The pivotal set of a Boolean function

Raphaël Cerf*

May 29, 2024

Abstract

We define the pivotal set of a Boolean function and we prove a fundamental inequality on its expected size, when the inputs are independent random coins of parameter p. We give two complete proofs of this inequality. Along the way, we obtain the classical Margulis–Russo formula. We give a short proof of the classical Hoeffding inequality for i.i.d. Bernoulli random variables, and we use it to derive more complex deviations inequalities associated to the pivotal set. We follow finally Talagrand's footsteps and we discuss a beautiful inequality that he proved in the uniform case.

1 Introduction

A boolean function is a function which takes only two values 0 and 1, it transforms a very complex input into a binary result. Consider for instance a function of an image which is meant to answer a complex question, like:
• Is there a trumpet somewhere in the image?
Or for more serious and important applications:
• Airport safety: Based on the scanner image, should a luggage be checked?
• Medicine: Is there a suspicion of cancer in the MRI image?
While the first question can be easily answered by anyone, the second can only be answered properly by qualified baggage screeners, and the third by skilled doctors. Today, neural networks are trained on huge databases using deep learning techniques and they achieve excellent results on these problems, which we only dreamed of twenty years ago. In each case, the outcome of the learning process can be seen as a boolean function of the pictures, albeit a very complex one, which depends on billions of parameters. A related problem is to guess the value of a boolean function when only partial information is available. The difficulty is to understand which part of the input of the boolean function plays a crucial role to decide its output. The analysis of boolean functions has grown into a very active field of research, due to its multiple applications in several branches of science (see for instance the book of O'Donnell [7]). Our goal here is to introduce the notion of the pivotal set, which is indeed pivotal in the understanding of a boolean function. We state and prove in full detail a basic and

*Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay.

classical inequality controlling the expected size of the pivotal set. This proof is not new, it is certainly known to experts in the field, however we feel it worth presenting and discussing the full details, all the more that introductory texts on Boolean functions typically rely on tools from the Fourier analysis to derive it. We shall then develop deviations estimates associated to this inequality, which will gradually lead us to a beautiful inequality due to Talagrand. This inequality is a part of a vast circle of ideas originating in harmonic analysis, which Talagrand developed to obtain several revolutionary inequalities. We will mention only two of these: Talagrand's famous concentration inequality and the isoperimetric inequality on the hypercube [9]. The inequality presented here relies on some concentration estimates, but it goes in the opposite direction to the isoperimetric inequality.

The project to revisit some intermediate results of Talagrand, with an adaptation to the biased case, started two years ago. The recent attribution of the Abel prize to Talagrand prompted me to conclude this project, indeed this is an opportunity to explain an important contribution of Talagrand which is less known than his other accomplishments. However, before embarking into this program, we need to introduce some notation and to precise the model we shall study.

The pivotal set. Let $n \geq 1$ and let f be a boolean function defined on $\Omega = \{0,1\}^n$, i.e.,

$$f : \omega = \big(\omega(1) \cdots \omega(n)\big) \mapsto f(\omega) \in \{0,1\}.$$

Given a configuration ω, we say that the i–th coordinate $\omega(i)$ of ω is pivotal if the value of $f(\omega)$ is decided by $\omega(i)$, or equivalently, if changing the value of $\omega(i)$ changes the value of f. More precisely, for $\omega \in \Omega$ and $i \in \{1,\dots,n\}$, we denote by ω^i (respectively ω_i) the configuration ω where the i–th component is set to 1 (respectively to 0), that is,

$$\omega^i(j) = \begin{cases} \omega(j) & \text{if } j \neq i \\ 1 & \text{if } j = i \end{cases}, \qquad \omega_i(j) = \begin{cases} \omega(j) & \text{if } j \neq i \\ 0 & \text{if } j = i \end{cases}.$$

The i–th coordinate $\omega(i)$ is said to be pivotal for the function f in the configuration ω if $f(\omega^i) \neq f(\omega_i)$. We denote by $\mathcal{P}(f,\omega)$ the set of the pivotal coordinates of ω for f. In general, little can be said on the pivotal set, because it depends in a complicated way on both ω and f. So we will focus on the specific situation where f is non–decreasing, meaning that

$$\forall \omega \in \Omega \quad \forall i \in \{1,\dots,n\} \qquad f(\omega_i) \leq f(\omega^i),$$

and the configuration ω is chosen randomly. The pivotal set $\mathcal{P}(f,\omega)$ becomes de facto a random subset of $\{1,\dots,n\}$.

Randomness. To go further, we must specify the random distribution of ω. We fix a parameter p in $[0,1]$ and, to decide the status of each component $\omega(i)$, we draw a coin of parameter p, the coins being independent. For $p = \frac{1}{2}$, this corresponds to the uniform distribution over Ω.

Influences. The probability that the i–th coordinate belongs to $\mathcal{P}(f,\omega)$ is called the influence of the coordinate i. It provides a mean of quantifying the importance of one specific coordinate to the output of the boolean function, and also to compare the relative importance of two distinct coordinates. The sum of the influences of all the coordinates is called the total influence, or the averaged sensitivity.

So, why is it important to control the pivotal set of a boolean function? Friedgut's junta theorem tells that, whenever the total influence of a function f is small, the function f can be approximated by a junta, that is a function depending on a small number of coordinates. With the help of this result, O'Donnell and Servidio [8] developed an algorithm able to learn a monotone function f to any constant accuracy, in time polynomial in n and in the decision tree size of f. On a more general level, the links between the structure of a boolean function f and the probabilistic properties of its pivotal set are complex and still mysterious. A famous theorem of Benjamini, Kalai and Schramm [1] states that, if the sum of the squares of the influences of a function is small, then it is noise sensitive: even if two inputs are highly correlated, the outputs of the function might be highly non–correlated. The noise sensitivity theorem relies on the deep inequality due to Talagrand that we shall present at the end of this text.

We come now to the first mathematical result presented here. A simple exchange between summation and expectation shows that the total influence is in fact equal to the expected size of the pivotal set. It turns out that, surprisingly, we can compute an explicit upper bound on the expected size of the pivotal set.

Theorem 1.1. *For any non–decreasing boolean function f, we have*

$$E\big(|\mathcal{P}(f)|\big) \,\leq\, \sqrt{\frac{nE(f)}{p(1-p)}}\,. \tag{1.1}$$

Naturally, the symbol E is the expectation with respect to the probability P on Ω defined by

$$\forall\omega\in\Omega \qquad P(\omega) \,=\, \prod_{1\leq i\leq n} p^{\omega(i)}(1-p)^{1-\omega(i)}\,,$$

and we remove the variable ω in the expectations, for instance

$$E(f) \,=\, \sum_{\omega\in\Omega} f(\omega)P(\omega)\,.$$

To the best of our knowledge, this result was proved for the first time by Friedgut and Kalai (see lemma 6.1 in [3]). Benjamini, Kalai and Schramm quote this result (theorem 3.3 in [1]) and they mentioned that it is a consequence of [2] (although it does not appear explicitly there).

Application to the majority function. Suppose that the configuration ω represents the votes in an election, and that the votes are independent and follow a Bernoulli distribution. The function f is taken to be the majority function, which is equal to 1 if there is a majority of votes equal to 1 and 0 otherwise. In this model, all the votes play the same role, so that $E\big(|\mathcal{P}(f)|\big) = nP(1 \text{ is pivotal})$ and theorem 1.1 yields that

$$P(1 \text{ is pivotal}) \,\leq\, \frac{1}{\sqrt{np(1-p)}}\,, \tag{1.2}$$

hence the probability that a fixed vote plays a decisive role goes to 0. Moreover, in the uniform case $p = 1/2$, the inequality (1.2) captures the correct order, indeed it can be proved that

$$P(1 \text{ is pivotal}) \,\sim\, \sqrt{\frac{2}{\pi n}} \qquad \text{as} \qquad n \to \infty\,.$$

However the constant in inequality (1.1) is not optimal. In fact, the original argument of Friedgut and Kalai [3] consisted in proving that the majority function maximizes $E\big(|\mathcal{P}(f)|\big)$ over all the boolean functions f.

Our first major goal is to provide the most elementary proof of theorem 1.1. The key idea consists in expressing the probability that a site i is pivotal with the help of a discrete derivative operator δ_i. As a by product, this proof yields another important formula due to Margulis and Russo. We shall next work to obtain a better control on the pivotal set, with a technique involving the Parseval formula in Fourier analysis. It is the strategy employed in modern textbooks on boolean functions to prove theorem 1.1 (see theorem 2.33 of [7]), in fact it yields a stronger result. In order to understand better the link between pivotal coordinates and the value of f, we shall compute the conditional expectation of the size of the pivotal set knowing the value of the function. Our second major goal is to obtain more complex deviations inequalities associated to the pivotal set. The key tool here is the classical Hoeffding inequality (to make the text self–contained, we give a short proof for the case of i.i.d. Bernoulli random variables). In the end, we extend to the Bernoulli product measure of parameter p an inequality that Talagrand proved in the case $p = 1/2$. We follow finally Talagrand's footsteps and we discuss a beautiful inequality that he proved in the uniform case $p = 1/2$. Talagrand's inequality is a far–reaching extension of theorem 1.1. The starting point is the deviations inequality that is developed in the second major goal of the text. The very statement and proof of the inequality testify to Talagrand's virtuosity. The noise sensitivity theorem of Benjamini, Kalai and Schramm [1] relies in a crucial way upon a further generalization of this inequality.

2 The discrete derivative δ_i

The first step to prove the inequality (1.1) is to express $|\mathcal{P}(f,\omega)|$ with the help of the discrete derivative of f along the i–th component, which is defined as

$$\forall \omega \in \Omega \qquad \delta_i f(\omega) \;=\; f(\omega^i) - f(\omega_i)\,. \tag{2.1}$$

Indeed, the i–th coordinate $\omega(i)$ of ω is pivotal if and only if $\delta_i f(\omega) = 1$, therefore

$$|\mathcal{P}(f,\omega)| \;=\; \sum_{i=1}^{n} \delta_i f(\omega)\,. \tag{2.2}$$

Taking the expectation and using the linearity, we obtain

$$E\big(|\mathcal{P}(f,\omega)|\big) \;=\; E\Big(\sum_{i=1}^{n} \delta_i f(\omega)\Big) \;=\; \sum_{i=1}^{n} E\big(\delta_i f(\omega)\big)\,. \tag{2.3}$$

We have now to evaluate

$$E\big(\delta_i f(\omega)\big) \;=\; E\big(f(\omega^i)\big) - E\big(f(\omega_i)\big)\,. \tag{2.4}$$

To that end, we will rely on a little trick, recorded in the next lemma. The simplicity of the calculation should not make us underestimate its power.

Lemma 2.1. *Let f be an arbitrary function defined on Ω with values in $\mathbb{R}$. For any $i \in \{\,1,\dots,n\,\}$, we have*

$$\sum_{\omega\in\Omega} f(\omega)\,\omega(i)\,P(\omega) \;=\; \sum_{\omega\in\Omega} p\,f(\omega^i)\,P(\omega)\,. \tag{2.5}$$

Proof. The identity (2.5) can be verified in several ways. Our presentation is a bit heavy, but it does not make appeal to any new notation. So, we write

$$\begin{aligned}
\sum_{\omega\in\Omega} f(\omega)\,\omega(i)\,P(\omega) \;&=\; \sum_{\omega(1),\cdots,\omega(n)\in\{0,1\}} f\big(\omega(1)\cdots\omega(n)\big)\,\omega(i) \prod_{1\le j\le n} p^{\omega(j)}(1-p)^{1-\omega(j)} \\
&= \sum_{\omega(1)}\cdots\sum_{\omega(i-1)}\sum_{\omega(i+1)}\cdots\sum_{\omega(n)} f\big(\omega(1)\cdots\omega(i-1)\,1\,\omega(i+1)\cdots\omega(n)\big) \\
&\qquad\qquad \times p \prod_{\substack{1\le j\le n\\ j\ne i}} p^{\omega(j)}(1-p)^{1-\omega(j)} \\
&= \sum_{\omega(1)}\cdots\sum_{\omega(i-1)}\sum_{\omega(i+1)}\cdots\sum_{\omega(n)} f\big(\omega^i\big)\,P(\omega^i)\,.
\end{aligned} \tag{2.6}$$

Now the summand $f\big(\omega^i\big)\,P(\omega^i)$ does not depend any more on $\omega(i)$, so we reintroduce artificially the summation over $\omega(i)$ by writing

$$f\big(\omega^i\big)\,P(\omega^i) \;=\; \sum_{\omega(i)} f\big(\omega^i\big)\,p\,P(\omega)\,. \tag{2.7}$$

We plug the equation (2.7) into (2.6) and we obtain the identity (2.5). $\square$

The next corollary presents the analogous formula dealing with $1-\omega(i)$ instead of $\omega(i)$. Of course, this formula could be obtained with a direct computation, but we will derive it from lemma 2.1.

Corollary 2.2. *For any function f from Ω to $\mathbb{R}$ and for any $i \in \{1,\dots,n\}$, we have*

$$\sum_{\omega\in\Omega} f(\omega)\,\big(1-\omega(i)\big)\,P(\omega) \;=\; \sum_{\omega\in\Omega}(1-p)\,f(\omega_i)\,P(\omega)\,. \tag{2.8}$$

Proof. For $\omega \in \Omega$, we denote by $\overline{\omega}$ the configuration obtained by flipping all the components of ω:

$$\forall j \in \{1,\dots,n\} \qquad \overline{\omega}(j) \;=\; 1-\omega(j)\,.$$

Of course, the distribution of ω under P is the same as the distribution of $\overline{\omega}$ under P_{1-p}, hence

$$\begin{aligned}\sum_{\omega\in\Omega} f(\omega)\,\big(1-\omega(i)\big)\,P(\omega) \;&=\; \sum_{\omega\in\Omega} f(\overline{\omega})\,\big(1-\overline{\omega}(i)\big)\,P_{1-p}(\omega)\\ &=\; \sum_{\omega\in\Omega} f(\overline{\omega})\,\omega(i)\,P_{1-p}(\omega)\,.\end{aligned} \tag{2.9}$$

We apply lemma 2.1 to the function $\omega \mapsto f(\overline{\omega})$ and the probability measure P_{1-p} and we get

$$\begin{aligned}\sum_{\omega\in\Omega} f(\overline{\omega})\,\omega(i)\,P_{1-p}(\omega) \;&=\; \sum_{\omega\in\Omega}(1-p)\,f\big(\overline{\omega^i}\big)\,P_{1-p}(\omega)\\ &=\;(1-p)\sum_{\omega\in\Omega} f(\overline{\omega}_i)\,P_{1-p}(\omega) \;=\;(1-p)\sum_{\omega\in\Omega} f\big(\omega_i\big)\,P_p(\omega)\,.\end{aligned} \tag{2.10}$$

Putting together formulas (2.9) and (2.10), we obtain (2.8). ☐

Taking advantage of lemma 2.1 and corollary 2.2, we have

$$E\big(f(\omega^i)\big) - E\big(f(\omega_i)\big) \;=\; \sum_{\omega\in\Omega} f(\omega)\Big(\frac{\omega(i)}{p} - \frac{1-\omega(i)}{1-p}\Big)\,P(\omega)\,. \tag{2.11}$$

3 Computation of $E\big(|\mathcal{P}(f)|\big)$

The righthand side of formula (2.4) has been computed in formula (2.11). In order to rewrite this last formula in a more concise form, we introduce the random variable X_i defined by

$$\forall \omega\in\Omega \qquad X_i(\omega) \;=\; \frac{\omega(i)}{p} - \frac{1-\omega(i)}{1-p}\,, \tag{3.1}$$

and, substituting (2.11) into (2.4), we obtain

$$E\big(fX_i\big) \;=\; E\big(\delta_i f\big)\,. \tag{3.2}$$

In the next step, we sum this formula over $i \in \{ 1, \dots, n \}$ and we get

$$\sum_{1 \le i \le n} E\big(fX_i\big) \;=\; \sum_{1 \le i \le n} E\big(\delta_i f\big)\,. \tag{3.3}$$

In the final step, we put the sum inside the expectation, we combine the above formula with the initial identity (2.2) and we get the formula stated in the next proposition.

Proposition 3.1. *For any non–decreasing boolean function f, we have*

$$E\big(|\mathcal{P}(f)|\big) \;=\; E\big(S_n f\big)\,, \tag{3.4}$$

where S_n is the random sum $S_n \;=\; X_1 + \cdots + X_n$.

In fact, lemma 2.1, corollary 2.2 and proposition 3.1 can be viewed as simple consequences of the following elementary formulas. Let $\omega(1)$ be a Bernoulli random variable with parameter p:

$$P(\omega(1)=0) = 1-p\,, \qquad P(\omega(1)=1) = p\,.$$

Let f be a function from $\{ 0, 1 \}$ to $\mathbb{R}$. We have

$$f(1) - f(0) \;=\; E\bigg(f(\omega(1))\Big(\frac{\omega(1)}{p} - \frac{1-\omega(1)}{1-p}\Big)\bigg)\,. \tag{3.5}$$

To prove directly formula (3.2), we compute $E\big(fX_i\big)$ with the help of Fubini's theorem. More precisely, we fix the variables $\omega(1), \dots, \omega(i{-}1), \omega(i{+}1), \dots, \omega(n)$ and we compute first the expectation with respect to $\omega(i)$. This amounts to take the conditional expectation of f given $\omega(1), \dots, \omega(i{-}1), \omega(i{+}1), \dots, \omega(n)$. From formula (3.5), we have

$$\begin{aligned} E\big(fX_i \,|\, \omega(1), \dots, \omega(i-1), \omega(i+1), \dots, \omega(n)\big) \;=&\\ \delta_i f\big(\omega(1), \dots, \omega(i-1), \cdot, \omega(i+1), \dots, \omega(n)\big)\,.& \end{aligned} \tag{3.6}$$

Naturally, the resulting function does not depend any more on the variable $\omega(i)$. Taking the expectation of formula (3.6), we obtain formula (3.2). To be frank, this is just another presentation of the computation done in the proof of lemma 2.1, which might look more natural for readers acquainted with the conditional expectation, but which is not as elementary.

4 Completion of the proof of theorem 1.1

We start from the formula of proposition 3.1:

$$E\big(|\mathcal{P}(f)|\big) \;=\; E\big(S_n f\big) \;=\; \sum_{\omega \in \Omega} S_n(\omega) f(\omega) P(\omega)\,. \tag{4.1}$$

A straightforward application of the Cauchy–Schwarz inequality gives

$$\Big|\sum_{\omega\in\Omega} S_n(\omega)f(\omega)P(\omega)\Big| \;\leq\; \sqrt{\Big(\sum_{\omega\in\Omega}\big(S_n(\omega)\big)^2 P(\omega)\Big)\Big(\sum_{\omega\in\Omega}\big(f(\omega)\big)^2 P(\omega)\Big)}\,.$$

This inequality can be rewritten in a more concise form with the help of the expectation E as

$$\big|E\big(S_n f\big)\big| \;\leq\; \sqrt{E\big((S_n)^2\big)E\big((f)^2\big)}\,. \tag{4.2}$$

The random variable S_n is the sum of n i.i.d. centered variables, so that

$$E\big((S_n)^2\big) \;=\; \sum_{1\leq i\leq n} E\big((X_i)^2\big) \;=\; \frac{n}{p(1-p)}\,. \tag{4.3}$$

Putting together (4.1), (4.2) and (4.3), we obtain the inequality stated in theorem 1.1.

5 The Margulis–Russo formula

We state here an important consequence of proposition 3.1.

Proposition 5.1. *For any non–decreasing boolean function f, we have*

$$\frac{d}{dp}E(f) \;=\; E\big(|\mathcal{P}(f)|\big)\,. \tag{5.1}$$

The Margulis–Russo formula (5.1) is very convenient to study the response of $E(f)$ to a slight variation of the parameter p, for instance to analyze threshold phenomena. Once we have proposition 3.1, the proof is straightforward. We take the derivative of $E(f)$ with respect to the parameter p:

$$\frac{d}{dp}E(f) \;=\; \frac{d}{dp}\sum_{\omega\in\Omega} f(\omega)\,P(\omega) \;=\; \sum_{\omega\in\Omega} f(\omega)\,\frac{d}{dp}P(\omega)\,.$$

We differentiate $P(\omega)$ either as an n–fold product, or through its logarithmic derivative. Indeed, we have

$$\forall\omega\in\Omega \qquad \ln P(\omega) \;=\; \sum_{1\leq i\leq n}\Big(\omega(i)\ln p + (1-\omega(i))\ln(1-p)\Big)\,.$$

Whichever way, we get, for any $\omega\in\Omega$,

$$\frac{d}{dp}P(\omega) \;=\; \sum_{1\leq i\leq n} X_i(\omega)\,P(\omega) \;=\; S_n(\omega)\,P(\omega)\,, \tag{5.2}$$

whence

$$\frac{d}{dp}E(f) \;=\; E\big(fS_n\big)\,,$$

and formula (5.1) follows immediately from proposition 3.1.

6 An alternative proof via Bessel's inequality

Techniques coming from the Fourier analysis play a central role in the study of boolean functions. We will show here how a stronger result than theorem 1.1 can be derived from the classical Bessel inequality. We use the random variable X_i defined in (3.1) and the formula (3.2) to write

$$P\big(i \text{ is pivotal for } f\big) \;=\; E\big(fX_i\big)\,. \tag{6.1}$$

We endow the space $\mathcal{F}(\Omega, \mathbb{R})$ of the functions from Ω to $\mathbb{R}$ with the scalar product

$$(f,g) \in \mathcal{F}(\Omega,\mathbb{R}) \mapsto \langle f, g\rangle \;=\; \sum_{\omega\in\Omega} f(\omega)g(\omega)P(\omega)\,.$$

We denote by $||\cdot||_2$ the associated norm on $\mathcal{F}(\Omega, \mathbb{R})$. This turns $\mathcal{F}(\Omega, \mathbb{R})$ into a Hilbert space. Moreover the n random variables $X_1, \dots, X_n$ form an orthogonal family, more precisely we have

$$\forall i,j \in \{\,1,\dots,n\,\} \qquad \langle X_i, X_j\rangle \;=\; \begin{cases} 0 & \text{if } i \neq j\,, \\ \dfrac{1}{p(1-p)} & \text{if } i = j\,. \end{cases} \tag{6.2}$$

From this perspective, the quantity $E\big(fX_i\big)$ is simply the scalar product $\langle f, X_i\rangle$ and a direct application of Bessel's inequality yields that

$$\sum_{1\leq i\leq n} \frac{\langle f, X_i\rangle^2}{\langle X_i, X_i\rangle^2} \;\leq\; \big(||f||_2\big)^2 \;=\; E(f)\,. \tag{6.3}$$

Combining the identities (6.1), the little computation (6.2) and inequality (6.3), we obtain the inequality stated in the next proposition.

Proposition 6.1. *For any non–decreasing boolean function f, we have*

$$\sum_{1\leq i\leq n} \Big(P\big(i \textit{ is pivotal for } f\big)\Big)^2 \;\leq\; \frac{E(f)}{p(1-p)}\,. \tag{6.4}$$

This last inequality (6.4) is in fact stronger than the inequality (1.1). Indeed, it follows from the Cauchy–Schwarz inequality that

$$E\big(|\mathcal{P}(f)|\big) \;=\; \sum_{1\leq i\leq n} P\big(i \text{ is pivotal for } f\big) \;\leq\; \sqrt{n\sum_{1\leq i\leq n}\Big(P\big(i \text{ is pivotal for } f\big)\Big)^2}\,,$$

and using (6.4) to bound the last term, we recover the inequality (1.1).

7 The conditional expectation of $|\mathcal{P}(f)|$

We will be interested in controlling the cardinality of the pivotal set for f whenever the function f is equal to 1. Our first step in this direction consists in derivating a formula involving the conditional expectation $E\big(|\mathcal{P}(f)|\,\big|\,f=1\big)$. We focus again on the specific situation where f is non–decreasing. In this case, whenever a coordinate i is pivotal for f, the value of f is equal to $\omega(i)$. Moreover the status of the coordinate i itself is independent of the fact that i is pivotal or not, therefore

$$P\big(i\in\mathcal{P}(f),\,f=1\big)\;=\;P\big(i\in\mathcal{P}(f),\,\omega(i)=1\big)\;=\;P\big(i\in\mathcal{P}(f)\big)\,p\,,\tag{7.1}$$

which we rewrite as

$$P\big(\,i\in\mathcal{P}(f)\,\big)\;=\;\frac{1}{p}P\big(i\in\mathcal{P}(f),\,f=1\big)\,.\tag{7.2}$$

Summing equation (7.2) over i, and putting the summation inside the expectation, we obtain

$$E\big(|\mathcal{P}(f)|\big)\;=\;\frac{1}{p}E\big(|\mathcal{P}(f)|\,f\big)\,.\tag{7.3}$$

The identities (3.4) and (7.3) yield

$$E\big(S_nf\big)\;=\;\frac{1}{p}E\big(|\mathcal{P}(f)|\,f\big)\,.\tag{7.4}$$

Dividing finally by $P(f=1)$, we obtain the formula stated in the next lemma.

Lemma 7.1. *For any non–decreasing boolean function f, we have*

$$E\big(S_n\,\big|\,f=1\big)\;=\;\frac{1}{p}E\big(|\mathcal{P}(f)|\,\big|\,f=1\big)\,.\tag{7.5}$$

The previous formula (7.5) is very interesting. Indeed, it tells us that, whenever the boolean function f takes the value 1, the expected size of the pivotal set coincides with the conditional expectation of S_n knowing that $f=1$. Up to an affine rescaling, the sum S_n is a binomial random variable and probabilists know very well how to control the binomial distribution in various regimes. Ultimately, we would like to link the size of the pivotal set $\mathcal{P}(f)$ to the probability $P(f=1)$. Our strategy to do so is the following. Suppose that the righthand side of (7.5) is large. This forces that the conditional law of S_n knowing that $f=1$ is concentrated on subsets of Ω where S_n is large, but these sets have small probability, so for the ratio $E(S_nf)/P(f=1)$ to be large, the event $f=1$ must also have a small probability.

8 Hoeffding's inequality

The key tool to complete the previous program is the classical Hoeffding inequality [4]. The general version of the inequality deals with sums of independent bounded variables. We restate next the inequality for the specific case which concerns us, namely the sum S_n. Recall that S_n is the random sum

$S_n = X_1 + \cdots + X_n$, where the variables $X_1, \ldots, X_n$ are i.i.d. with distribution given by

$$\forall i \in \{1, \ldots, n\} \qquad P\Big(X_i = \frac{1}{p}\Big) = p\,, \qquad P\Big(X_i = -\frac{1}{1-p}\Big) = 1-p\,.$$

We provide a detailed proof of Hoeffding's inequality in this case. The proof is short (one page) and elementary, but it is quite tricky.

Proposition 8.1. *For any $n \geq 1$ and any $p \in [0,1]$, we have*

$$\forall u > 0 \qquad P\big(|S_n| \geq u\big) \leq 2\exp\Big(-\frac{2p^2(1-p)^2u^2}{n}\Big)\,. \tag{8.1}$$

Proof. Let $u > 0$ be fixed and let $\lambda > 0$ be an additional parameter, to be chosen later. We write

$$\begin{aligned} P\big(S_n \geq u\big) = P\big(\lambda S_n \geq \lambda u\big) &= P\big(\exp\big(\lambda S_n\big) \geq \exp(\lambda u)\big) \\ &\leq \exp(-\lambda u) E\big(\exp\big(\lambda S_n\big)\big)\,, \end{aligned} \tag{8.2}$$

where we have applied the Markov inequality in the last step. Using the fact that the variables X_i are i.i.d., we compute then

$$E\big(\exp\big(\lambda S_n\big)\big) = \prod_{1 \leq i \leq n} E\big(\exp\big(\lambda X_i\big)\big) = \phi(\lambda)^n\,, \tag{8.3}$$

where we have set $\phi(\lambda) = E\big(\exp\big(\lambda X_1\big)\big)$. The explicit expression of $\phi(\lambda)$ is

$$\phi(\lambda) = E\big(\exp\big(\lambda X_1\big)\big) = p\exp\big(\lambda/p\big) + (1-p)\exp\big(-\lambda/(1-p)\big)\,.$$

Using the convexity of the exponential, we see that $\phi(\lambda) \geq 1$ for any $\lambda \in \mathbb{R}$. We take advantage of this inequality to perform a little symmetrization trick: for any $\lambda > 0$, we have

$$\phi(\lambda) \leq \phi(\lambda)\phi(-\lambda) = p^2 + 2p(1-p)\cosh\Big(\frac{\lambda}{p(1-p)}\Big) + (1-p)^2\,. \tag{8.4}$$

The function cosh is even and satisfies $1 \leq \cosh(x) \leq \exp(x^2/2)$, so from the inequality (8.4) we deduce that

$$\phi(\lambda) \leq \cosh\Big(\frac{\lambda}{p(1-p)}\Big) \leq \exp\Big(\frac{\lambda^2}{2p^2(1-p)^2}\Big)\,. \tag{8.5}$$

Substituting this inequality back in (8.3) and (8.2), we obtain

$$P\big(S_n \geq u\big) \leq \exp\Big(-\lambda u + \frac{n\lambda^2}{2p^2(1-p)^2}\Big)\,. \tag{8.6}$$

It remains to choose the optimal value for λ. Obviously, we obtain the best inequality with $\lambda = up^2(1-p)^2/n$, i.e.,

$$\forall u > 0 \qquad P\big(S_n \geq u\big) \leq \exp\Big(-\frac{p^2(1-p)^2u^2}{2n}\Big)\,. \tag{8.7}$$

The same inequality holds for $P\big(S_n \leq -u\big)$ (it suffices to change p into $1-p$ and to apply the previous inequality to $-S_n$). The sum of these two inequalities yields the inequality (8.1) and creates the factor 2 there. To be honest, our proof yields a suboptimal constant in the exponential: the 2 comes in the denominator in formula (8.7). A little refinement in the above chain of inequalities would give the inequality (8.1) with the 2 in the numerator. □

9 An exponential inequality

With the Hoeffding inequality in the hand, we shall complete the program described at the end of section 7 and we shall prove the following general exponential inequality, which does not even require that the boolean function f is monotone.

Proposition 9.1. *For any boolean function f defined on $\{0,1\}^n$, we have*

$$\forall p\in[0,1]\qquad P(f=1)\,\leq\,2\exp\Big(-\frac{1}{2n}\Big(p(1-p)E\big(S_n\,\big|\,f=1\big)\Big)^2\Big)\,. \tag{9.1}$$

Proof. We start with

$$E\big(|S_n|\,\big|\,f=1\big)\;=\;\frac{1}{P(f=1)}\int_{\{f=1\}}|S_n|\,dP\,. \tag{9.2}$$

Using Fubini's theorem, we rewrite the integral in (9.2) as follows:

$$\int_{\{f=1\}}|S_n|\,dP\;=\;\int_\Omega\Big(\int_0^{+\infty}1_{u\leq|S_n|}\,du\Big)f\,dP\;=\;\int_0^{+\infty}P\big(|S_n|\geq u,f=1\big)\,du\,.$$

Let $t>0$ and let us split this integral in two:

$$\begin{aligned}\int_0^{+\infty}P\big(|S_n|\geq u,f=1\big)\,du\;&=\;\int_0^t\cdots+\int_t^{+\infty}\cdots\\&\leq\;tP(f=1)+\int_t^{+\infty}P\big(|S_n|\geq u\big)\,du\,.\end{aligned} \tag{9.3}$$

Reporting (9.3) in (9.2), we have

$$E\big(|S_n|\,\big|\,f=1\big)\;\leq\;\frac{1}{P(f=1)}\int_t^{+\infty}P\big(|S_n|\geq u\big)\,du\,+\,t\,. \tag{9.4}$$

Now S_n is the sum of n i.i.d. centered random variables $X_1,\dots,X_n$ satisfying

$$\forall i\in\{1,\dots,n\}\qquad -\frac{1}{1-p}\;\leq\;X_i\;\leq\;\frac{1}{p}\,.$$

Substituting Hoeffding's inequality (8.1) into (9.4), we obtain

$$E\big(|S_n|\,\big|\,f=1\big)\;\leq\;\frac{2}{P(f=1)}\int_t^{+\infty}\exp\Big(-\frac{2p^2(1-p)^2u^2}{n}\Big)\,du\,+\,t\,. \tag{9.5}$$

We make the change of variable $2p(1-p)u = \sqrt{n}v$, so that (9.5) becomes

$$E\big(|S_n| \,\big|\, f=1\big) \;\leq\; \frac{\sqrt{n}}{p(1-p)P(f=1)} \int_{\frac{2p(1-p)t}{\sqrt{n}}}^{+\infty} \exp\Big(-\frac{v^2}{2}\Big)\,dv \,+\, t\,. \tag{9.6}$$

Recalling that this inequality holds for any $t>0$, we replace t by $\frac{\sqrt{n}t}{2p(1-p)}$ and we get

$$E\big(|S_n| \,\big|\, f=1\big) \;\leq\; \frac{\sqrt{n}}{p(1-p)} \bigg(\frac{1}{P(f=1)} \int_{t}^{+\infty} \exp\Big(-\frac{v^2}{2}\Big)\,dv \,+\, \frac{t}{2}\bigg)\,. \tag{9.7}$$

The righthand quantity admits a unique global minimum at the point $t^* = \sqrt{2\ln(2/P(f=1))}$ and we conclude that

$$E\big(|S_n| \,\big|\, f=1\big) \;\leq\; \frac{1}{p(1-p)}\sqrt{2n\ln\frac{2}{P(f=1)}}\,. \tag{9.8}$$

We rewrite this inequality in exponential form as

$$P(f=1) \;\leq\; 2\exp\Big(-\frac{p^2(1-p)^2}{2n}\Big(E\big(|S_n| \,\big|\, f=1\big)\Big)^2\Big)\,. \tag{9.9}$$

We use finally the fact that

$$\Big|E\big(S_n \,\big|\, f=1\big)\Big| \;\leq\; E\big(|S_n| \,\big|\, f=1\big)\,, \tag{9.10}$$

to obtain the inequality (9.1) of the proposition. $\square$

In the case of a non–decreasing boolean function f, thanks to lemma 7.1, we can rewrite $pE\big(S_n \,\big|\, f=1\big)$ as $E\big(|\mathcal{P}(f)| \,\big|\, f=1\big)$ and we get immediately the following corollary.

Corollary 9.2. *For any non–decreasing boolean function f defined on $\{0,1\}^n$, we have*

$$\forall p\in[0,1] \qquad P(f=1) \;\leq\; 2\exp\Big(-\frac{1}{2n}\Big((1-p)E\big(|\mathcal{P}(f)| \,\big|\, f=1\big)\Big)^2\Big)\,. \tag{9.11}$$

This is a deviations inequality: it tells us that, if the conditional expectation of the cardinality of the pivotal set is larger than $\sqrt{n}$, then it becomes extremely unlikely that the function f takes the value 1.

10 Following Talagrand's footsteps

In fact, the technique employed in the proof of proposition 9.1 to bound the expectation $E\big(|S_n| \,\big|\, A\big)$ has been used routinely in various contexts by Talagrand. It is presented in his most recent book (see lemma 2.3.2 in [11]). Notably, Talagrand used this technique to derive a beautiful inequality on the correlation on increasing sets in his influential work [10]. However, only the very beginning of the proof of the correlation inequality involves this technique, i.e., the first step of the proof of Proposition 2.2 in [10]. This kind of upper bound is such a routine for Talagrand that the detailed proof of our proposition 9.1 corresponds essentially to 6 lines in [10]! So, what is the point of the previous computations? A major difference with [10] is that we deal here with the Bernoulli product measure of parameter p, whereas Talagrand was considering the uniform case $p = 1/2$. So the subgaussian inequality of the uniform case is replaced here by Hoeffding's inequality. Another difference is that the constants in Talagrand's inequality are not explicit (Talagrand did not care about their values), whereas we do provide explicit values here. Let us proceed a bit further along the computations of Talagrand. We implement now the second step of Proposition 2.2 in [10] in our biased context and we see where it leads. We first improve proposition 9.1 to get the following deviations inequality.

Proposition 10.1. *For any boolean function f, we have*

$$P(f=1) \,\leq\, 2\exp\Big(-\frac{1}{2}\sum_{1\leq i\leq n}\Big(p(1-p)E\big(X_i f\,\big|\,f=1\big)\Big)^2\Big)\,. \tag{10.1}$$

Proof. The proof follows the same strategy than the proof of proposition 9.1, the difference being that, instead of S_n, we work with the random variable $\widetilde{S}_n$ given by

$$\widetilde{S}_n \,=\, \sum_{1\leq i\leq n}\alpha_i X_i\,, \tag{10.2}$$

where the α_i's are non–negative real numbers such that $\sum_{1\leq i\leq n}(\alpha_i)^2 = 1$. Let $t > 0$. Formula (9.4) holds for $\widetilde{S}_n$ as well:

$$E\big(|\widetilde{S}_n|\,\big|\,f=1\big) \,\leq\, \frac{1}{P(f=1)}\int_t^{+\infty} P\big(|\widetilde{S}_n|\geq u\big)\,du \,+\, t\,. \tag{10.3}$$

Now $\widetilde{S}_n$ is the sum of n i.i.d. centered random variables which satisfy

$$\forall i\in\{\,1,\dots,n\,\}\qquad -\frac{\alpha_i}{1-p} \,\leq\, \alpha_i X_i \,\leq\, \frac{\alpha_i}{p}\,. \tag{10.4}$$

Applying the more general Hoeffding's inequality [4] to $\widetilde{S}_n$, we get

$$\forall u>0\qquad P\big(|\widetilde{S}_n|\geq u\big) \,\leq\, 2\exp\Big(-2p^2(1-p)^2u^2\Big)\,. \tag{10.5}$$

We perform then exactly the same step as in the proof of proposition 9.1, except that the factor n is replaced by 1. This way we obtain the following inequality:

$$E\big(|\widetilde{S}_n|\,\big|\,f=1\big) \,\leq\, \frac{1}{p(1-p)}\sqrt{2\ln\frac{2}{P(f=1)}}\,. \tag{10.6}$$

We use the inequality $\big|E\big(\widetilde{S}_n \,\big|\, f=1\big)\big| \leq E\big(|\widetilde{S}_n| \,\big|\, f=1\big)$ and we rewrite the inequality (10.6) in exponential form to get

$$P(f=1) \,\leq\, 2\exp\Big(-\frac{1}{2}p^2(1-p)^2\Big(E\big(\widetilde{S}_n \,\big|\, f=1\big)\Big)^2\Big)\,. \tag{10.7}$$

Using the linearity of the expectation, we have

$$E(\widetilde{S}_n f) \,=\, \sum_{1\leq i\leq n} \alpha_i E(X_i f)\,,$$

whence, dividing by $P(f=1)$,

$$E\big(\widetilde{S}_n \,\big|\, f=1\big) \,=\, \sum_{1\leq i\leq n} \alpha_i E\big(X_i f \,\big|\, f=1\big)\,. \tag{10.8}$$

The time has come to choose the α_i's. We take

$$\forall i\in\{\,1,\dots,n\,\}\qquad \alpha_i \,=\, \frac{E\big(X_i f \,\big|\, f=1\big)}{\sqrt{\displaystyle\sum_{1\leq j\leq n}\Big(E\big(X_j f \,\big|\, f=1\big)\Big)^2}}\,.$$

With this choice, the equation (10.8) together with the inequality (10.7) yield the inequality (10.1) of the proposition. □

Naturally, proposition 10.1 is stronger than proposition 9.1: a routine application of the Cauchy–Schwarz inequality shows that the inequality (9.1) is implied by the inequality (10.1). In the specific case where the function f is non–decreasing, we have, thanks to formulas (6.1) and (7.2),

$$E\big(X_i f \,\big|\, f=1\big) \,=\, \frac{1}{p}P\big(i\in\mathcal{P}(f) \,\big|\, f=1\big)\,, \tag{10.9}$$

so that proposition 10.1 yields the following corollary, which is an improvement of corollary 9.2.

Corollary 10.2. *For any non–decreasing boolean function f, we have*

$$P(f=1) \,\leq\, 2\exp\Big(-\frac{1}{2}\sum_{1\leq i\leq n}\Big((1-p)P\big(i\in\mathcal{P}(f) \,\big|\, f=1\big)\Big)^2\Big)\,. \tag{10.10}$$

At this point, we have generalized to the Bernoulli product measure of parameter p an inequality that Talagrand proved in the case where $p=1/2$. Indeed, denoting by $\mu=P_{1/2}$ the uniform measure on $\{0,1\}^n$, Talagrand's inequality can be stated as follows: there exists a universal constant K such that, for any boolean function f, we have

$$\sum_{1\leq i\leq n}\Big(\int fX_i\,d\mu\Big)^2 \,\leq\, K\mu(f=1)^2\ln\frac{e}{\mu(f=1)}\,. \tag{10.11}$$

Clearly, we can rewrite the inequality (10.11) in an exponential form to get an inequality similar to (10.1), albeit with non explicit constants (remember that Talagrand did not care about their values). In fact, the main purpose of Talagrand was to improve this first inequality (stated in Proposition 2.2 of [10]). This first inequality was just a little warm–up before starting the real work!

11 The magic of Talagrand

So, how did Talagrand proceed to go further? Noticing that proposition 10.1 holds for any boolean function, he applied it to the indicator function of the event $\{\, k \in \mathcal{P}(f)\,\}$, where f is a non–decreasing boolean function. This yields

$$P\big(\, k \in \mathcal{P}(f)\,\big) \,\leq\, 2\exp\Big(-\frac{1}{2}\sum_{1\leq i\leq n} \Big(p(1-p)E\big(X_i 1_{\{\, k\in \mathcal{P}(f)\,\}}\,\big|\, k \in \mathcal{P}(f)\,\big)\Big)^2\Big)\,.$$

The variable X_k and the event $\{\, k \in \mathcal{P}(f)\,\}$ are independent, therefore

$$E\big(X_k 1_{\{\, k\in \mathcal{P}(f)\,\}}\big) \,=\, 0\,.$$

Using formula (6.1) in the reverse direction, we can rewrite the expectation for $i \neq k$ as follows:

$$\forall i \neq k \qquad E\big(X_i 1_{\{\, k\in \mathcal{P}(f)\,\}}\big) \,=\, E\big(X_i X_k f\,\big)\,, \tag{11.1}$$

and the previous inequality becomes, in a non exponential form similar to Talagrand's first inequality (10.11),

$$\sum_{\substack{1\leq i\leq n\\ i\neq k}} \Big(E\big(X_i X_k f\,\big)\Big)^2 \,\leq\, \frac{2}{p^2(1-p)^2}\Big(P\big(k \in \mathcal{P}(f)\big)\Big)^2 \ln\Big(\frac{2}{P\big(\, k \in \mathcal{P}(f)\,\big)}\Big)\,.$$

Next, we note $c(p) = 2p^{-2}(1-p)^{-2}$ and we sum this inequality over k to get

$$\sum_{\substack{1\leq i,k\leq n\\ i\neq k}} \Big(E\big(X_i X_k f\,\big)\Big)^2 \,\leq\, c(p) \sum_{1\leq k\leq n}\Big(P\big(k \in \mathcal{P}(f)\big)\Big)^2 \ln\Big(\frac{2}{P\big(\, k \in \mathcal{P}(f)\,\big)}\Big)\,.$$

Finally, in one of those feats of strength for which only he has the secret, Talagrand proved the following much stronger inequality for the uniform measure $\mu = P_{1/2}$: there exists a universal constant K such that, for any non–decreasing boolean function f, we have

$$\sum_{\substack{1\leq i,k\leq n\\ i\neq k}} \Big(\int f X_i X_k \, d\mu\Big)^2 \,\leq\, K \sum_{1\leq k\leq n} \mu\big(k \in \mathcal{P}(f)\big)^2 \ln \frac{K}{\sum_{1\leq k\leq n}\mu\big(k \in \mathcal{P}(f)\big)^2}\,. \tag{11.2}$$

The considerable improvement lies in the presence of the sum inside the logarithm. Talagrand's proof is beautiful and mysterious. It rests on a complicated intermediate result which involves a sort of bootstrap mechanism. The inequality (11.2) attracted the interest of several researchers and led to numerous unexpected developments. For instance, the whole theory of noise sensitivity of Benjamini, Kalai and Schramm [1] was born with a generalization of this inequality. An important question was whether inequality (11.2) holds for the product measure P for $p \neq 1/2$. It is known that the inequalities involving the uniform measure can in principle be generalized to the Bernoulli product measure. Keller [5] has designed a procedure to deduce automatically some inequalities for the biased measure from inequalities involving the uniform measure. Anyway, the natural approach consists in adapting the proofs written for the uniform measure, as we did in this paper for the first inequality of Talagrand. However, this might turn out to be very difficult in some cases. After more than twenty years, Talagrand's inequality has been extended to the biased case by Keller and Kindler [6], with a new proof which is simpler than the original proof of Talagrand. Even more, they managed to prove the inequality for any boolean function, non necessarily non–decreasing. However, upon inspection, it seems that the proof of Talagrand would work also for any boolean function!

References

[1] Itai Benjamini, Gil Kalai, and Oded Schramm, *Noise sensitivity of Boolean functions and applications to percolation*, Inst. Hautes Études Sci. Publ. Math. (1999), no. 90, 5–43. MR 1813223

[2] J. T. Chayes, L. Chayes, Daniel S. Fisher, and T. Spencer, *Finite-size scaling and correlation lengths for disordered systems*, Phys. Rev. Lett. **57** (1986), 2999–3002.

[3] Ehud Friedgut and Gil Kalai, *Every monotone graph property has a sharp threshold*, Proc. Am. Math. Soc. **124** (1996), no. 10, 2993–3002.

[4] Wassily Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. **58** (1963), 13–30.

[5] Nathan Keller, *A simple reduction from a biased measure on the discrete cube to the uniform measure*, Eur. J. Comb. **33** (2012), no. 8, 1943–1957.

[6] Nathan Keller and Guy Kindler, *Quantitative relation between noise sensitivity and influences*, Combinatorica **33** (2013), no. 1, 45–71.

[7] Ryan O'Donnell, *Analysis of Boolean functions*, Cambridge University Press, New York, 2014. MR 3443800

[8] Ryan O'Donnell and Rocco A. Servedio, *Learning monotone decision trees in polynomial time*, SIAM J. Comput. **37** (2007), no. 3, 827–844 (English).

[9] M. Talagrand, *Isoperimetry, logarithmic Sobolev inequalities on the discrete cube, and Margulis' graph connectivity theorem*, Geom. Funct. Anal. **3** (1993), no. 3, 295–314 (English).

[10] Michel Talagrand, *How much are increasing sets positively correlated?*, Combinatorica **16** (1996), no. 2, 243–258.

[11] ______, *Upper and lower bounds for stochastic processes. Decomposition theorems*, Ergeb. Math. Grenzgeb., vol. 60, Cham: Springer, 2021.