we will define two relaxed notions related to the capacity. A lower-bound of the capacity will be any c for which a solution is guaranteed to exist whenever $k < cn$. In the same way, an upper bound of the capacity is any c for which no solution will exist with high probability whenever $k > cn$. It is not hard to see that the capacity exists if and only if there exists a lower bound and an upper bound of the capacity which coincide.

It is also worth noting that in addition to this potential threshold, one may consider algorithmic thresholds that will characterize the difficulty or complexity of finding a solution. One expects that the closer to the threshold (from below) the harder it will be to find a solution even if it is guaranteed to exist.

1.6 Some history

When $S_X = S_w = \mathbb{S}^n = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ and $\kappa = 0$ we will use the notation $P_{s,s}(n,k)$ to denote the probability of finding a solution when the patterns are in $\mathbb{S}^n$ and the weights are also in $\mathbb{S}^n$ (the little s stands for $\mathbb{S}^n$). This is the problem of covering $\mathbb{S}^n$ by half spaces generated by patterns from $\mathbb{S}^n$.

In the same way, $P_{b,s}(n,k)$ will be used in the case where $S_X = \mathbb{S}^n$ and $S_w = \{\pm 1\}^n$, this is the problem of covering $\mathbb{S}^n$ by half spaces generated by patterns from $\{\pm 1\}^n$.

In the sixties, it was proven in (Wendel, 1962) and then indirectly in (Cover, 1965) that :

$$P_{s,s}(n,k) = 2^{-k+1} \sum_{i=0}^{n-1} \binom{k-1}{i} \tag{3}$$

(3) is is known as "the function counting theorem" for example in (Cover, 1965).

One of the consequences of formula (3) is that the capacity exists and is equal to 2, this was rediscovered later by Gardner (Gardner, 1988) using the replica trick.

The problem of covering $\mathbb{S}^n$ by binary vectors from $\{\pm 1\}^n$ proved to be a little bit more difficult than the first problem. This problem was, reportedly, first considered by Erdos, and then solved in (Furedi, 1986) who showed that it was, up to some event with a vanishing probability, the same as the first problem (thus they have the same capacity = 2), exactly :

$$P_{b,s}(n,k) = P_{s,s}(n,k) + O(n^{-1/2}) \tag{4}$$

In (4), the term $O(n^{-1/2})$ is the probability that an $n \times n$ $\{\pm 1\}$ matrix taken uniformly at random is singular, this fact was shown in (Komlos, 1967) and then improved to $O(0.999^n)$ in (Kahn, Komlos, & Szemeredi, 1993). The 0.999 was improved by Van Vu and Terence Tao to 0.985, and to $\frac{3}{4}$ and then to $\frac{1}{\sqrt{2}}$ in 2009 ((Tao & Vu, 2005) for example).

1.7 Main contributions

The main contribution of this work is extending the algorithm in (Kim & Roche, 1998) which was engineered for a stability parameter $\kappa = 0$ to general $\kappa \geq 0$.
We also show that the majority rule is optimal amongst all local rules (see 2.2 for a definition), the first consequence of this is that to do better one needs to consider non-local rules. Another consequence is that one cannot use the same principle that was used in (Kim & Roche, 1998) to enhance the majority rule on another local rule as it will be sub-optimal.

2 Basic results

There are two approaches that we will discuss here for solving (1). The first one is what is know as the **majority rule**, the second one we will refer to it, for reasons that will be made known, as the **enhanced majority rule** or **EMR** for briefness.

2.1 Upper bound from the first moment method

Theorem 2.1. *For* $c > \dfrac{ln(2)}{-ln(H(\kappa))}$ *:* $\lim_{n\to\infty} P_\kappa(n, cn) = 0$

Proof : denote

$$Z_n = \sum_{w\in\{\pm 1\}^n} I(\langle w, X_i\rangle \geq \kappa\sqrt{n} \quad \forall i \in \{1,\dots,k\}))$$

So :

$$\begin{aligned} E[Z_n] &= 2^n Pr(< w_0, X_1 >\geq \kappa\sqrt{n})^k \\ &= 2^n H_n(\kappa)^k \\ &= exp\{-n(-c\, ln(H_n(\kappa)) - ln(2))\} \quad ; c = \frac{k}{n} \end{aligned}$$

As $H_n(\kappa) \to \int_\kappa^\infty \dfrac{dt}{\sqrt{2\pi}} exp(-\dfrac{t^2}{2}) := H(\kappa)$ from the central limit theorem, the theorem holds.
For $\kappa = 0$, (Kim & Roche, 1998) proved an upper bound equal to 0.9963 which is better than 1 implied by theorem 2.1.

2.2 The majority rule

The majority rule is the following algorithm :

Algorithm 1. ***Input*** $X \in \{\pm 1\}^{k\times n}$
For $j \in \{1, \cdots, n\}$*:*
  Set $w_j^{maj} = sign(\sum_{i=1}^k X_{ij}) \quad \forall j \in \{1, \cdots, n\}$
Output w^{maj}

Where $X_{ij} = (X_i)_j$, the j-th component of the i-th pattern.
The majority rule is a **local rule** i.e. the j-th component is a function of only the j-th column of X.

2.2.1 Analysis

Notice that the fact of the majority rule being local means that the k components of the vector Xw^{maj} are the sum of n independent and identically distributed random variables in $\mathbb{Z}^k$, it is a **random walk** in $\mathbb{Z}^k$.
To be more precise,

$$Xw^{maj} = \sum_{j=1}^{n} Y_j$$

Where $Y_j = (X_{1j}w_j^{maj}, \dots, X_{kj}w^{maj})^\top$, the components of Y_j are identically distributed (due to symmetry, they are actually what we call exchangeable variables) but they are not independent, they are all correlated. In fact they are **weakly correlated** as we will see.
We have the following computations :

$$Pr(X_{1j}w_j^{maj} = +1) = Pr(\sum_{i=1}^{k} X_{ij}X_{1j} \geq 0) = Pr(\sum_{i=2}^{k} X_{ij}X_{1j} \geq -1) = \frac{1}{2}(1 + \sqrt{\frac{2}{\pi k}}) + o(k^{-1})$$

This derivation is true only for k odd, but it is not difficult to see that when k is even the term $\sqrt{2/(\pi k)}$ (the average of $Y_j^{(1)}$) will be $(3/2)\sqrt{2/(\pi k)}$ which is even greater, this is due to the symmetry of the problem and actually has nothing to do with the convention $sign(0) = +1$. In general the case k even, is not treated because it is easier for the above reason.
The covariance matrix of Y_j is thus as follows:
$var(Y_j^{(i)}) = 1 - 2/(\pi k) + o(k^{-2})$ and $cov(Y_j^{(i)}, Y_j^{(l)}) = o(k^{-1})$ for $i \neq l$ (hence why we say weakly correlated).

2.2.2 Main results

(Fang & Venkatesh, 1998) proved theorem 2.2 using a large deviation local limit theorem (Richter, 1958),(Chaganty, 1986) on random walks in $\mathbb{Z}^k$ (see theorem A.1) for a stability parameter $\kappa = 0$, we verified that the theorem can be extended to $\kappa \geq 0$ without changing anything.

Theorem 2.2. *Let Z_n be the number of rows r violating the inequality $\langle w^{maj}, X_r \rangle \geq \kappa\sqrt{n}$. Then for $k_n = \frac{n}{\pi log(n)}(1 + \frac{(3/2)loglog(n) + log(2\lambda\pi^{3/2})}{log(n)} + O(\frac{loglog(n)^2}{log(n)^2})$. We have :*

$$Z_n \to Poisson(\lambda)$$

This implies in particular that $k_s(n) = n/(\pi log(n))$ is a sharp threshold for the majority rule on problem (1) i.e :
$\forall C < \frac{1}{\pi} : lim_{n\to\infty} P_\kappa(n, \frac{C}{log(n)}n) = 1$ and $\forall C > \frac{1}{\pi} : lim_{n\to\infty} P_\kappa(n, \frac{C}{log(n)}n) = 0$
The fact that we don't need to change anything to go from the case $\kappa = 0$ to $\kappa \geq 0$ can be easily seen from the fact the average of $\sum_{j=1}^{n} Y_j$ is of the order of $n/\sqrt{k} \sim \sqrt{nlog(n)}$ and the variance is of the order $\sqrt{n}$, the worst we can do is $\sqrt{nlog(n)} - \sqrt{n}/u_n = \sqrt{n}(\sqrt{log(n)} - 1/u_n) \geq \kappa\sqrt{n}$ for a suitable choice of $u_n \in o(1)$.
For details on this,refer to Appendix A.

2.2.3 Local rules : beyond the majority rule?

The majority rule is a local rule as discussed before. Let us consider a local rule f .i.e: based on the column X_i we construct $w_i = f(X_i)$. Due to the symmetry of the problem all local rules should be symmetric functions. As the only power sum that is not a constant on $\{\pm 1\}$ variables is their sum, using Girard-Newton formulas we deduce that f only depends on the sum of its inputs, hence all local rules should be functions of the sum, more precisely, there exists a function $g_f \in \{\pm 1\}^{\{k-2l, l=0,\dots,k\}}$ such that $f(x_1,\dots,x_k) = g_f \circ SUM(x_1,\dots,x_k) = g_f(\sum_{j=1}^k x_j)$.
For a given local rule determined by the function g, the variables $\{x_i g(\sum_{j=1}^k x_j)\}_i$ are identically distributed ± 1 random variables. We have the following identities :

$$\mu_g = E[x_1 g(\sum_{j=1}^k x_j)] = 2Pr(x_1 g(\sum_{j=1}^k x_j) = +1) - 1$$

$$cov(x_\alpha g(\sum_{j=1}^k x_j), x_\beta g(\sum_{j=1}^k x_j)) = \mathbb{I}(\alpha = \beta) - \mu_g^2$$

This implies that the best local rule is the rule determined by $g^* \in argmax_g\, \mu_g$.

Theorem 2.3. ***Optimality of the majority rule***
$argmax_g\, \mu_g = \{sign\}$ where sign is the sign function with the freedom of choosing $sign(0)$ to be either $+1$ or -1. In other terms, the majority rule $f(x_1,\dots,x_k) = sign(\sum_{j=1}^k x_j)$ is optimal amongst all local rules.

Proof: see appendix A.1.
This suggests that non-local rules should be considered if we hope to approach a constant capacity as suggested by simulations (Krauth & Opper, 1989) and the replica trick from statistical physics (Krauth & Mézard, 1989b).
Another consequence of this result is that even non-local rules that are built using local rules other than the majority rule in the same fashion as the construction of the enhanced majority rule (2.3) based on the majority rule will be sub-optimal to the enhanced majority rule.

2.3 The enhanced majority rule

In (Kim & Roche, 1998) algorithm 2 which can be seen as an enhancement of the majority rule (hence the name enhanced majority rule) was introduced. By dividing the columns of the matrix X into N blocks and giving the right only to some carefully chosen rows to vote at each step, they were able to leverage the already known threshold of the majority rule in theorem 2.2. As seen in the analysis of the majority rule, the average of Xw^{maj} is of the order $n/\sqrt{k}$ which means that k needs to be small compared to n, but as we want this to work for a constant capacity (k of the same order as n), one of the possibilities is to correct the bad rows (the ones who have a small inner product, this will be made clear).
Before stating what the EMR algorithm is, first we need to introduce some notations. Let $C_0 \cup \dots \cup C_N$ be a partition of $\{1,\dots,n\}$ such that $card(C_i) = n_i$ (and evidently $\sum_{i=0}^N n_i = n$).

We will denote $X(i:j)$ the matrix formed by columns of X from column blocks i to j, i.e. $X(i:j)$ is formed out of columns of X starting from $\sum_{s=0}^{i-1} n_s$ till column $\sum_{s=0}^{j} n_s$. In the same manner, for a vector $z \in \mathbb{R}^n$, $z(i:j)$ will mean the same thing. We will denote $S(i:j) = X(i:j)w(i:j)$. We will refer to $S(0:j)$ as the **cumulative inner products** up to step j.
The EMR algorithm is constituted of N steps. For step $s = 0$ we use the majority rule seen before on $X(0:0)$ to construct $w(0:0)$. At each step $s \geq 1$, we construct the n_s entries of $w(s:s)$ from the block $X(s:s) \in \{\pm 1\}^{k \times n_s}$ by only letting $k_s \leq k$ rows with the smallest cumulative inner product up to step $s-1$ vote.
The EMR algorithm is the following :

Algorithm 2. ***Input*** $X, N, \{k_s, n_s\}$
For s *in* $\{0, \cdots, N\}$ *:*
  If $s = 0$ *then use the majority rule to select* $z(0:0)$
  Else $s \neq 0$ *:*
    Let R_s *be the indices of the* k_s *smallest rows of* $S(0:s-1)$.
    Set : $w_j = sign(\sum_{i \in R_s} X_{ij}) \quad \forall j \in C_s$ *Where* $C_s = \{\sum_{i=0}^{s-1} n_i, \cdots, \sum_{i=0}^{s} n_i\}$
Output w

2.3.1 Main results

Let $f_0 = 1$, $f_1 = 1/200$ and $f_s = 10^{-2^s}$. Denote $A = \sum_{s=0}^{N} f_s$ and let $n_s = (f_s/A)n$. Let $k_0 = c_0 n = 0.005n$, $k_1 = 10^{-8}n$ and $k_s = f_s^3 n$. (we need actually to make the quantities n_s, k_s integers, which is not difficult but as it will make things more complicated than necessary, we will not do it here, the exact formulas can be found in (Kim & Roche, 1998)).
Notice the following good approximations : $n_s \approx f_s n$, $k_s \approx f_s^3 n$, these approximations will be useful in the proofs.

Theorem 2.4. *Using Algorithm 2 with* $N = \lfloor log_2(0.01 log_{10}(n)) \rfloor$, n_s *and* k_s *defined as above, for all* $c \leq c_0 := 0.005$ *:*

$$lim_{n \to \infty} n(1 - P_0(n, cn)) = 0$$

We extend theorem 2.4 for $\kappa \geq 0$

Theorem 2.5. *Generalization of 2.4 to* $\kappa \geq 0$
Using Algorithm 2 with $k_{s,\kappa} = (1 + \sqrt{\frac{\pi c_0}{2}}\kappa)k_s$ *and* N *and* n_s *the same as before, we have the following result:*

$$\forall c \leq c(\kappa) := \frac{c_0}{(1 + \sqrt{\frac{\pi c_0}{2}}\kappa)^2} \; : \; lim_{n \to \infty} n(1 - P_\kappa(n, cn)) = 0$$

2.4 Basic ideas of the proofs

For detailed proofs refer to the Appendix B.3 and B.4. We start here by discussing the basic ideas for the proof of theorem 2.4 and then move to theorem 2.5. We

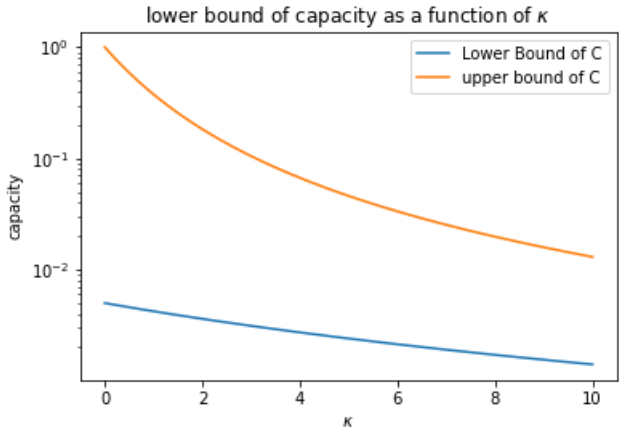

Figure 1: Plot of the lower bound in Th 2.4
and the upper bound in Th 2.1

leverage the fact that when using the majority rule all the variables are exchangeable, this means their joint distribution is unchanged under any permutation. This seemingly simple fact is very powerful (for this see B.1). In particular, with exchangeable random variables, we can upper bound their empirical cumulative distribution function (empirical cdf) by upper bounding their cdf. We can also replace exchangeable variables by i.i.d random variables as long as we only consider a few of them at a given time, with a small price to pay that vanishes as n, k get big.
Using an induction on the steps $s = 0, \dots, N$, the percentage of rows with cumulative inner products up to step s smaller than a threshold $T_s = \sqrt{n}/2^{s-1}$ will be shown to be upper bounded by k_{s+1}/k. The last fact implies in particular that all the rows that will not vote at the next step will be at least as great as the threshold T_s. The last step N, is treated differently. The rows that will not vote are guaranteed with a high probability to be greater than $T_N = \theta(\sqrt{n}/log(n))$, but because n_N is small ($n^{0.99} \leq n_N \leq n^{0.995}$, obtained using the definition of N), with high probability the contribution of the last step will not make any of the cumulative inner products negative. For the rows that will vote at the last step, anyway, we could not have done as worse as $-\sqrt{n}/o(1)$ (up to step $N-1$), because k_N is very small compared to n_N, the positive shift $\sqrt{2/(\pi k_N)}n_N$ will dominate and make the overall inner product positive with high probability.
For proving theorem 2.5, we use the same ideas. In this case, the threshold T_s will be changed to $T_{s,\kappa} = T_s + \kappa\sqrt{n}$, $k = c_0 n$ to $k_\kappa = c(\kappa)n$ and k_s to $k_{s,\kappa}$ which will be chosen such that $k_{s,\kappa}/k_\kappa = k_s/k$. This way if we make sure the new upper bounds of the empirical cdf of the cumulative inner products are tighter than before, we will satisfy directly the upper bound $k_{s,\kappa}/k_\kappa$. The new upper bounds are tighter because the new threshold $T_{s,\kappa}$ is just a translation of the old one and the new positive shift $\sqrt{2/(\pi k_{s,\kappa})} = (1 + \sqrt{\pi c_0/2}\kappa)\sqrt{2/(\pi k_s)}$ is greater than the positive shift for $\kappa = 0$, this makes sense as the constraints now are harder and thus more effort is needed.

2.5 Extension : An adaptive algorithm

From the ideas of the proof of theorem 2.4 and theorem 2.5, we can modify algorithm 2 to choose automatically the rows that should vote at each step s, for this

we need to define the following quantity :$T_{s,\kappa} = \sqrt{n}/2^{s-1} + \kappa\sqrt{n}$.

Algorithm 3. ***Input*** $X, N, \{n_s\}$
For s *in* $\{0, \cdots, N\}$ *:*
  If $s = 0$ *then use the majority rule to select* $z(0:0)$
  Else $s \neq 0$ *:*
    Let R_s *be the indices of the rows of* $S(0:s-1)$ *smaller than* $T_{s,\kappa}$.
    Set : $w_j = sign(\sum_{i \in R_s} X_{ij}) \quad \forall j \in C_s$ *Where* $C_s = \{\sum_{i=0}^{s-1} n_i, \cdots, \sum_{i=0}^{s} n_i\}$
Output w

Theorem 2.6. *The conclusion of theorem 2.5 holds when using Algorithm 3 with the same* N *and* n_s *.*

2.6 How simulations are done?

In (Krauth & Opper, 1989) a simulation of problem 1 was proposed. In the simulation, instead of asking what is the threshold or capacity for a fixed κ, we ask up to what value of κ can we go for a fixed capacity c (i.e $k = cn$).
The following quantity $\kappa_n(c) = E_X[\max_{w \in \{\pm 1\}^n} \min_{i=1,\ldots,cn} \langle X_i, w\rangle/\sqrt{n}]$ was simulated. κ_n is a decreasing piece-wise constant function (constant on intervals of size $1/n$). If we denote $Y_n(c) = \max_{w \in \{\pm 1\}^n} \min_{i=1,\ldots,cn} \langle X_i, w\rangle/\sqrt{n}$, then if we can prove a statement of the form : for n big enough we have with high probability $Y_n(c) \geq K(c)$ and the bound is tight, then we would be able to say that $c(\kappa) = K^{-1}(\kappa)$, and this will tell us for example if the bound is indeed sharp (when $\kappa \mapsto c(\kappa)$ is strictly decreasing). This is difficult, so we follow what good sense suggests : for big n, the curve of κ_n should be near the curve of K.
We run the simulation 1000 times for $n = 11, 13, 15$, here is what we found : This

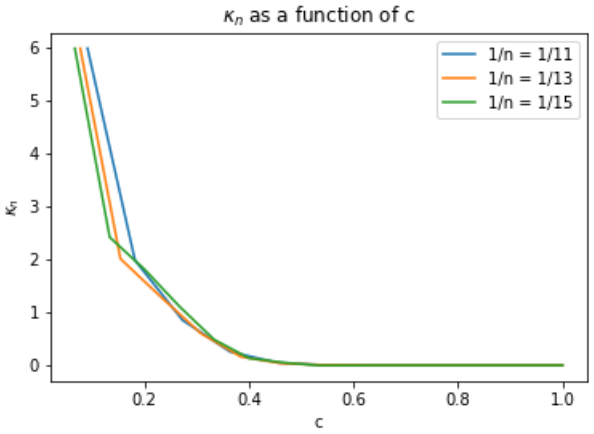

Figure 2: simulated curve of κ_n for $n = 11, 13, 15$ average over 1000 independent trials.

suggests a capacity (for $\kappa = 0$) around 0.7 obtained by a linear interpolation of c as a function of κ and looking for the output when $\kappa = 0$ (the estimation from statistical physics is 0.83).

3 Conclusion

We discussed the problem of linearly separating binary patterns with a binary perceptron, the problem is very interesting from a mathematical perspective as it shows how a seemingly simple problem can be very challenging to mathematicians and computer scientists alike.
In nowadays deep learning, the binary networks that are considered are very different from the problem that was studied here. In general, we mean binarized neural networks (Hubara, Courbariaux, Soudry, El-Yaniv, & Bengio, 2016) which are motivated by being less demanding in memory (with a small price of less precision) and thus offer a good solution for integrating deep learning solutions in small devices. These models are not 100% discreet to still be able to use SGD variants and back-propagation for learning, this luxury is not possible for completely discreet models which makes the task a lot more difficult as we have seen.
To summarize what has been done in this work, the story begun with the rigorous analysis of the majority rule applied to problem 1 with $\kappa = 0$ in (Fang & Venkatesh, 1998) where it has been shown a transition when the number of constraints k_n scales as $n/(\pi log(n))$, we showed that the same transition is maintained for any $\kappa \geq 0$. (Kim & Roche, 1998) proposed an enhanced version of the majority rule that solves problem 1 with $\kappa = 0$ up to a number of constraints $k = 0.005n$, we extended their work to the general case of $\kappa \geq 0$ and $k = c(\kappa)n$ (Th. 2.4). We also proposed a small modification of the algorithm (Alg.3) that makes it possible to choose automatically what and how many rows will vote at each step of the algorithm.
While it might be thought the same can be done for a simple rule that would be similar to the majority rule, we show that as long as the new rule is local, it will be sup-optimal to the majority rule, even more than that, the "enhanced" (based on the procedure in (Kim & Roche, 1998)) version of any local rule will be sup-optimal compared to the enhanced majority rule (Alg.2) and its adaptive variant (Alg.3).

A Proofs related to the majority rule

A.1 Proof of optimality of the majority rule

Maximizing $\mu_g = 2Pr(x_1 g(\sum_{j=1}^{k} x_j) = +1) - 1$ is equivalent to maximizing $Pr(x_1 g(\sum_{j=1}^{k} x_j) = +1)$.

$$\begin{aligned}
Pr(x_1 g(\sum_{j=1}^{k} x_j) = +1) &= \frac{1}{2}(Pr(g(1 + \sum_{j=2}^{k} x_j)) + Pr(g(-1 + \sum_{j=2}^{k} x_j))) \\
&= 2^{-k} \sum_{l=0}^{k-1} \binom{k-1}{l} \{\mathbb{I}(g(k-2l) = +1) + \mathbb{I}(g(k-2l-2) = -1)\} \\
&= 2^{-k} \{\mathbb{I}(g(k) = +1) + \mathbb{I}(g(-k) = -1)\} \\
&+ 2^{-k} \sum_{l=1}^{k-1} \{\binom{k-1}{l} \mathbb{I}(g(k-2l) = +1) + \binom{k-1}{l-1} \mathbb{I}(g(k-2l) = -1)\} \\
&= 2^{-k} \{\mathbb{I}(g(k) = +1) + \mathbb{I}(g(-k) = -1)\} \\
&+ 2^{-k} \sum_{l=1}^{k-1} \binom{k-1}{l} \{\mathbb{I}(g(k-2l) = +1) + \frac{l}{k-l} \mathbb{I}(g(k-2l) = -1)\}
\end{aligned}$$

g maximizes the above sum if and only if $g(k) = +1, g(-k) = -1$ and $(g(k-2l) = 2 \times \mathbb{I}(\frac{l}{k-l} < 1) - 1 = 2 \times \mathbb{I}(l < \frac{k}{2}) - 1 \ \forall l \in \{1, \ldots, k-1\}$ or $g(k-2l) = 2 \times \mathbb{I}(l \leq \frac{k}{2}) - 1 \ \forall l \in \{1, \ldots, k-1\}$). This means that g is the $sign$ function, with the freedom of assigning to zero either $+1$ or -1.

A.2 Proof of the capacity of the majority rule

Theorem A.1. *Local version of the central limit theorem Let $\{U_{nj}\}_{j=1}^{n}$ be n i.i.d random variables taking values in $\{0,1\}^k$ such that $E[U_{nj}] = \mu_n$ and $cov(U_{nj}) = V_n \to V$ non-singular.*
Let $R_n = \sum_{j=1}^{n} U_{nj}$ and denote $\xi_n = \frac{R_n - n\mu_n}{\sqrt{n}}$, then for $\epsilon_n = o(n^{1/6})$:

$$p(\xi_n = \epsilon_n) \sim n^{-k/2} \phi_{V_n}(\epsilon_n)$$

*as $n \to \infty$ (ϕ_{V_n} is the **pdf** of a Gaussian r.v with covariance matrix V_n)*

Proof: is similar to the proof of the central limit theorem with some minor changes. It starts by writing the probability mass function as an inverse Fourier transform of the characteristic function and then using Laplace integration formula with mathematical details to make sure everything is good, the probability mass function is shown to be equivalent to the normal probability density function with the same mean and same variance which depend this time on n hence why we call it a local theorem).
A consequence of the last theorem is this lemma :

Lemma A.2. ***Large deviation global limit theorem***
Let $\epsilon_n = (\epsilon_n^{(1)}, \ldots, \epsilon_n^{(k)})$ *be any sequence such that* $1 \ll \epsilon_n^{(\alpha)} \ll n^{1/6}$ *as* $n \to \infty$*, then under the assumptions of the previous theorem:*

$$Pr(\xi_n \geq \epsilon_n) \sim \Phi_{V_n}(-\epsilon_n) \text{ and } Pr(\xi_n \leq -\epsilon_n) \sim \Phi_{V_n}(-\epsilon_n)$$

Where Φ_{V_n} *is the* ***cdf*** *of a Gaussian r.v with covariance matrix* V_n

In the case of the majority rule, $Xw^{maj} = 2\sum_{j=1}^{n} U_{nj} - n$, where the random variables U_j are i.i.d and take values in $0, 1$.
$\mu_n = E[U_{nj}^{\alpha}] = 1/2 + 1/(\sqrt{2\pi k}) + O(k^{-1})$ and $cov(U_{nj}^{\alpha}, U_{nj}^{\beta}) = (1/4)\mathbb{I}(\alpha = \beta) - 1/(2\pi k) + O(k^{-3/2})$ Which means :

$$V_n = \begin{bmatrix} \sigma_n^2 & \rho_n & \ldots & \rho_n \\ \rho_n & . & . & \rho_n \\ . & . & . & . \\ \rho_n & . & \rho_n & \sigma_n^2 \end{bmatrix}$$

The matrix with $\sigma_n^2 = (1/4) - 1/(2\pi k_n) + O(k_n^{-3/2})$ on the diagonal and $\rho_n = O(k_n^{-1})$ on the off-diagonal entries.

Let m be a fixed number and $1 \leq \alpha_1 < \ldots < \alpha_m \leq k_n$, we denote :

$$p_m = Pr(\langle w^{maj}, X_{\alpha_1} \rangle < \kappa\sqrt{n}, \ldots, \langle w^{maj}, X_{\alpha_m} \rangle < \kappa\sqrt{n})$$

The probability that m distinct constraints are violated at the same time.
Let $n^{2/3} \ll k_n \ll n$, then

$$\begin{aligned} p_m &= Pr(\sum_{j=1}^{n} U_{nj}^{(1)} < \frac{n + \kappa\sqrt{n}}{2}, \ldots, \sum_{l=1}^{n} U_{nj}^{(m)} < \frac{n + \kappa\sqrt{n}}{2}) \\ &= Pr(\sum_{j=1}^{n} U_{nj} < \frac{n + \kappa\sqrt{n}}{2}) = Pr(R_n < \frac{n + \kappa\sqrt{n}}{2}) \\ &= Pr(\xi_n < (\frac{n + \kappa\sqrt{n}}{2} - n\mu_n)\frac{1}{\sqrt{n}}) \\ &= Pr(\xi_n < -\{(\mu_n - \frac{1}{2})\sqrt{n} - \frac{\kappa}{2}\}) = Pr(\xi_n < -\epsilon_n) \end{aligned}$$

where $\epsilon_n = (\mu_n - \frac{1}{2})\sqrt{n} - \frac{\kappa}{2} = \sqrt{\frac{n}{2\pi k_n}} + O(\sqrt{n}k_n^{-1})$ so $1 \ll \epsilon_n \ll n^{1/6}$, we can apply the previous lemma

$$\begin{aligned} p_m &= Pr(\xi_n < -\epsilon_n) \\ &\sim \Phi_{V_n}(-\epsilon_n) = \Phi_{\sigma_n^2 A_{\theta_n}}(-\epsilon_n) \\ &= \Phi_{A_{\theta_n}}(-\epsilon_n/\sigma_n) \\ &\sim \Phi_{A_{\theta_n}}(-\sqrt{\frac{2n}{\pi k_n}}) \end{aligned}$$

$A_{\theta_n} = \sigma_n^{-2} V_n$ is a matrix with diagonal entries equal to 1 and off-diagonal entries equal to $\theta_n = \rho_n/\sigma_n^2 \to 0$.

Lemma A.3. ***multivariate Mill's ration***
$\Phi_{A_n}(-\epsilon_n) \sim \epsilon_n^{-k}\phi_{A_n}(-\epsilon_n)$ *(an equivalence relation between the cdf and the pdf of a Gaussian), this is true for matrices of the type* $A_n = A(\theta_n)$ *with* $\theta_n \to 0$,

Using this lemma,we prove that :

$$p_m \sim (\frac{\sqrt{k_n}}{2\sqrt{n}} e^{-n/(\pi k_n)})^m$$

Which means that when $n \to \infty$ all the dependencies introduced by the majority rule vanish.
Let $Z_n = \sum_{i=1}^{k_n} \mathbb{I}(\langle w^{maj}, X_i\rangle < \kappa\sqrt{n})$, the number of rows that violate the constraint.
There is a classical result that says that in order to show $Z_n \to Poisson(\lambda)$ it is equivalent to show that $E[(Z_n)_m := Z_n(Z_n-1)\cdots(Z_n-m+1)] \to \lambda^k$ for all m.
As $E[(Z_n)_m] = m!\binom{k_n}{m}p_m$ (because of exchangeability), it suffices to show that for all m : $\binom{k_n}{m}p_m \to \frac{\lambda^m}{m!}$, this is not difficult to verify for the threshold given in theorem 2.2.

B Basic Lemmas and more details on the proofs

B.1 Exchangeable random variables

The random variables $X_{ij}w_j$ are dependent which is not good. But while we lost the independence, we did not lose everything, the variables $X_{ij}w_j$ are still what we call exchangeable variables.

Definition B.1 (**Exchangeable variables**). *We say that a sequence* $X_1, \cdots, X_n$ *of random variables is exchangeable if for all permutations* π, *the sequence has the same distribution as* $X_{\pi(1)}, \cdots, X_{\pi(n)}$.

We have the following interesting lemmas that apply to this type of sequences.

Lemma B.1. *Lower bounding exchangeable variables by iid variables*
Let t *be odd, and* b *even with* $b \leq t^{0.1}$.
Let $\psi_1, \cdots, \psi_t$ *be iid symmetric Bernoulli random variables (i.e.* $Pr(\psi_i = 1) = Pr(\psi_i = -1) = 0.5$*) and let* $B = \{m_1, \cdots, m_b\}$ *any be element subset of* $\{1, \cdots, n\}$.
There exists t_0 *independent of* b, *such that for all* $t \geq t_0$ *we can define iid rv's* $\xi_1, \cdots, \xi_b$ *iid rv's such that :*

$$\xi_i \sim Bern(\frac{1}{2}(1 + \sqrt{\frac{2}{\pi t}}(1 - t^{-1/8}))$$

And :

$$\xi_j \leq \psi_{m_j} sign(\sum_{i=1}^{t} \psi_i) \quad \forall j = 1, \cdots, b$$

Lemma B.2. *Controlling* $0-1$ *exchangeable variables by there sum.*
Let $\xi_1, \cdots, \xi_t$ *be exchangeable* $0-1$ *variables and let* $b = t^{0.1}$ *and suppose there exists* $q \in (0,1)$ *such that :*

$$Pr(\xi_1 = \cdots = \xi_b = 1) \leq q^b$$

Then there exits t_0 independent of b such that :

$$Pr(\sum_{i=1}^{t} \xi_i \geq t^{3/5} + (1 + t^{-1/12})qt) \leq exp(-t^{1/70})$$

A basic idea of the proof of lemma B.2 is the following Markov type inequality :

$$Pr(\sum_{i=1}^{t} \xi_i \geq T) \leq \frac{\binom{t}{b} Pr(\xi_1 = \cdots = \xi_b = 1)}{\binom{T}{b}}$$

Which can be proven easily by applying Markov after the following remark:

$$Pr(\sum_{i=1}^{t} \xi_i \geq T) \leq Pr(\sum_{A,|A|=b} \mathbb{I}(\forall i \in A : \xi_i = 1) \geq \binom{T}{b})$$

This last Lemma, shows that we can control the empirical cdf of a sequence of t exchangeable random variables by only controlling at most $b = t^{0.1}$ of these random variables.

Lemma B.3. *Suppose that $\xi_1, ..., \xi_t$ are real-valued exchangeable r.v.'s drawn from the same set of size (at most) M. (M can be a constant or can depend arbitrarily on t.) Also suppose $Pr(\xi_1 \leq u, ..., \xi_b \leq u) \leq [G(u)]^b$ for all $u \in \mathbb{R}$, where $b = 2\lfloor (1/2)t^{0.1}$ and G is a non-negative and non-decreasing real-valued function (a cdf for example).*
Let $\hat{F}$ be the empirical c.d.f. of the ξ's. Then there exists an absolute constant t_0 such that for all $t \geq t_0$,

$$Pr(\hat{F}(u) \leq t^{-2/5} + (1 + t^{-1/12})G(u) \forall u \in \mathbb{R}) \geq 1 - Mexp(t^{-1/70})$$

This last lemma which is a corollary of the lemma just before it shows that upper-bounding the cdf yields an upper bound on the empirical cdf of a collection of exchangeable random variables taking values on a finite set.

Lemma B.4. *Let $U_1, ..., U_t$ be exchangeable real-valued r.v.'s taking values on a finite set (whose size can grow with t), and suppose that, w.h.p., their sample c.d.f. is dominated by $\hat{F}$. Then we can define i.i.d. r.v.'s $W_1, ..., W_t$ (defined on the same space as the U_i 's) with c.d.f. G defined as*

$$G(u) = min\{t^{-1/6} + \hat{F}(u), 1\}\mathbb{I}(u \geq c_1)$$

Where c_1 is the minimum value that can be taken by the U_i 's. with high probability we have the following :

$$U_i \geq W_i \quad \forall i \in \{1, \ldots, t\}$$

Using the last two lemmas we can go from bounds on cdfs to bounds on empirical cdfs, and we can also change a given cdf by its empirical cdf in a given bound by adding negligible terms.
For dealing with Binomial random variables, we need the following definitions and lemmas.

Definition B.2. *For m, j integers and $p \in [0,1]$ we define :*

$$b(m,p,j) = \binom{m}{j} p^j (1-p)^{m-j}$$

And :

$$B(m,p,j) = \sum_{i=0}^{j} b(m,p,i)$$

Lemma B.5. *(**DeMoivre Laplace theorem**). As before, m, j are integers and $p \in (0,1)$ and $\epsilon_1 \in (0, \frac{1}{6})$. Then the central limit theorem is valid if the following conditions are met :*

$$m^{-1/3+\epsilon_1} < p < 1 - m^{-1/3+\epsilon_1} \text{ and } |j - mp| \leq (mp(1-p))^{2/3-\epsilon_1}$$

Furthermore, under the same conditions, there exists $M(\epsilon_1)$ (does not depend on p) such that for all $m \geq M(\epsilon_1)$:

$$\sum_{i \geq mp + (mp(1-p))^{2/3-\epsilon_1}} b(m,p,i) \leq exp(-m^{(4/3)(1/6-\epsilon_1)})$$

and by symmetry

$$\sum_{i \leq mp - (mp(1-p))^{2/3-\epsilon_1}} b(m,p,i) \leq exp(-m^{(4/3)(1/6-\epsilon_1)})$$

This means basically that a binomial random variable $Bin(m,p)$ will concentrate around its mean mp with a variance equal to $mp(1-p)$ when $m \to \infty$ and it will stay in the interval $(mp-(mp(1-p))^{2/3-\epsilon_1}, mp+(mp(1-p))^{2/3-\epsilon_1})$ with high probability (vanishing exponentially with m).

Lemma B.6. *Bounding binomial random variables*
For any $\epsilon > 0$ there exists $M(\epsilon)$ such that $\forall m \geq M(\epsilon)$, $\forall j \in \{1, \cdots, m\}$ and $\forall p \in [1/10, 9/10]$:

$$B(m,p,j) \leq exp(-m^{1/6}) + (1+\epsilon)\Phi((j-mp)/(mp(1-p))^{1/2})$$

Lemma B.7. *Bounding the sum of random variables*
Let Ψ, Ψ', ξ, ξ' be independent random variables such that :

$$F_\xi \leq a + bF_{\xi'} \quad \text{and} \quad F_\Psi \leq a + bF_{\Psi'}$$

Where F_X means the cdf of the random variable X, and a, b, c, d are all positive numbers.
Then :

$$F_{\xi+\Psi} \leq (a+bc) + (bd)F_{\xi'+\Psi'} \quad \text{and} \quad F_{\xi+\Psi} \leq (c+da) + (bd)F_{\xi'+\Psi'}$$

Proof :

$$\begin{aligned} F_{\xi+\Psi}(\eta) &= Pr(\xi + \Psi \leq \eta) \\ &= \int Pr(\xi \leq \eta - w) \mathrm{d}F_{\Psi}(w) \\ &= \int F_{\xi}(\eta - w) \mathrm{d}F_{\Psi}(w) \\ &\leq a + b \int F_{\xi'}(\eta - w) \mathrm{d}F_{\Psi}(w) \\ &= a + bF_{\xi'+\Psi} \leq a + b(c + dF_{\xi'+\Psi'}) \end{aligned}$$

B.2 Proof

Let R be a uniform random variable taking values in the set of row indices of X : $\{1, \cdots, k\}$. We denote $S_R = X_R.w$ to mean the inner product of the R-th row of X by w constructed using Algorithm 1. We will use the following notation : $P_{emp}(S \leq \eta) = Pr(S_R \leq \eta \mid X) = (1/k) \sum_{r=1}^{k} \mathbb{I}(S_r \leq \eta)$ it is the empirical cdf of the random variables $\{S_r, r \in \{1, \cdots, k\}\}$.
Using the above two lemmas on exchangeable variables with $t = k$ one can prove the following result holds with high probability :

$$P_{emp}(S \leq \eta) \leq (9/8)k^{-2/5} + (5/4)\Phi((\eta - \mu_0)/\sigma_0) \tag{5}$$

Φ is the cdf of a Gaussian r.v $\mathcal{N}(0, 1)$; $\mu_0 = n\gamma_0 = n\sqrt{\frac{2}{\pi k}}(1 - k^{-1/8})$ and $\sigma_0^2 = n(1 - \gamma_0^2)$.

B.2.1 Proof of inequality (5)

Using lemma B.1, with $\psi_i = X_{ij}$ and $t = k$ for each $i \in 1, \cdots, b$, there exists $\xi_{i1}, \cdots, \xi_{ib}$ iid such that :

$$\xi_{ij} \leq X_{ij} z_j = X_{ij} sign(\sum_{m=1}^{k} X_{mj})$$

and

$$\xi_{ij} \sim Bern(\frac{1}{2}(1 + \gamma_0))$$

And this for all $i \in \{1, \cdots, b\} = B$ and $j \in \{1, \cdots, n\}$
Then for all $\eta \in \mathbb{R}$:

$$\begin{aligned} Pr(S_i \leq \eta \; \forall i \in B)^{1/b} &= Pr(\sum_{j=1}^{n} X_{ij} z_j \leq \eta \; \forall i \in B)^{1/b} \\ &\leq Pr(\sum_{j=1}^{n} \xi_{ij} \leq \eta \; \forall i \in B)^{1/b} \\ &= Pr(\sum_{j=1}^{n} \xi_{1j} \leq \eta \; \forall i \in B) \\ &\leq exp(-n^{1/6}) + (9/8)\Phi((\eta - \mu_0)/\sigma_0) \quad \text{(Lemma B.6)} \end{aligned}$$

Using Lemma (B.2) we obtain that with high probability:

$$P_{emp}(S \leq \eta) \leq k^{-2/5} + (1 + k^{-1/12})(exp(-n^{1/6}) + (9/8)\Phi((\eta - \mu_0)/\sigma_0))$$

Which shows the result.

B.3 Proofs for the EMR strategy

B.3.1 Notations

As before, R will mean a random row sampled uniformly from $\{1, \cdots, k\}$. R_s as before is the set of rows that participate in the voting at step s, now $S_R(i:j) = X_R(i:j)z(i:j)$ is the inner product of the R-th row of $X(i:j)$ and $z(i:j)$. The following notation will be used for conditioning over the random row index R : $S(i:j \mid A)$ will mean that R is conditioned to be in the subset of row indices A (which can be for example R_s or $\bar{R}_s$). Also $S(i:j)$ will mean $S_R(i:j)$ (i.e. R is any row taking uniformly from the set of rows). And as before :

$$P_{emp}(S(i,j \mid A) \leq \eta) := (1/ \mid A \mid) \sum_{r \in A} \mathbb{I}(S_l(i:j)) \leq \eta)$$

It should be noted in particular that this equation holds true :

$$P_{emp}(S(0:j+1) \leq \eta) = \frac{k_{j+1}}{k} P_{emp}(S(0:j+1 \mid R_{j+1}) \leq \eta) + (1 - \frac{k_{j+1}}{k}) P_{emp}(S(0:j+1 \mid \bar{R}_{j+1}) \leq \eta) \tag{6}$$

This equality combined with the following recursive relation : $S(0, j+1) = S(0, j) + S(j+1, j+1)$ makes it possible to use induction to upper bound $P_{emp}(S(0,s) \leq \eta)$ at each step s of the Algorithm.

Let the following quantities be defined as $T_s = \frac{\sqrt{n}}{2^{s-1}}$, $k_s \approx \frac{n}{1000^{2^s}}$, $n_s \approx \frac{n}{10^{2^s}}$ for each $s = 0, \cdots, N \approx loglog(n)/log(2)$ (the reason behind these choices will be explained shortly).

Following the same steps as in the proof in chapter B.2.1, the same inequality as 5 is obtained, At each step $s \in \{0, \cdots, N\}$:

$$P_{emp}(S(s:s \mid R_s) \leq \eta) \leq (9/8)k_s^{-2/5} + (5/4)\Phi((\eta - \mu(s:s))/\sigma(s:s)) \tag{7}$$

Where $\mu(s:s) = n_s\gamma_s = n_s\sqrt{\frac{2}{\pi k_s}}(1 - k_s^{-1/8})$ and $\sigma(s:s)^2 = n_s(1 - \gamma_s^2)$.

Then using an induction argument based on equation (6), we obtain this general inequality :

$$P_{emp}(S(0:s) \leq \eta) \leq 3^s n^{-1/10} + \Phi((\eta - \mu_{s,1})/\sigma_{s,1}) + (3/2)(1 + n^{-1/10})^s \Phi((\eta - \mu_{s,2})/\sigma_{s,2}) \tag{8}$$

And

$$P_{emp}(S(0:s) \leq T_s) \leq \frac{k_{s+1}}{k} \tag{9}$$

Where $\mu_{s,1} = T_{s-1}$, $\sigma_{s,1}^2 = n_s$, $\mu_{s,2} = n_s\gamma_s \sim n_s\sqrt{\frac{2}{\pi k_s}}$, $\sigma_{s,1}^2 = \sum_{j=0}^{s} n_j$.

(8 and 9) hold with high probability (w.h.p) for all $s \in \{0, \cdots, N-1\}$ and for all $\eta \leq T_s$.

(9) is very important as it implies that for the next step $s+1$, w.h.p

$$P_{emp}(S(0:s \mid \bar{R}_{s+1}) \geq T_s) = 1$$

Which simply means that : $\forall r \in \bar{R}_{s+1} \quad S_r(0:s) \geq T_s) = 1$
The last step is as important as the first step. In fact (8) is only valid up to $s = N-1$. For any row $r \in \bar{R}_N$ we have the following :

$$\begin{aligned} Pr(S_r(0:N) \leq 0) &= Pr(S_r(0:N-1) + S_r(N:N) \leq 0) \\ &\leq Pr(S_r(N:N) \leq -T_{N-1}) \\ &\leq exp(-n_N^{1/6}) + (3/2)\Phi(-\frac{T_{N-1}}{\sqrt{n_N}}) \text{ (Using Lemma B.6)} \\ &\to 0 \text{ (as } n \to \infty) \end{aligned}$$

For the other rows we have the following, we first need to bound $S_r(0:N-1)$ like before.

Lemma B.8. *For each $r \in \{1, \cdots, k\}$ and every $\epsilon \in (0, 1/6)$, there exists $n_0(\epsilon)$ such that for all $n \geq n_0(\epsilon)$:*

$$Pr(S_r(0:N-1) \leq -n^{1/2+\epsilon}) \leq exp(-n^{\epsilon})$$

Using this lemma, for any row $r \in R_N$:

$$\begin{aligned} Pr(S_r(0:N) \leq 0) &= Pr(S_r(0:N-1) + S_r(N:N) \leq 0) \\ &\leq Pr(S_r(N:N) \leq n^{1/2+\epsilon}) \\ &\leq exp(-n_N^{1/6}) + (3/2)\Phi(\frac{n^{1/2+\epsilon} - n_N\gamma_N}{\sqrt{n_N(1-\gamma_N^2)}}) \text{ (Using Lemma B.6)} \end{aligned}$$

Notice $n_N\gamma_N = \sqrt{\frac{2}{\pi k_N}} n_N > n^{0.502} \gg n^{1/2+\epsilon}$ which means that for ϵ small enough (say 0.001) $\frac{n^{1/2+\epsilon} - n_N\gamma_N}{\sqrt{n_N(1-\gamma_N^2)}} \to -\infty$
All in all, $lim_{n\to\infty} n Pr(S_r(0:N) \leq 0) = 0$

B.3.2 Sketch of the proof of 8 and 9

A proof by induction is used.
For $s = 0$, inequality (7) can be easily seen to imply inequality (8). In fact :
$T_0 - \mu_{0,1} \simeq (2 - \sqrt{\frac{2n}{\pi k}})\sqrt{n} = (2 - \sqrt{\frac{400}{\pi}})\sqrt{n} \leq 0$
This shows that we can replace $\sigma_{0:0}$ by $\sigma_{0,1}$ without changing the inequality (as $\sigma_{0:0} \leq \sigma_{0,1}$), it shows also that (9) holds for $s = 0$.
now if we suppose that 8 and 9 hold up to step s then we use the following argument to prove that it will still hold for the next step :

$$P_{emp}(S(0:s+1) \leq \eta) = \frac{k_{s+1}}{k} P_{emp}(S(0:s+1 \mid R_{s+1}) \leq \eta) + (1 - \frac{k_{s+1}}{k}) P_{emp}(S(0:s+1 \mid \bar{R}_{s+1}) \leq \eta) \quad (6)$$

Bounding the second term is easier

$$(1-\frac{k_{s+1}}{k})P_{emp}(S(0:s+1 \mid \bar{R}_{s+1}) \leq \eta) \leq P_{emp}(S(0:s+1 \mid \bar{R}_{s+1}) \leq \eta)$$
$$= P_{emp}(\underbrace{S(0:s \mid \bar{R}_{s+1})}_{w.h.p \geq T_s} + S(s+1:s+1 \mid \bar{R}_{s+1}) \leq \eta)$$
$$\overbrace{\leq}^{w.h.p} P_{emp}(S(s+1:s+1 \mid \bar{R}_{s+1}) \leq \eta - T_s)$$

The variables $\{S_r(s+1:s+1), r \in \bar{R}_{s+1}\}$ are all independent conditioned on knowing $\bar{R}_{s+1}$, their cdf can be bounded using Lemma B.6. Using Lemmas B.3 we can bound their empirical cdf. In the end we will have :

$$(1-\frac{k_{s+1}}{k})P_{emp}(S(0:s+1 \mid \bar{R}_{s+1}) \leq \eta) \preceq \Phi((\eta - T_s)/\sqrt{n_{s+1}}) = \Phi((\eta - \mu_{s+1,1})/\sigma_{s+1,1})$$

Where we use $\preceq$ instead of $\leq$ to say that there are negligible constants that should be added and that will no change anything in the big n, big k regime.
Bounding the first term is more trickier.

$$\frac{k_{s+1}}{k}P_{emp}(S(0:s+1 \mid R_{s+1}) \leq \eta) = \frac{k_{s+1}}{k}P_{emp}(S(0:s \mid R_{s+1}) + S(s+1:s+1 \mid R_{s+1}) \leq \eta)$$

Basically, using the interchangeability of the cumulative rows, we prove in a fashion similar to Lemma B.1 that we can define iid random variables $\{W_i\}$ and $\{\Psi_i\}$ such that :

$$W_i \leq S_i(0:s) \quad \text{and} \quad \Psi_i \leq S_i(s+1:s+1)$$

Where each W_i (resp. Ψ_i) has a statistical cdf bounded by the right hand side of eq (8) (resp. eq (7) by changing s to $s+1$). Note that:

$$P_{emp}(S(0:s \mid R_{s+1}) \leq \alpha) \leq \frac{k}{k_{s+1}}P_{emp}(S(0:s) \leq \alpha)$$

We are able to bound $Pr(W_i \leq \alpha)$ and $Pr(\Psi_i \leq \alpha)$ using Lemmas B.2,B.3,B.4. By lemmaB.8 we can bound the cdf of the sum of W_i and Ψ_i and by lemmaB.2 we will transform it to an upper bound of the empirical cdf of $\{S_r(0:s+1), r \in R_{s+1}\}$. In the end we will have the following :

$$\frac{k_{s+1}}{k}P_{emp}(S(0:s+1 \mid R_{s+1}) \leq \eta) \preceq \Phi(\frac{\eta - (\mu_{s,1} + \gamma_{s+1}n_{s+1})}{\sqrt{\sigma_{s,1}^2 + n_{s+1}}}) + (1+\epsilon)\Phi(\frac{\eta - (\mu_{s,2} + \gamma_{s+1}n_{s+1})}{\sqrt{\sigma_{s,2}^2 + n_{s+1}}})$$

All in all :

$$P_{emp}(S(0:s+1) \leq \eta) \preceq \Phi(\frac{\eta - T_s}{\sqrt{n_{s,1}}})$$
$$+ \Phi(\frac{\eta - (\mu_{s,1} + \gamma_{s+1}n_{s+1})}{\sqrt{\sigma_{s,1}^2 + n_{s+1}}})$$
$$+ (1+\epsilon)\Phi(\frac{\eta - (\mu_{s,2} + \gamma_{s+1}n_{s+1})}{\sqrt{\sigma_{s,2}^2 + n_{s+1}}})$$

The term in the middle $\Phi(\frac{\eta - (\mu_{s,1} + \gamma_{s+1} n_{s+1})}{\sqrt{\sigma_{s,1}^2 + n_{s+1}}})$ can be absorbed in the last term by increasing ϵ, this is true because they are left tails of Gaussians with means of the same order and the variance of the middle term is negligible compared to the variance of the last term (This means the variance is dominated by step $s = 0$).
So :

$$P_{emp}(S(0:s+1) \leq \eta) \preceq \Phi(\frac{\eta - T_s}{\sqrt{n_{s,1}}}) + (1+\epsilon)\Phi(\frac{\eta - (\mu_{s,2} + \gamma_{s+1} n_{s+1})}{\sqrt{\sigma_{s,2}^2 + n_{s+1}}})$$

All that is left is to verify that indeed $P_{emp}(S(0:s+1) \leq \eta) \leq k_{s+2}/k$, this can be done by using upper bounds left tails of Gaussian r.v's.

B.4 Case $\kappa \geq 0$

In this case, let $T_{s,\kappa} = T_s + \kappa\sqrt{n}$.
We want to prove, exactly as before, something of the form :

$$P_{emp}(S(0:s) \leq \eta_\kappa) \preceq \Phi((\eta_\kappa - \mu_{s,1,\kappa})/\sigma_{s,1,\kappa}) + (1+\epsilon)\Phi(\frac{\eta_\kappa - \mu_{s,2,\kappa}}{\sigma_{s,2,\kappa}}) \quad \forall \eta_\kappa \leq T_{s,\kappa}$$

And

$$P_{emp}(S(0:s) \leq T_{s,\kappa}) \leq k_{s+1,\kappa}/k_\kappa$$

It is difficult to keep all the new parameters free and look for them, instead will keep n_s the same and only change k_s and k in the following way : $\frac{k_{s,\kappa}}{k_\kappa} = \frac{k_s}{k}$ where $k_\kappa = c(\kappa)n$, we will use also the following notation $\eta = `\eta_\kappa - \kappa\sqrt{n}$.
$c(\kappa)$ is chosen so that nothing changes at the first step step $s = 0$ .i.e:

$$\implies \frac{T_{s,\kappa} - \sqrt{\frac{2}{\pi k_{0,\kappa}}} n_0}{\sqrt{n_0}} = \frac{T_s - \sqrt{\frac{2}{\pi k_0}} n_0}{\sqrt{n_0}} \implies c(\kappa) := \frac{c_0}{(1 + \sqrt{\frac{\pi c_0}{2}}\kappa)^2}$$

As in the last proof, using the induction we will arrive at

$$P_{emp}(S(0:s+1) \leq \eta_\kappa) \preceq \Phi(\frac{\eta_\kappa - T_{s,\kappa}}{\sqrt{n_{s,1}}}) + \Phi(\frac{\eta_\kappa - (\mu_{s,1,\kappa} + \gamma_{s+1,\kappa} n_{s+1})}{\sqrt{\sigma_{s,1}^2 + n_{s+1}}}) + (1+\epsilon)\Phi(\frac{\eta_\kappa - (\mu_{s,2,\kappa} + \gamma_{s+1,\kappa} n_{s+1})}{\sqrt{\sigma_{s,2}^2 + n_{s+1}}})$$

Note that

- $\eta_\kappa \leq T_{s,\kappa} \iff \eta \leq T_s$

- $\gamma_{s,\kappa} = (1 + \sqrt{\frac{\pi c_0}{2}}\kappa)\gamma_s$
- $\eta_\kappa - T_{s,\kappa} = \eta - T_s$
- $\eta_\kappa - (\mu_{s,1,\kappa} + \gamma_{s+1,\kappa} n_{s+1}) = \eta - (\mu_{s,1} + \gamma_{s+1,\kappa} n_{s+1}) \leq \eta - (\mu_{s,1} + \gamma_{s+1} n_{s+1})$
- $\eta_\kappa - (\mu_{s,2,\kappa} + \gamma_{s+1,\kappa} n_{s+1}) \leq \eta - (\mu_{s,2} + \gamma_{s+1} n_{s+1})$

From the above points and the fact that Φ is a strictly increasing function, we are assured that we will be dominated by the bounds that we had for the case $\kappa = 0$, and thus the choice $\frac{k_{s,\kappa}}{k_\kappa} = \frac{k_s}{k}$ makes everything work perfectly.
For the last step $s = N$, we do the same as before.

References

Chaganty, N. R. (1986). Multidimensional large deviation local limit theorems. *JOURNAL OF MULTIVARIATE ANALYSIS 20, 19O204*.

Coja-Oghlan, A. (2009). Random constraint satisfaction problems. In S. B. Cooper & V. Danos (Eds.), *Proceedings fifth workshop on developments in computational models–computational models from nature, DCM 2009, rhodes, greece, 11th july 2009* (Vol. 9, pp. 32–37). Retrieved from https://doi.org/10.4204/EPTCS.9.4 doi: 10.4204/EPTCS.9.4

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans.Electronic Comput.14, 326-331*.

Fang, S., & Venkatesh, S. (1998). The capacity of majority rule. *Random Structures and Algorithms 12 , 83.109*.

Furedi, Z. (1986). Random polytopes in the d-dimensional cube. *Discrete Comput.Geom.1, 315-319*.

Gardner, E. (1988). The space of interactions in neural networks models. *J. Phys. A: Math. Gen., 21:257–270*.

Gardner, E., & Derrida, B. (1988). Optimal storage properties of neural network models. *J. Phys. A: Math. Gen. 21 (1988) 271-284*.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Binarized neural networks. *https://papers.nips.cc/paper/2016/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf*.

Kahn, J., Komlos, J., & Szemeredi, E. (1993). Singularity probability for random ±1 matrices. *preprin*.

Kim, J. H., & Roche, J. R. (1998). Covering cubes by random half cubes, with applications to binary neural networks. *Journal of Computer and System Sciences 56, 223-252*.

Komlos, J. (1967). On the determinant of (0, 1) matrices. *Studia Sci.Math.Hung.2, 7-21*.

Krauth, W., & Mézard, M. (1989a). Storage capacity of memory networks with binary couplings. *Journal de Physique, 50 (20), pp.3057-3066.*.

Krauth, W., & Mézard, M. (1989b). Storage capacity of memory networks with binary couplings. *J. Physique 50, 3057-3066*.

Krauth, W., & Opper, M. (1989). Critical storage capacity of the $j = \pm 1$ neural network. *J. Phys. A: Math. Gen. 22 , L519-L523*.

Richter, W. (1958). Multidimensional local limit theorems for large deviations. *Theory Probab. Appl., 3, 100-106*.

Tao, T., & Vu, V. (2005). On the singularity probability of random bernoulli matrices. *https://arxiv.org/abs/math/0501313*.

Wendel, J. (1962). A problem in geometric probability. *Math.Scand.11, 109-111*.