

6

The Entropy Method

In Chapter 3 we saw that the Efron–Stein inequality served as a powerful tool for bounding the variance of functions of several independent random variables. In many cases, however, it is reasonable to expect that, as in the case of sums of bounded random variables, the tail probabilities decrease at an exponential speed, a phenomenon the Efron–Stein inequality fails to capture. In Chapter 5 we have seen that logarithmic Sobolev inequalities, together with Herbst’s argument, may be used to derive exponential concentration inequalities. However, the logarithmic Sobolev inequalities presented there are only valid for functions of either Bernoulli or Gaussian random variables and therefore the scope of the concentration inequalities obtained is significantly more limited than that of the Efron–Stein inequality.

The purpose of this chapter is to attempt to generalize the methodology based on logarithmic Sobolev inequalities that allows one to prove exponential concentration bounds that hold for functions of *arbitrary* independent random variables. A way to achieve this is by trying to mimic the procedure that worked for functions of Bernoulli and Gaussian random variables, that is, to start with a logarithmic Sobolev inequality and then, according to Herbst’s trick, apply it to exponential functions of the random variable of interest. Since exact analogs of the Bernoulli and Gaussian logarithmic Sobolev inequalities do not always exist, we need to resort to appropriate modifications. Luckily, the sub-additivity of entropy (see Theorems 4.10 and 4.22) holds in a great generality and indeed, this inequality serves as our starting point. Then, by bounding the right-hand side of the inequality of Theorem 4.10, we obtain an appropriate *modified logarithmic Sobolev inequality* which, in turn, can be used via Herbst’s argument to derive exponential concentration inequalities.

We term the proof method described above the *entropy method*, and the purpose of this chapter is to define its basis and to show some of the simplest powerful concentration bounds one can achieve using this method. In Chapters 11, 12, 14, and 15 we elaborate the entropy method and show various extensions.

As in Chapter 3, we investigate the concentration behavior of a real-valued random variable $Z = f(X_1, \dots, X_n)$ where X_1, \dots, X_n are independent random variables taking values in a measurable space \mathcal{X} and $f : \mathcal{X}^n \rightarrow \mathbb{R}$ is a function.

The main purpose of the entropy method for proving concentration inequalities is to apply the sub-additivity of entropy (Theorems 4.10 and 4.22) for the positive random variable $Y = e^{\lambda Z}$ where λ is a real number. Recall that by the sub-additivity of entropy,

$$\text{Ent}(Y) \leq E \sum_{i=1}^n \text{Ent}^{(i)}(Y)$$

or, equivalently,

$$\begin{aligned} & E[Y \log Y] - (EY) \log(EY) \\ & \leq \sum_{i=1}^n E \left[E^{(i)}[Y \log Y] - \left(E^{(i)}Y \right) \log \left(E^{(i)}Y \right) \right] \end{aligned} \quad (6.1)$$

where $E^{(i)}$ denotes integration with respect to the distribution of X_i only. Then, normalizing by $Ee^{\lambda Z}$ and denoting the logarithmic moment-generating function of $Z - EZ$ by $\psi(\lambda) = \log Ee^{\lambda(Z-EZ)}$, the left-hand side of this inequality becomes

$$\frac{\text{Ent}(e^{\lambda Z})}{Ee^{\lambda Z}} = \lambda \psi'(\lambda) - \psi(\lambda). \quad (6.2)$$

Our strategy is based on using (6.2) the sub-additivity of entropy and then univariate calculus to derive upper bounds for the derivative of $\psi(\lambda)$. By solving the obtained differential inequality, we obtain tail bounds via Chernoff's bounding.

To achieve this in a convenient way, we need some further bounds for the right-hand side of the inequality above. This is the purpose of Section 6.3 in which, relying on the sub-additivity of entropy, we prove some basic results which will serve as our starting point. These results are reminiscent of the classical logarithmic Sobolev inequalities discussed in Chapter 5, where it is shown that concentration inequalities follow from logarithmic Sobolev inequalities by *Herbst's argument*. Here we formalize this argument.

Proposition 6.1 (HERBST'S ARGUMENT) *Let Z be an integrable random variable such that for some $v > 0$, we have, for every $\lambda > 0$,*

$$\frac{\text{Ent}(e^{\lambda Z})}{Ee^{\lambda Z}} \leq \frac{\lambda^2 v}{2}.$$

Then, for every $\lambda > 0$,

$$\log Ee^{\lambda(Z-EZ)} \leq \frac{\lambda^2 v}{2}.$$

Proof The condition of the proposition means, via (6.2), that

$$\lambda \psi'(\lambda) - \psi(\lambda) \leq \frac{\lambda^2 v}{2},$$

or equivalently,

$$\frac{1}{\lambda} \psi'(\lambda) - \frac{1}{\lambda^2} \psi(\lambda) \leq \frac{\nu}{2}.$$

Setting $G(\lambda) = \lambda^{-1} \psi(\lambda)$, we see that the differential inequality becomes $G'(\lambda) \leq \nu/2$. Since $G(\lambda) \rightarrow 0$ as $\lambda \rightarrow 0$, this implies $G(\lambda) \leq \lambda\nu/2$, and the result follows. \square

First, we present in Section 6.1 two simple direct methods to bound the right-hand side of the inequality of the sub-additivity of entropy and use Herbst's argument to conclude. This permits us to derive the celebrated *bounded differences inequality*, a simple prototypical exponential concentration inequality for functions of bounded differences that has found countless applications. We also present a sharper version in which the bounded differences assumption is significantly relaxed.

In Section 6.4 we present the first and simplest application of these modified logarithmic Sobolev inequalities. This first example is surprisingly powerful as it may be used to prove exponential concentration in many interesting cases. We describe some applications. The obtained inequalities reach further than the bounded differences inequality as they are able to handle much more general functions than just those having the bounded-differences property. A simple but useful application for convex Lipschitz functions of independent random variables is presented in Section 6.6.

In Section 6.7 we return to the class of self-bounding functions introduced in Section 3.3 and prove an exponential concentration inequality, thus providing a significant sharpening of Corollary 3.7. The notion of self-bounding function is generalized and further investigated in Section 6.11.

In Sections 6.8, 6.9, and 6.13 we use the entropy method to prove inequalities that may be considered as exponential versions of the Efron–Stein inequality. Various concentration results are shown here under different conditions with the purpose of demonstrating the flexibility of the entropy method.

We close the chapter by proving Janson's celebrated inequality for the lower tail probabilities of random Boolean polynomials. Even though Janson's inequality is not based on the entropy method, its proof shows some similarities with the techniques we use throughout the chapter.

6.1 The Bounded Differences Inequality

As a first illustration of the entropy method, we derive an exponential concentration inequality for functions of bounded differences. Unlike the Bernoulli and Gaussian concentration inequalities of Chapter 5, this inequality is distribution free: apart from independence, nothing else is required from the random variables X_1, \dots, X_n .

Recall that a function $f : \mathcal{X}^n \rightarrow \mathbb{R}$ has the *bounded differences property* if for some nonnegative constants c_1, \dots, c_n ,

$$\sup_{\substack{x_1, \dots, x_n, \\ x'_i \in \mathcal{X}}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

In Chapter 3, as a corollary of the Efron–Stein inequality, we saw that if f has the bounded differences property, then $Z = f(X_1, \dots, X_n)$ satisfies $\text{Var}(Z) \leq (1/4) \sum_{i=1}^n c_i^2$ (see Corollary 3.2). The *bounded differences inequality* shows that such functions satisfy a sub-Gaussian tail inequality in which the role of the variance factor is played by the Efron–Stein upper bound of the variance $v = (1/4) \sum_{i=1}^n c_i^2$.

Theorem 6.2 (BOUNDED DIFFERENCES INEQUALITY) *Assume that the function f satisfies the bounded differences assumption with constants c_1, \dots, c_n and denote*

$$v = \frac{1}{4} \sum_{i=1}^n c_i^2.$$

Let $Z = f(X_1, \dots, X_n)$ where the X_i are independent. Then

$$P\{Z - EZ > t\} \leq e^{-t^2/(2v)}.$$

Note that since the bounded differences assumption is symmetric, Z also satisfies the lower-tail inequality

$$P\{Z - EZ < -t\} \leq e^{-t^2/(2v)}.$$

The proof combines sub-additivity of entropy, Hoeffding’s lemma (Lemma 2.2) and Herbst’s argument. The following way of looking at Hoeffding’s lemma may illuminate the use of the sub-additivity of entropy: if Y is a random variable taking its values in $[a, b]$, then we know from Lemma 2.2 that $\psi''(\lambda) \leq (b-a)^2/4$ for every $\lambda \in \mathbb{R}$, where $\psi(\lambda) = \log Ee^{\lambda(Y-EY)}$. Hence,

$$\lambda\psi'(\lambda) - \psi(\lambda) = \int_0^\lambda \theta\psi''(\theta)d\theta \leq \frac{(b-a)^2\lambda^2}{8},$$

which means that

$$\frac{\text{Ent}(e^{\lambda Y})}{Ee^{\lambda Y}} \leq \frac{(b-a)^2\lambda^2}{8}. \quad (6.3)$$

By Proposition 6.1, this inequality implies Hoeffding’s inequality, that is, $\psi(\lambda) \leq (b-a)^2\lambda^2/8$ for all λ . Thus, (6.3) is a way of rephrasing Hoeffding’s inequality, which is stronger than the usual one.

Proof Recall that by the sub-additivity of entropy (6.1),

$$\text{Ent}(e^{\lambda Z}) \leq E \sum_{i=1}^n \text{Ent}^{(i)}(e^{\lambda Z})$$

where $\text{Ent}^{(i)}$ denotes conditional entropy, given $X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. By the bounded differences assumption, given $X^{(i)}$, Z is a random variable whose range is in an interval of length at most c_i , so by (6.3),

$$\frac{\text{Ent}^{(i)}(e^{\lambda Z})}{E^{(i)} e^{\lambda Z}} \leq \frac{c_i^2 \lambda^2}{8}.$$

Hence, by the sub-additivity of entropy,

$$\text{Ent}(e^{\lambda Z}) \leq E \left[\sum_{i=1}^n \left(\frac{c_i^2 \lambda^2}{8} \right) E^{(i)} e^{\lambda Z} \right] = \sum_{i=1}^n \frac{c_i^2 \lambda^2}{8} E e^{\lambda Z},$$

or equivalently,

$$\frac{\text{Ent}(e^{\lambda Z})}{E e^{\lambda Z}} \leq \frac{\lambda^2 v}{2}.$$

Proposition 6.1 allows us to conclude that

$$\psi(\lambda) = \log E e^{\lambda(Z-EZ)} \leq \frac{\lambda^2 v}{2}.$$

Finally, by Markov's inequality,

$$P\{Z > EZ + t\} \leq e^{\psi(\lambda) - \lambda t} \leq e^{\lambda^2 v / 2 - \lambda t}.$$

Choosing $\lambda = t/v$, the upper bound becomes $e^{-t^2/(2v)}$. □

This extends Corollary 3.2 to an exponential concentration inequality. Thus, the applications of Corollary 3.2 in all examples of functions with bounded differences shown in Section 3.2 (such as bin packing, the length of the longest common subsequence, the L_1 error of the kernel density estimate, etc.) are improved in an essential way without further work.

Next we describe another application which is the simplest example of a concentration inequality for sums of independent vector-valued random variables.

Example 6.3 (A Hoeffding-type inequality in Hilbert space) As an illustration of the power of the bounded differences inequality, we derive a Hoeffding-type inequality for sums of random variables taking values in a Hilbert space. In particular, let X_1, \dots, X_n be independent zero-mean random variables taking values in a separable Hilbert space such that $\|X_i\| \leq c_i/2$ with probability one and denote $v = (1/4) \sum_{i=1}^n c_i^2$. Then, for all $t \geq \sqrt{v}$,

$$P \left\{ \left\| \sum_{i=1}^n X_i \right\| > t \right\} \leq e^{-(t-\sqrt{v})^2/(2v)}.$$

This follows simply by observing that, by the triangle inequality, $Z = \left\| \sum_{i=1}^n X_i \right\|$ satisfies the bounded differences property with constants c_i , and therefore

$$\begin{aligned} P \left\{ \left\| \sum_{i=1}^n X_i \right\| > t \right\} &= P \left\{ \left\| \sum_{i=1}^n X_i \right\| - E \left\| \sum_{i=1}^n X_i \right\| > t - E \left\| \sum_{i=1}^n X_i \right\| \right\} \\ &\leq \exp \left(- \frac{(t - E \left\| \sum_{i=1}^n X_i \right\|)^2}{2v} \right). \end{aligned}$$

The proof is completed by observing that, by independence,

$$E \left\| \sum_{i=1}^n X_i \right\| \leq \sqrt{E \left\| \sum_{i=1}^n X_i \right\|^2} = \sqrt{\sum_{i=1}^n E \|X_i\|^2} \leq \sqrt{v}.$$

The next example illustrates a surprising application in which the bounded differences inequality is applied in a quite unexpected context.

Example 6.4 (SPECTRAL MEASURE OF RANDOM HERMITIAN MATRICES) Let $H = (H_{ij})$ be an $n \times n$ random Hermitian matrix such that the vectors $(H_i)_{1 \leq i \leq n}$ are independent, where $H_i = (H_{ij})_{1 \leq j \leq i}$. Let L_H denote the empirical spectral measure of H (i.e. the probability measure that gives mass r/n to an eigenvalue of H with multiplicity r). Given a bounded function $g : \mathbb{R} \rightarrow \mathbb{R}$ that has total variation $\|g\|_{TV} \leq 1$, we are interested in the concentration of the random variable $Z = \int g dL_H$. Recall that the total variation of a function $g : \mathbb{R} \rightarrow \mathbb{R}$ is defined by

$$\|g\|_{TV} = \sup_{n=1,2,\dots} \sup_{x_1 < \dots < x_n} \sum_{i=1}^{n-1} |f(x_{i+1}) - f(x_i)|.$$

Remarkably, much can be said about Z without imposing any moment assumption on the entries of the matrix. The argument is surprisingly simple. Indeed, for every $x = (x_1, \dots, x_n)$ such that $x_i \in \mathbb{C}^{i-1} \times \mathbb{R}$ for all i , denote by $H(x)$ the Hermitian matrix given by $(H(x))_{ij} = x_{ij}$ for $1 \leq j \leq i \leq n$ and define the function f by

$$f(x) = \int g dL_{H(x)}.$$

The random variable of interest Z is just $f(H_1, \dots, H_n)$ and it remains to establish the bounded differences property for f to get a concentration inequality of Z around its mean. To this end, we apply the following deterministic rank inequality for spectral measures (which relies on the Cauchy interlacing theorem, see Exercises 6.2 and 6.3 below). Let A and B denote Hermitian matrices. If one denotes by F_A and F_B the distribution functions related to the spectral measures L_A and L_B , then

$$\|F_A - F_B\|_\infty \leq \frac{\text{rank}(A - B)}{n}.$$

Integrating by parts (noting that $F_A - F_B$ tends to 0 at $-\infty$ and $+\infty$), one has

$$\left| \int g dL_A - \int g dL_B \right| = \left| \int (F_A - F_B) dg \right| \leq \|F_A - F_B\|_\infty,$$

where the last inequality comes from the fact that the absolute total mass of the Stieljes measure dg equals $\|g\|_{TV} \leq 1$. Combining the two inequalities above, we find that for every x and x' ,

$$|f(x) - f(x')| \leq \frac{\text{rank}(H(x) - H(x'))}{n}.$$

Now if x' differs from x only in the i -th coordinate, the matrix $H(x) - H(x')$ has all zero entries, except maybe for one row and one column which proves that $\text{rank}(H(x) - H(x')) \leq 2$. This shows that f satisfies the bounded differences condition with $c_i = 2/n$ for all i and, therefore, the bounded differences inequality tells us that Z is a sub-Gaussian random variable with variance factor $1/n$. Consequently $P\{|Z - EZ| \geq t\} \leq 2e^{-nt^2/2}$ for all $t > 0$.

6.2 More on Bounded Differences

Next we show a more flexible variant of the bounded differences inequality of Theorem 6.2. It relaxes the bounded differences condition in that differences need not be bounded by “hard” constants c_i but rather by quantities that are allowed to depend on x , as long as the sum of their squares are bounded. More precisely, we say that a function $f : \mathcal{X}^n \rightarrow \mathbb{R}$ has the *x -dependent bounded differences property* if there exists a constant $\nu > 0$ such that for all $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ there exist n functions of $n - 1$ variables $c_1, \dots, c_n : \mathcal{X}^{n-1} \rightarrow [0, \infty)$, such that for $1 \leq i \leq n$,

$$\sup_{\substack{x'_i \in \mathcal{X} \\ x''_i \in \mathcal{X}}} |f(x_1, \dots, x_{i-1}, x''_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i(x^{(i)}),$$

and $(1/4) \sum_{i=1}^n c_i^2(x^{(i)}) \leq \nu$ for all $x \in \mathcal{X}^n$. Here $x^{(i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ stands for the $(n - 1)$ -vector obtained by dropping the i -th component of x .

Clearly, the Efron–Stein inequality still implies that if f has the x -dependent bounded differences property, then $Z = f(X_1, \dots, X_n)$ satisfies $\text{Var}(Z) \leq \nu$. The next sub-Gaussian tail inequality extends Theorem 6.2 to such functions.

Theorem 6.5 *Assume that the function f satisfies the x -dependent bounded differences property with constant ν . Let $Z = f(X_1, \dots, X_n)$ where the X_i are independent. Then for all $t > 0$,*

$$P\{Z - EZ \geq t\} \leq e^{-t^2/(2\nu)}.$$

Proof Since the proof is a simple extension of that for bounded differences inequality, we will only sketch it. By the x -dependent bounded differences assumption, for fixed $X^{(i)}$, conditionally, Z is a random variable whose range is in an interval of length at most $c_i(X^{(i)})$ so by (6.3),

$$\frac{\text{Ent}^{(i)}(e^{\lambda Z})}{E^{(i)}e^{\lambda Z}} \leq \frac{c_i^2(X^{(i)})\lambda^2}{8}$$

and by (6.1),

$$\text{Ent}(e^{\lambda Z}) \leq \sum_{i=1}^n E \left[\left(\frac{c_i^2(X^{(i)})\lambda^2}{8} \right) E^{(i)}e^{\lambda Z} \right] = \sum_{i=1}^n E \left[\left(\frac{c_i^2(X^{(i)})\lambda^2}{8} \right) e^{\lambda Z} \right].$$

Since $(1/4) \sum_{i=1}^n c_i^2(x^{(i)}) \leq v$, this inequality implies that

$$\frac{\text{Ent}(e^{\lambda Z})}{Ee^{\lambda Z}} \leq \frac{\lambda^2 v}{2}$$

and the announced inequality follows by using Herbst's argument as we did at the end of the proof of Theorem 6.2. \square

6.3 Modified Logarithmic Sobolev Inequalities

In this section we present a simple inequality with the purpose of bringing sub-additivity of entropy into a more manageable form, providing a versatile tool for deriving exponential concentration inequalities. This tool will help us prove inequalities under much more flexible conditions than bounded differences. This is achieved by further developing the right-hand side of Eq. (6.1). The obtained inequalities are closely related to the *logarithmic Sobolev inequalities* that we met in Chapter 5, but there we were restricted to functions of Bernoulli or Gaussian random variables.

Our first modified logarithmic Sobolev inequality follows from the sub-additivity and the variational formulation of entropy. Throughout the entire chapter, we consider independent random variables X_1, \dots, X_n taking values in some space \mathcal{X} , a real-valued function $f: \mathcal{X}^n \rightarrow \mathbb{R}$, and the random variable $Z = f(X_1, \dots, X_n)$. As in Section 3.1, we denote $Z_i = f_i(X^{(i)}) = f_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ where $f_i: \mathcal{X}^{n-1} \rightarrow \mathbb{R}$ is an arbitrary function.

Theorem 6.6 (A MODIFIED LOGARITHMIC SOBOLEV INEQUALITY) *Let $\phi(x) = e^x - x - 1$. Then for all $\lambda \in \mathbb{R}$,*

$$\lambda E[Z e^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] \leq \sum_{i=1}^n E[e^{\lambda Z} \phi(-\lambda(Z - Z_i))].$$

Proof We bound each term on the right-hand side of the sub-additivity of entropy (6.1). To do this, recall that by the variational formula of entropy given in Corollary 4.17, for any nonnegative random variable Y and for any $u > 0$,

$$E[Y \log Y] - (EY) \log(EY) \leq E[Y \log Y - Y \log u - (Y - u)].$$

We use this bound conditionally. It implies that if Y_i is a positive function of the random variables $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, then

$$E^{(i)}[Y \log Y] - (E^{(i)}Y) \log(E^{(i)}Y) \leq E^{(i)}[Y(\log Y - \log Y_i) - (Y - Y_i)].$$

Applying the above inequality to the variables $Y = e^{\lambda Z}$ and $Y_i = e^{\lambda Z_i}$, one obtains

$$E^{(i)}[Y \log Y] - (E^{(i)}Y) \log(E^{(i)}Y) \leq E^{(i)}[e^{\lambda Z} \phi(-\lambda(Z - Z_i))]$$

and the proof is completed by (6.1). \square

6.4 Beyond Bounded Differences

Simplicity and generality make the bounded differences inequality attractive and it has become a universal tool as witnessed by its countless applications. However, it is possible to improve this simple inequality in various ways, and the entropy method provides a versatile tool. In this section we first give a simple example that is quite easy to obtain from the modified logarithmic Sobolev inequalities of the previous section yet has numerous interesting applications. Its proof is essentially identical to that of Theorem 5.3 but thanks to the generality of Theorem 6.6, we do not need to restrict ourselves to functions of Bernoulli random variables.

Here we consider a general real-valued function of n independent random variables $Z = f(X_1, \dots, X_n)$ and Z_i denotes an $X^{(i)}$ -measurable random variable defined by $Z_i = \inf_{x_i'} f(X_1, \dots, x_i', \dots, X_n)$.

Theorem 6.7 Assume that Z is such that there exists a constant $v > 0$ such that, almost surely,

$$\sum_{i=1}^n (Z - Z_i)^2 \leq v.$$

Then for all $t > 0$,

$$P\{Z - EZ > t\} \leq e^{-t^2/(2v)}.$$

Proof The result follows easily from the modified logarithmic Sobolev inequality proved in the previous section. Observe that for $x > 0$, $\phi(-x) \leq x^2/2$, and therefore, for all $\lambda > 0$, Theorem 6.6 implies

$$\begin{aligned} \lambda E[Z e^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] &\leq E \left[e^{\lambda Z} \sum_{i=1}^n \frac{\lambda^2}{2} (Z - Z_i)^2 \right] \\ &\leq \frac{\lambda^2 \nu}{2} E e^{\lambda Z}, \end{aligned}$$

where we used the assumption of the theorem. The obtained inequality has the same form as the one we already faced in the proof of Theorem 6.2 and the proof may be finished in an identical way. \square

By replacing f by $-f$ in the theorem above, we see that if Z is such that

$$\sum_{i=1}^n (Z - Z_i)^2 \leq \nu$$

with $Z_i = \sup_{x'_i} f(X_1, \dots, X_{i-1}, x'_i, X_{i+1}, \dots, X_n)$, then one obtains an analogous bound for the lower tail

$$P\{Z < EZ - t\} \leq e^{-t^2/(2\nu)}.$$

As a consequence, if the condition

$$\sum_{i=1}^n (Z - Z_i)^2 \leq \nu$$

is satisfied both for $Z_i = \inf_{x'_i} f(X_1, \dots, X_{i-1}, x'_i, X_{i+1}, \dots, X_n)$ and for $Z_i = \sup_{x'_i} f(X_1, \dots, X_{i-1}, x'_i, X_{i+1}, \dots, X_n)$, one has the two-sided inequality

$$P\{|Z - EZ| > t\} \leq 2e^{-t^2/(2\nu)}.$$

To understand why this inequality is a significant step forward in comparison with Theorem 6.2, simply observe that the conditions of Theorem 6.7 do not require that f should have bounded differences. All they require is that

$$\sup_{\substack{x_1, \dots, x_n, \\ x'_1, \dots, x'_n \in \mathcal{X}}} \sum_{i=1}^n (f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n))^2 \leq \nu.$$

The quantity ν may be interpreted as an upper bound for the Efron–Stein estimate of the variance $\text{Var}(Z)$. Many of the inequalities proved by the entropy method in this chapter have a similar flavor: a sub-Gaussian (or sometimes sub-gamma) tail bound where the role of the variance factor is played by a suitable upper bound based on the Efron–Stein inequality.

Note, however, that if f satisfies the bounded differences assumption (or the x -dependent bounded differences assumption), then Theorems 6.2 and 6.5 provide better constants in the exponent. To illustrate why Theorem 6.7 is an essential improvement, recall the example of the largest eigenvalue of a random symmetric matrix, as described in Example 3.14. For this example Theorem 6.5 fails to provide a meaningful inequality.

Example 6.8 (THE LARGEST EIGENVALUE OF A RANDOM SYMMETRIC MATRIX)

As in Example 3.14, we consider a random symmetric real matrix A with entries X_{ij} , $1 \leq i \leq j \leq n$ where the X_{ij} are independent random variables with absolute value bounded by 1. Let $Z = \lambda_1$ denote the largest eigenvalue of A . In Section 3.14, we have already seen that, almost surely,

$$\sum_{1 \leq i \leq j \leq n} (Z - Z_{ij})^2 \leq 16.$$

We used this estimate and the Efron–Stein inequality to conclude that $\text{Var}(Z) \leq 16$. Using Theorem 6.7, we get, without further work, the sub-Gaussian tail estimate

$$P\{Z > EZ + t\} \leq e^{-t^2/32}.$$

Clearly, the bounded differences inequality is useless here as it is impossible to handle the individual differences $Z - Z'_{ij}$ in a meaningful way, while the sum of their squares is bounded by 16. In Section 8.2 we return to this example, re-prove the exponential tail inequality with a different method and derive a corresponding lower-tail inequality.

6.5 Inequalities for the Lower Tail

In the previous section we showed that the condition

$$\sum_{i=1}^n \left(f(X_1, \dots, X_n) - \inf_{x'_i} f(X_1, \dots, X_{i-1}, x'_i, X_{i+1}, \dots, X_n) \right)^2 \leq v$$

guarantees a sub-Gaussian behavior for the upper tail probabilities $P\{Z > EZ + t\}$. To obtain an analogous bound for the lower tail probabilities $P\{Z < EZ - t\}$, however, one needs a condition of the form

$$\sum_{i=1}^n \left(f(X_1, \dots, X_n) - \sup_{x'_i} f(X_1, \dots, X_{i-1}, x'_i, X_{i+1}, \dots, X_n) \right)^2 \leq v.$$

In many interesting cases, only one of the two quantities can be controlled easily, although one would like to handle both upper and lower tails. This is possible under an additional condition of bounded differences. Here we show a simple version of such a result. Note that it is not quite a sub-Gaussian but rather a sub-Poisson bound. As we point out in subsequent sections, there are some important applications in which sub-Gaussian lower tail

bounds hold. In particular, in Section 6.11 below, we show a general sub-Gaussian lower tail inequality under some additional conditions (see Corollary 6.24). For more discussion and related results, we refer to Chapters 7, 9, and 15.

Theorem 6.9 *Assume that X_1, \dots, X_n are independent and $Z = f(X_1, \dots, X_n)$ is such that there exists a constant $v > 0$ such that, almost surely,*

$$\sum_{i=1}^n (Z_i - Z)^2 \leq v$$

where $Z_i = \sup_{x'_i} f(X_1, \dots, X_{i-1}, x'_i, X_{i+1}, \dots, X_n)$. Assume also that $Z_i - Z \leq 1$ almost surely for all $i = 1, \dots, n$. Then for all $t > 0$,

$$\mathbf{P}\{Z - \mathbf{E}Z > t\} \leq e^{-vh(t/v)} \leq e^{-t^2/(2(v+t/3))}$$

where $h(x) = (1+x)\log(1+x) - x$ for $x > -1$.

Proof Our starting point is, once again, the modified logarithmic Sobolev inequality of Theorem 6.6. In order to bound the right-hand side of that inequality, we need to bound $\mathbf{E}[e^{\lambda Z} \phi(-\lambda(Z - Z_i))]$ with Z_i defined above. The key observation is that $\phi(x)/x^2 = (e^x - x - 1)/x^2$ is an increasing function of x and therefore, for any $\lambda > 0$,

$$\frac{\phi(-\lambda(Z - Z_i))}{\lambda^2(Z - Z_i)^2} \leq \frac{\phi(\lambda)}{\lambda^2}$$

where we used the fact that $Z_i - Z \leq 1$. Thus, by Theorem 6.6, for $\lambda > 0$, we have

$$\begin{aligned} \frac{d}{d\lambda} \left(\frac{1}{\lambda} \log \mathbf{E} e^{\lambda Z} \right) &\leq \frac{1}{\lambda^2 \mathbf{E} e^{\lambda Z}} \sum_{i=1}^n \mathbf{E} [e^{\lambda Z} \phi(-\lambda(Z - Z_i))] \\ &\leq \frac{\phi(\lambda)}{\mathbf{E} e^{\lambda Z}} \mathbf{E} \left[e^{\lambda Z} \sum_{i=1}^n (Z - Z_i)^2 \right] \\ &\leq v \phi(\lambda) \end{aligned}$$

where we used the hypothesis of the theorem. The proof can now be finished as in Theorem 6.7, by integrating the bound above. We thus obtain

$$\mathbf{E} e^{\lambda(Z - \mathbf{E}Z)} \leq e^{\phi(\lambda)v}.$$

The upper bound is just the moment-generating function of a centered Poisson(v) random variable and the tail bounds follow from the calculations shown in Sections 2.2 and 2.7. \square

Of course, by replacing f by $-f$, we get the analog result that if

$$\sum_{i=1}^n \left(f(X_1, \dots, X_n) - \inf_{x'_i} f(X_1, \dots, X_{i-1}, x'_i, X_{i+1}, \dots, X_n) \right)^2 \leq v$$

(i.e. under the same condition as in Theorem 6.7) and also

$$f(X_1, \dots, X_n) - \inf_{x'_i} f(X_1, \dots, X_{i-1}, x'_i, X_{i+1}, \dots, X_n) \leq 1,$$

then for all $0 < t$,

$$P\{Z < EZ - t\} \leq e^{-t^2/(2(v+t/3))}.$$

This bound explains the title of the section.

6.6 Concentration of Convex Lipschitz Functions

In Section 5.4 we proved the fundamental result that any Lipschitz function of a canonical Gaussian vector has sub-Gaussian tails. The entropy method presented in the previous sections allows us to extend this to much more general product distributions, though we need an extra convexity condition on the Lipschitz function. This is analogous to the relationship of the “convex” Poincaré inequality of Section 3.5 to the Gaussian Poincaré inequality presented in Section 3.7. We state the result for functions of n independent random variables taking values in $[0, 1]^n$. However, the same proof extends easily to functions of n independent vector-valued random variables under appropriate Lipschitz and convexity assumptions (see Exercise 6.5).

Recall that $f : [0, 1]^n \rightarrow \mathbb{R}$ is said to be *separately convex* if, for every $i = 1, \dots, n$, it is a convex function of i -th variable if the rest of the variables are fixed.

Theorem 6.10 *Let X_1, \dots, X_n be independent random variables taking values in the interval $[0, 1]$ and let $f : [0, 1]^n \rightarrow \mathbb{R}$ be a separately convex function such that $|f(x) - f(y)| \leq \|x - y\|$ for all $x, y \in [0, 1]^n$. Then $Z = f(X_1, \dots, X_n)$ satisfies, for all $t > 0$,*

$$P\{Z > EZ + t\} \leq e^{-t^2/2}.$$

Proof We may assume without loss of generality that the partial derivatives of f exist. (Otherwise one may approximate f by a smooth function via a standard argument.) Theorem 6.7 suffices to bound the random variable $\sum_{i=1}^n (Z - Z_i)^2$ where $Z_i = \inf_{x'_i} f(X_1, \dots, x'_i, \dots, X_n)$. However, we have already shown in the proof of Theorem 3.17 that

$$\sum_{i=1}^n (Z - Z_i)^2 \leq \|\nabla(f(X))\|^2 \leq 1$$

where at the last step we used the Lipschitz property of f . Therefore, Theorem 6.7 is applicable with $v = 1$. \square

Note that a naive bound using the Lipschitz condition would only give the bound $\sum_{i=1}^n \left(f(X) - f(\bar{X}^{(i)}) \right)^2 \leq 4n$. The convexity assumption provides an immense improvement over this simple bound.

Example 6.11 (THE LARGEST SINGULAR VALUE OF A RANDOM MATRIX) Consider again Example 3.18, that is, let Z be the largest singular value of an $m \times n$ matrix with independent entries X_{ij} ($i = 1, \dots, m, j = 1, \dots, n$) taking values in $[0, 1]$. As we pointed out, Z is a convex function of the X_{ij} , which is also Lipschitz, so Theorem 6.10 implies

$$P\{Z > EZ + t\} \leq e^{-t^2/2}.$$

Here, we assumed that all entries of the matrix A are independent. This assumption may be weakened at the price of obtaining a weaker sub-Gaussian bound. The same argument may be used to establish concentration properties of the largest singular value of a matrix whose columns are independent vectors, but the components of these vectors are not necessarily independent (see Exercise 6.6).

6.7 Exponential Inequalities for Self-Bounding Functions

In this section we revisit self-bounding functions introduced in Section 3.3. Recall that a function $f : \mathcal{X}^n \rightarrow \mathbb{R}$ is said to have the self-bounding property if, for some functions $f_i : \mathcal{X}^{n-1} \rightarrow \mathbb{R}$, for all $x = (x_1, \dots, x_n) \in \mathcal{X}^n$, and for all $i = 1, \dots, n$,

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1$$

and

$$\sum_{i=1}^n \left(f(x) - f_i(x^{(i)}) \right) \leq f(x),$$

where, as usual, $x^{(i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$. If X_1, \dots, X_n are independent random variables taking values in \mathcal{X} and $Z = f(X_1, \dots, X_n)$ for a self-bounding function f , then the Efron–Stein inequality implies $\text{Var}(Z) \leq EZ$. We have seen several interesting examples of self-bounding functions, including various configuration functions, Rademacher averages (Section 3.3), and the combinatorial entropies introduced in Section 4.5. Here, building on the modified logarithmic Sobolev inequality of Theorem 6.6, we obtain exponential concentration bounds for self-bounding functions.

To state the main result of this section, recall the definition of the following two functions that we have already seen in Bennett’s inequality and in the modified logarithmic Sobolev inequalities above:

$$h(u) = (1 + u) \log(1 + u) - u, \quad u \geq -1$$

and

$$\phi(v) = \sup_{u \geq -1} (uv - h(u)) = e^v - v - 1.$$

Theorem 6.12 Assume that Z satisfies the self-bounding property. Then for every $\lambda \in \mathbb{R}$,

$$\log E e^{\lambda(Z-EZ)} \leq \phi(\lambda)EZ.$$

Moreover, for every $t > 0$,

$$P\{Z \geq EZ + t\} \leq \exp\left(-h\left(\frac{t}{EZ}\right)EZ\right)$$

and for every $0 < t \leq EZ$,

$$P\{Z \leq EZ - t\} \leq \exp\left(-h\left(-\frac{t}{EZ}\right)EZ\right).$$

By recalling that $h(u) \geq u^2/(2 + 2u/3)$ for $u \geq 0$ (we have already used this in the proof of Bernstein's inequality; see Exercise 2.8) and observing that $h(u) \geq u^2/2$ for $u \leq 0$, we obtain the following immediate, perhaps more transparent, corollaries: for every $t > 0$,

$$P\{Z \geq EZ + t\} \leq \exp\left(-\frac{t^2}{2EZ + 2t/3}\right)$$

and for every $0 < t \leq EZ$,

$$P\{Z \leq EZ - t\} \leq \exp\left(-\frac{t^2}{2EZ}\right).$$

In these sub-gamma tail bounds the variance factor EZ is the Efron–Stein upper bound of the variance $\text{Var}(Z)$.

Proof We first invoke the modified logarithmic Sobolev inequality (Theorem 6.6).

Since the function ϕ is convex with $\phi(0) = 0$, for any λ and any $u \in [0, 1]$, $\phi(-\lambda u) \leq u\phi(-\lambda)$. Thus, since $Z - Z_i \in [0, 1]$, we have, for every λ , $\phi(-\lambda(Z - Z_i)) \leq (Z - Z_i)\phi(-\lambda)$ and therefore, Theorem 6.6 and the condition $\sum_{i=1}^n (Z - Z_i) \leq Z$ imply that

$$\begin{aligned} \lambda E[Z e^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] &\leq E\left[\phi(-\lambda) e^{\lambda Z} \sum_{i=1}^n (Z - Z_i)\right] \\ &\leq \phi(-\lambda) E[Z e^{\lambda Z}]. \end{aligned}$$

Define, for $\lambda \in \mathbb{R}$, $F(\lambda) = Ee^{\lambda(Z-EZ)}$. Then the inequality above becomes

$$[\lambda - \phi(-\lambda)] \frac{F'(\lambda)}{F(\lambda)} - \log F(\lambda) \leq \phi(-\lambda)EZ,$$

which, writing $G(\lambda) = \log F(\lambda)$, implies

$$(1 - e^{-\lambda}) G'(\lambda) - G(\lambda) \leq \phi(-\lambda)EZ.$$

For $\lambda \geq 0$ this inequality is equivalent to

$$\left(\frac{G(\lambda)}{e^\lambda - 1} \right)' \leq EZ \cdot \left(\frac{-\lambda}{e^\lambda - 1} \right)'.$$

The last differential inequality is straightforward to solve and we obtain, for $\lambda > \lambda_0 > 0$,

$$G(\lambda) \leq (e^\lambda - 1) \left(\frac{G(\lambda_0)}{e^{\lambda_0} - 1} + EZ \left(\frac{\lambda_0}{e^{\lambda_0} - 1} - \frac{\lambda}{e^\lambda - 1} \right) \right).$$

Letting λ_0 tend to 0 and observing that $\lim_{\lambda_0 \rightarrow 0} \lambda_0 / (e^{\lambda_0} - 1) = 1$ and that, by l'Hospital's rule, $\lim_{\lambda_0 \rightarrow 0} G(\lambda_0) / (e^{\lambda_0} - 1) = E[Z - EZ] = 0$, for $\lambda \geq 0$, we get

$$G(\lambda) \leq \phi(\lambda)EZ.$$

Proceeding in a similar way for $\lambda \leq 0$, we obtain the first inequality of the theorem.

On the right-hand side we recognize the moment-generating function of a centered Poisson random variable with parameter EZ . The probability bounds are the corresponding Poisson tail inequalities and are obtained by Chernoff's bounding, as calculated in Section 2.2. \square

Theorem 6.12 provides concentration inequalities for any function satisfying the self-bounding property. In Sections 3.3 and 4.5 several examples of such functions are discussed. Here we mention one more example.

Example 6.13 (MAXIMAL DEGREE IN A RANDOM GRAPH) Consider the Erdős–Rényi $G(n, p)$ model of a random graph. In this model a graph of n vertices is obtained if each one of the $m = \binom{n}{2}$ possible edges is selected, independently, with probability p . The degree of a vertex is the number of edges adjacent to that vertex. Note that the degree of any vertex is a binomial $(n-1, p)$ random variable. Let D denote the maximal degree of any vertex in the graph. Clearly, D is a configuration function, so Theorem 6.12 applies. See Exercise 6.14 for properties of D .

Next we write out explicitly what the theorem implies for combinatorial entropies, defined in Section 4.5.

Theorem 6.14 Assume that $h(x) = \log_b |tr(x)|$ is a combinatorial entropy such that for all $x \in \mathcal{X}^n$ and $i \leq n$,

$$h(x) - h(x^{(i)}) \leq 1.$$

If $X = (X_1, \dots, X_n)$ is a vector of n independent random variables taking values in \mathcal{X} , then the random combinatorial entropy $Z = h(X)$ satisfies

$$P\{Z \geq EZ + t\} \leq \exp\left(-\frac{t^2}{2EZ + 2t/3}\right),$$

and

$$P\{Z \leq EZ - t\} \leq \exp\left(-\frac{t^2}{2EZ}\right).$$

Moreover,

$$E \log_b |tr(X)| \leq \log_b E|tr(X)| \leq \frac{b-1}{\log b} E \log_b |tr(X)|.$$

Note that the left-hand side of the last statement follows from Jensen's inequality, while the right-hand side follows by taking $\lambda = \log b$ in the first inequality of Theorem 6.12. One of the examples of combinatorial entropies, defined in Section 4.5, is VC entropy. For the random VC entropy $T(X)$, we obtain

$$E \log_2 T(X) \leq \log_2 ET(X) \leq (\log_2 e) E \log_2 T(X).$$

This last statement shows that the expected VC entropy $E \log_2 T(X)$ and the annealed VC entropy $\log_2 ET(X)$ are tightly connected, regardless of the class of sets \mathcal{A} and the distribution of the X_i 's.

The same inequality holds for the logarithm of the number of increasing subsequences of a random permutation (see Section 4.5 for the definitions).

6.8 Symmetrized Modified Logarithmic Sobolev Inequalities

One of the most useful forms of the Efron–Stein inequality establishes an upper bound for the variance of $Z = f(X_1, \dots, X_n)$ in terms of the behavior of the random variables $Z - Z'_i$ where $Z'_i = f(X_1, \dots, X'_i, \dots, X_n)$ is obtained by replacing the variable X_i by an independent copy X'_i (see Theorem 3.1). The purpose of the next few sections is the search for exponential concentration inequalities involving the differences $Z - Z'_i$. The following symmetrized modified logarithmic Sobolev inequality is at the basis of such exponential tail inequalities.

Theorem 6.15 (SYMMETRIZED MODIFIED LOGARITHMIC SOBOLEV INEQUALITIES)

For all $\lambda \in \mathbb{R}$,

$$\lambda E[Z e^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] \leq \sum_{i=1}^n E[e^{\lambda Z} \phi(-\lambda(Z - Z'_i))]$$

where $\phi(x) = e^x - x - 1$. Moreover, denoting $\tau(x) = x(e^x - 1)$, for all $\lambda \in \mathbb{R}$,

$$\begin{aligned} \lambda E[Z e^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] &\leq \sum_{i=1}^n E[e^{\lambda Z} \tau(-\lambda(Z - Z'_i)_+)], \\ \lambda E[Z e^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] &\leq \sum_{i=1}^n E[e^{\lambda Z} \tau(\lambda(Z'_i - Z)_+)]. \end{aligned}$$

Proof The first inequality is proved exactly as for Theorem 6.6, simply by noting that, like Z_i, Z'_i is also independent of X_i . To prove the second and third inequalities, write

$$e^{\lambda Z} \phi(-\lambda(Z - Z'_i)) = e^{\lambda Z} \phi(-\lambda(Z - Z'_i)_+) + e^{\lambda Z} \phi(\lambda(Z'_i - Z)_+).$$

By symmetry, the conditional expectation of the second term, conditioned on $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, may be written as

$$\begin{aligned} E^{(i)}[e^{\lambda Z} \phi(\lambda(Z'_i - Z)_+)] &= E^{(i)}[e^{\lambda Z'_i} \phi(\lambda(Z - Z'_i)_+)] \\ &= E^{(i)}[e^{\lambda Z} e^{-\lambda(Z - Z'_i)} \phi(\lambda(Z - Z'_i)_+)]. \end{aligned}$$

Summarizing, we have

$$\begin{aligned} E^{(i)}[e^{\lambda Z} \phi(-\lambda(Z - Z'_i))] &= E^{(i)}\left[\left(\phi(-\lambda(Z - Z'_i)_+) + e^{-\lambda(Z - Z'_i)} \phi(\lambda(Z - Z'_i)_+)\right) e^{\lambda Z}\right]. \end{aligned}$$

The second inequality of the theorem follows simply by noting that $\phi(x) + e^x \phi(-x) = x(e^x - 1) = \tau(x)$. The last inequality follows similarly. \square

6.9 Exponential Efron–Stein Inequalities

Recall that by the Efron–Stein inequality, if $X = (X_1, \dots, X_n)$ is a vector of independent random variables, then the variance of $Z = f(X)$ is bounded as

$$\text{Var}(Z) \leq \frac{1}{2} \sum_{i=1}^n E[(Z - Z'_i)^2].$$

If we denote by $E'[\cdot] = E[\cdot|X]$ expectation with respect to the variables X'_1, \dots, X'_n only, then by introducing the random variables

$$V^+ = \sum_{i=1}^n E'[(Z - Z'_i)_+^2]$$

and

$$V^- = \sum_{i=1}^n E'[(Z - Z'_i)_-^2],$$

the Efron–Stein inequality can be written in either one of the equivalent forms

$$\text{Var}(Z) \leq EV^+ \quad \text{and} \quad \text{Var}(Z) \leq EV^-.$$

The message of the next theorem is that upper bounds for the moment-generating function of the random variables V^+ and V^- may be translated into exponential concentration inequalities for Z . In a sense, these may be understood as exponential versions of the Efron–Stein inequality.

Theorem 6.16 *Let $Z = f(X_1, \dots, X_n)$ be a real-valued function of n independent random variables. Let $\theta, \lambda > 0$ be such that $\theta\lambda < 1$ and $Ee^{\lambda V^+/\theta} < \infty$. Then*

$$\log Ee^{\lambda(Z-EZ)} \leq \frac{\lambda\theta}{1-\lambda\theta} \log Ee^{\lambda V^+/\theta}.$$

Next assume that Z is such that $Z'_i - Z \leq 1$ for every $1 \leq i \leq n$. Then for all $\lambda \in (0, 1/2)$,

$$\log Ee^{\lambda(Z-EZ)} \leq \frac{2\lambda}{1-2\lambda} \log Ee^{\lambda V^-}.$$

Proof The proof of the first statement is based on the second inequality of Theorem 6.15.

To apply this inequality, we need to establish appropriate upper bounds for the quantity $\sum_{i=1}^n E[e^{\lambda Z} \tau(-\lambda(Z - Z'_i)_+)]$ appearing on the right-hand side. By noting that $\tau(-x) \leq x^2$ for all $x \geq 0$, we see that it suffices to bound

$$\sum_{i=1}^n E[e^{\lambda Z} \lambda^2 (Z - Z'_i)_+^2] = \lambda^2 E[V^+ e^{\lambda Z}].$$

In previous applications of the entropy method, our strategy was to relate $E[V^+ e^{\lambda Z}]$ to quantities expressed as a functional of the random variable Z . Here our approach is different: we bound the right-hand side by something that involves the moment-generating function of Z and a functional of V^+ . In order to do this, we “decouple” the random variables $e^{\lambda Z}$ and V^+ .

The duality formula of the entropy given in Theorem 4.13 serves as an ideal tool for this purpose. Recall that the duality formula implies that for any random variable W such that $Ee^W < \infty$,

$$E[(W - \log Ee^W) e^{\lambda Z}] \leq \text{Ent}(e^{\lambda Z}),$$

or equivalently,

$$E[We^{\lambda Z}] \leq E[e^{\lambda Z}] \log E[e^W] + \text{Ent}(e^{\lambda Z}).$$

A natural choice for W is λV^+ but it is advantageous to introduce a free parameter $\theta > 0$ and apply the “decoupling” inequality above with $W = \lambda V^+/\theta$. Now the symmetrized modified logarithmic Sobolev inequality becomes

$$\text{Ent}(e^{\lambda Z}) \leq \lambda\theta (E[e^{\lambda Z}] \log E[e^{\lambda V^+/\theta}] + \text{Ent}(e^{\lambda Z})).$$

Rearranging, and writing $\rho(\lambda) = \log Ee^{\lambda V^+}$ for the logarithmic moment generating function of V^+ , we have

$$(1 - \lambda\theta) \text{Ent}(e^{\lambda Z}) \leq \lambda\theta \rho(\lambda/\theta) Ee^{\lambda Z}$$

which, of course, is only meaningful if $\lambda\theta < 1$. If, as before, we let $G(\lambda) = \log Ee^{\lambda(Z-EZ)}$, then the previous inequality becomes

$$\lambda G'(\lambda) - G(\lambda) \leq \frac{\lambda\theta}{1 - \lambda\theta} \rho(\lambda/\theta).$$

This differential inequality is of the form that we have already encountered and indeed, by Lemma 6.25,

$$G(\lambda) \leq \lambda\theta \int_0^\lambda \frac{\rho(u/\theta)}{u(1 - u\theta)} du.$$

Since $\rho(0) = 0$, the convexity of ρ implies that $\rho(u/\theta)/(u(1 - u\theta))$ is a non-decreasing function and therefore

$$G(\lambda) \leq \frac{\theta\lambda\rho(\lambda/\theta)}{1 - \lambda\theta},$$

and the first inequality of the theorem follows.

To prove the second statement of the theorem, we start with the last inequality of Theorem 6.15 which may be written as

$$\text{Ent}(e^{\lambda Z}) \leq \sum_{i=1}^n E \left[e^{\lambda Z} \lambda^2 (Z'_i - Z)_+^2 \frac{e^{\lambda(Z'_i - Z)_+} - 1}{\lambda(Z'_i - Z)_+} \right].$$

Since $(e^x - 1)/x$ is an increasing function, the conditions $Z'_i - Z \leq 1$ and $\lambda < 1/2$ imply that

$$\text{Ent}(e^{\lambda Z}) \leq \lambda^2 \sum_{i=1}^n E[e^{\lambda Z} (Z'_i - Z)^2_+ 2(e^{1/2} - 1)] \leq 2\lambda^2 E[e^{\lambda Z} V^-].$$

The rest of the proof is the same as for the first inequality of the theorem. \square

6.10 A Modified Logarithmic Sobolev Inequality for the Poisson Distribution

In the previous sections we derived modifications of the Gaussian logarithmic Sobolev inequality that allowed us to prove concentration inequalities for functions of independent random variables of arbitrary distribution. For certain specific distributions, apart from the normal distribution, sharper inequalities are available. Here we show such a “modified logarithmic Sobolev inequality” for Poisson random variables. Recall that X has a Poisson distribution with parameter $\mu > 0$ if X takes nonnegative integer values and for every $k = 0, 1, \dots$, $P\{X = k\} = \mu^k e^{-\mu} / k!$.

If f is a real-valued function defined on the set of nonnegative integers \mathbb{N} , then define the *discrete derivative* of f at $x \in \mathbb{N}$ by $Df(x) = f(x+1) - f(x)$. If one wanted to establish a “discrete” analog of the Gaussian logarithmic Sobolev inequality, one would hope to prove that all functions $f : \mathbb{N} \rightarrow \mathbb{R}$, $\text{Ent}(f^2(X)) \leq \kappa E[|Df(X)|^2]$ for some constant κ . Unfortunately, such a result is not true if X is Poisson because the supremum of $\text{Ent}((f(X))^2) / E[(Df(X))^2]$ is infinite.

However, Theorem 6.15 may be used to prove the following modified logarithmic Sobolev inequalities for Poisson distributions, which is a refinement of the Poisson Poincaré inequality of Exercise 3.21.

Theorem 6.17 (POISSON LOGARITHMIC SOBOLEV INEQUALITY) *Let X be a Poisson random variable and let $f : \mathbb{N} \rightarrow (0, \infty)$. Then*

$$\text{Ent}(f(X)) \leq (EX)E[Df(X)D \log f(X)],$$

and

$$\text{Ent}[f(X)] \leq (EX)E\left[\frac{|Df(X)|^2}{f(X)}\right].$$

The theorem may be proved in a way similar to that with which we proved the Gaussian logarithmic Sobolev inequality: first we establish an inequality for the Bernoulli distribution (see the lemma below) and then use the convergence of the binomial distribution to Poisson. We leave the details of the proof to the reader.

Lemma 6.18 (MODIFIED LOGARITHMIC SOBOLEV INEQUALITIES FOR BERNOULLI DISTRIBUTIONS) *For any function $f : \{0, 1\} \rightarrow (0, \infty)$, let $\nabla f(x) = f(1-x) - f(x)$.*

Let $p \in (0, 1)$, and let X be a Bernoulli random variable with parameter p (i.e., $P\{X = 1\} = 1 - P\{X = 0\} = p$). Then

$$\text{Ent}(f(X)) \leq p(1-p)E[\nabla f(X)\nabla \log f(X)]$$

and

$$\text{Ent}(f(X)) \leq p(1-p)E\left[\frac{|\nabla f(X)|^2}{f(X)}\right].$$

Proof We only prove the first inequality. The proof of the second is left as an exercise. Let X' be an independent copy of X . Let $q = 1 - p$. By the first inequality of Theorem 6.15, taking $\lambda = 1$ and $Z = \log f(X)$,

$$\begin{aligned} \text{Ent}(f(X)) &\leq E[f(X)\phi(\log(f(X')/f(X)))] \\ &= E[f(X') - f(X) - f(X)(\log(f(X')) - \log(f(X)))] \\ &= pq[-f(1)(\log(f(0)) - \log(f(1)))] + pq[-f(0)(\log(f(1)) - \log(f(0)))] \\ &= pqE[\nabla f(X)\nabla \log f(X)]. \quad \square \end{aligned}$$

It is easy to deduce from Theorem 6.17 that the square root of a Poisson random variable X satisfies

$$\log Ee^{\lambda(\sqrt{X} - E\sqrt{X})} \leq \nu(e^\lambda - 1)$$

where $\nu = (EX)E[1/(4X + 1)]$. This represents an improvement over what can be obtained from Theorem 6.29 below (see Exercise 6.12).

6.11 Weakly Self-Bounding Functions

Self-bounding functions, discussed in Section 6.7, appear naturally in numerous applications including configuration functions and combinatorial entropies. Theorem 6.12 is quite satisfactory as it cannot be improved in this generality and its proof is rather simple. However, one often faces functions that only satisfy slightly weaker conditions. A prime example, presented in Chapter 7, is the squared “convex distance.” In order to handle this example, as well as various other naturally emerging cases, we generalize the definition of self-bounding functions in two different ways. This section is dedicated to inequalities for such generalized self-bounding functions. The proofs are variants of the entropy method, all based on the modified logarithmic Sobolev inequality of Theorem 6.6. However, the resulting differential inequality for the moment-generating function is not always as easy to solve as in Theorems 6.7 and 6.12, and most of our effort is devoted to the solution of these differential inequalities.

We distinguish two notions of generalized self-bounding functions. In both of the following definitions, a and b are nonnegative constants.

A nonnegative function $f : \mathcal{X}^n \rightarrow [0, \infty)$ is called *weakly (a, b) -self-bounding* if there exist functions $f_i : \mathcal{X}^{n-1} \rightarrow [0, \infty)$ such that for all $x \in \mathcal{X}^n$,

$$\sum_{i=1}^n \left(f(x) - f_i(x^{(i)}) \right)^2 \leq af(x) + b.$$

On the other hand, we say that a function $f : \mathcal{X}^n \rightarrow [0, \infty)$ is *strongly (a, b) -self-bounding* if there exist functions $f_i : \mathcal{X}^{n-1} \rightarrow [0, \infty)$ such that for all $i = 1, \dots, n$ and all $x \in \mathcal{X}^n$,

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1,$$

and

$$\sum_{i=1}^n \left(f(x) - f_i(x^{(i)}) \right) \leq af(x) + b.$$

Clearly, a self-bounding function is strongly $(1, 0)$ -self-bounding and every strongly (a, b) -self-bounding function is weakly (a, b) -self-bounding. In both cases, the Efron–Stein inequality implies $\text{Var}(Z) \leq aEZ + b$. Indeed, this quantity appears as a variance factor in the exponential bounds established below.

We present three inequalities. The simplest is an inequality for the upper tails of weakly (a, b) -self-bounding functions.

Theorem 6.19 *Let $X = (X_1, \dots, X_n)$ be a vector of independent random variables, each taking values in a measurable set \mathcal{X} , let $a, b \geq 0$ and let $f : \mathcal{X}^n \rightarrow [0, \infty)$ be a weakly (a, b) -self-bounding function. Let $Z = f(X)$. If, in addition, $f_i(x^{(i)}) \leq f(x)$ for all $i \leq n$ and $x \in \mathcal{X}^n$, then for all $0 \leq \lambda \leq 2/a$,*

$$\log \mathbb{E} e^{\lambda(Z-EZ)} \leq \frac{(aEZ + b)\lambda^2}{2(1 - a\lambda/2)}$$

and for all $t > 0$,

$$\mathbb{P}\{Z \geq EZ + t\} \leq \exp\left(-\frac{t^2}{2(aEZ + b + at/2)}\right).$$

Proof Once again, our starting point is the modified logarithmic Sobolev inequality. Write $Z_i = f_i(X^{(i)})$. The main observation is that for $x \geq 0$, $\phi(-x) \leq x^2/2$. Since $Z - Z_i \geq 0$, for $\lambda > 0$, by further bounding the right-hand side of the inequality of Theorem 6.6, we obtain

$$\begin{aligned}\lambda E[Ze^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] &\leq \frac{\lambda^2}{2} E\left[e^{\lambda Z} \sum_{i=1}^n (Z - Z_i)^2\right] \\ &\leq \frac{\lambda^2}{2} E[(aZ + b)e^{\lambda Z}]\end{aligned}$$

where we use the assumption that f is weakly (a, b) -self-bounding. Introducing $G(\lambda) = \log Ee^{\lambda(Z-EZ)}$, the inequality obtained above may be re-arranged to read

$$\left(\frac{1}{\lambda} - \frac{a}{2}\right) G'(\lambda) - \frac{G(\lambda)}{\lambda^2} \leq \frac{\nu}{2}$$

where we write $\nu = aEZ + b$.

To finish the proof, simply observe that the left-hand side is just the derivative of the function $(1/\lambda - a/2) G(\lambda)$. Using the fact that $G(0) = G'(0) = 0$, and that $G'(\lambda) \geq 0$ for $\lambda > 0$, integrating this differential inequality leads to

$$G(\lambda) \leq \frac{\nu\lambda^2}{2(1 - a\lambda/2)} \quad \text{for all } \lambda \in [0, 2/a).$$

This shows that $Z - EZ$ is a sub-gamma random variable with variance factor $\nu = aEZ + b$ and scale parameter $a/2$. The tail bound follows from the calculations shown in Section 2.4. \square

The next theorem provides lower tail inequalities for weakly (a, b) -self-bounding functions. This will become essential for proving the convex distance inequality in Section 7.4.

Theorem 6.20 *Let $X = (X_1, \dots, X_n)$ be a vector of independent random variables, each taking values in a measurable set \mathcal{X} , let $a, b \geq 0$ and let $f : \mathcal{X}^n \rightarrow [0, \infty)$ be a weakly (a, b) -self-bounding function. Let $Z = f(X)$ and define $c = (3a - 1)/6$. If, in addition, $f(x) - f_i(x^{(i)}) \leq 1$ for each $i \leq n$ and $x \in \mathcal{X}^n$, then for $0 < t \leq EZ$,*

$$P\{Z \leq EZ - t\} \leq \exp\left(-\frac{t^2}{2(aEZ + b + c \cdot t)}\right).$$

Note that if $a \geq 1/3$, then the left tail is sub-Gaussian with variance proxy $aEZ + b$, while for $a < 1/3$ we will only obtain a sub-gamma tail bound.

The proof of this theorem is shown below, together with the proof of the following upper tail inequality for strongly (a, b) -self-bounding functions.

Theorem 6.21 Let $X = (X_1, \dots, X_n)$ be a vector of independent random variables, each taking values in a measurable set \mathcal{X} , let $a, b \geq 0$ and let $f : \mathcal{X}^n \rightarrow [0, \infty)$ be a strongly (a, b) -self-bounding function. Let $Z = f(X)$ and define $c = (3a - 1)/6$. Then for all $\lambda \geq 0$,

$$\log E e^{\lambda(Z-EZ)} \leq \frac{(aEZ + b)\lambda^2}{2(1 - c_+\lambda)}$$

and for all $t > 0$,

$$P\{Z \geq EZ + t\} \leq \exp\left(-\frac{t^2}{2(aEZ + b + c_+t)}\right).$$

In this upper tail bound we observe a similar phenomenon as in Theorem 6.20 but with a different sign. If $a \leq 1/3$, then the upper tail of a strongly (a, b) -self-bounding function is purely sub-Gaussian.

Our starting point is once again the modified logarithmic Sobolev inequality of Theorem 6.6.

If $\lambda \geq 0$ and f is strongly (a, b) -self-bounding, then, using $Z - Z_i \leq 1$ and the fact that for all $x \in [0, 1]$, $\phi(-\lambda x) \leq x\phi(-\lambda)$,

$$\begin{aligned} \lambda E[Z e^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] &\leq \phi(-\lambda) E\left[e^{\lambda Z} \sum_{i=1}^n (Z - Z_i)\right] \\ &\leq \phi(-\lambda) E[(aZ + b) e^{\lambda Z}]. \end{aligned}$$

For any $\lambda \in \mathbb{R}$, define $G(\lambda) = \log E e^{(\lambda Z - EZ)}$. Then the previous inequality may be written as the differential inequality

$$[\lambda - a\phi(-\lambda)] G'(\lambda) - G(\lambda) \leq \nu\phi(-\lambda), \quad (6.4)$$

where $\nu = aEZ + b$.

On the other hand, if $\lambda \leq 0$ and f is weakly (a, b) -self-bounding, then since $\phi(x)/x^2$ is nondecreasing over \mathbb{R}^+ , $\phi(-\lambda(Z - Z_i)) \leq \phi(-\lambda)(Z - Z_i)^2$ so

$$\begin{aligned} \lambda E[Z e^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] &\leq \phi(-\lambda) E\left[e^{\lambda Z} \sum_{i=1}^n (Z - Z_i)^2\right] \\ &\leq \phi(-\lambda) E[(aZ + b) e^{\lambda Z}]. \end{aligned}$$

This again leads to the differential inequality (6.4) but this time for $\lambda \leq 0$.

When $a = 1$, this differential inequality can be solved exactly as we saw it in the proof of Theorem 6.12, and one obtains the sub-Poissonian inequality

$$G(\lambda) \leq \nu\phi(\lambda).$$

However, when $a \neq 1$, it is not obvious what kind of bounds for G should be expected. If $a > 1$, then $\lambda - a\phi(-\lambda)$ becomes negative when λ is large enough. Since both $G'(\lambda)$ and $G(\lambda)$ are nonnegative when λ is nonnegative, (6.4) becomes trivial for large values of λ . Hence, at least when $a > 1$, there is no hope to derive Poissonian bounds from (6.4) for positive values of λ (i.e. for the upper tail).

The following lemma, proved in Section 6.12 below, is the key to the proof of both Theorems 6.20 and 6.21. It shows that if f satisfies a self-bounding property, then on the relevant interval, the logarithmic moment-generating function of $Z - EZ$ is upper bounded by v times a function G_γ defined by

$$G_\gamma(\lambda) = \frac{\lambda^2}{2(1 - \gamma\lambda)} \quad \text{for every } \lambda \text{ such that } \gamma\lambda < 1$$

where $\gamma \in \mathbb{R}$ is a real-valued parameter. In the lemma below we mean $c_+^{-1} = \infty$ (resp. $c_-^{-1} = \infty$) when $c_+ = 0$ (resp. $c_- = 0$).

Lemma 6.22 *Let $a, v > 0$ and let G be a solution of the differential inequality*

$$[\lambda - a\phi(-\lambda)] H'(\lambda) - H(\lambda) \leq v\phi(-\lambda).$$

Define $c = (a - 1/3)/2$. Then, for every $\lambda \in (0, c_+^{-1})$

$$G(\lambda) \leq vG_{c_+}(\lambda)$$

and for every $\lambda \in (-\theta, 0)$

$$G(\lambda) \leq vG_{-c_-}(\lambda)$$

where $\theta = c_-^{-1}(1 - \sqrt{1 - 6c_-})$ if $c_- > 0$ and $\theta = a^{-1}$ whenever $c_- = 0$.

The proof is given in the next section. Equipped with this lemma, it is now easy to obtain Theorems 6.20 and 6.21.

Proof of Theorem 6.20. We have to check that the condition $\lambda > -\theta$ is harmless. Since $\theta < c_-^{-1}$, by continuity, for every $t > 0$,

$$\sup_{u \in (0, \theta)} \left(tu - \frac{u^2 v}{2(1 - c_- u)} \right) = \sup_{u \in (0, \theta]} \left(tu - \frac{u^2 v}{2(1 - c_- u)} \right).$$

Note that we are only interested in values of t that are smaller than $EZ \leq v/a$. Now the supremum of

$$tu - \frac{u^2 v}{2(1 - c_- u)}$$

as a function of $u \in (0, c_-^{-1})$ is achieved either at $u_t = t/v$ (if $c_- = 0$) or at $u_t = c_-^{-1} \left(1 - (1 + (2tc_-/v))^{-1/2} \right)$ (if $c_- > 0$).

It is time to take into account the restriction $t \leq v/a$. In the first case, when $u_t = t/v$, it implies that $u_t \leq a^{-1} = \theta$, while in the second case, since $a = (1 - 6c_-)/3$ it implies that $1 + (2tc_-/v) \leq (1 - 6c_-)^{-1}$ and therefore $u_t \leq c_-^{-1} (1 - \sqrt{1 - 6c_-}) = \theta$. In both cases $u_t \leq \theta$ which means that for every $t \leq v/a$

$$\sup_{u \in (0, \theta]} \left(tu - \frac{u^2 v}{2(1 - c_- u)} \right) = \sup_{u \in (0, c_-^{-1})} \left(tu - \frac{u^2 v}{2(1 - c_- u)} \right)$$

and the result follows. \square

Proof of Theorem 6.21. The upper-tail inequality for strongly (a, b) -self-bounding functions follows from Lemma 6.22 and Markov's inequality by routine calculations, exactly as in the proof of Bernstein's inequality when $c_+ > 0$, and it is straightforward when $c_+ = 0$. \square

Example 6.23 (THE SQUARE OF A REGULAR FUNCTION) To illustrate the use of the results of this section, consider a function $g : \mathcal{X}^n \rightarrow \mathbb{R}$ and assume that there exists a constant $v > 0$ and that there are measurable functions $g_i : \mathcal{X}^{n-1} \rightarrow \mathbb{R}$ such that for all $x \in \mathcal{X}^n$, $g(x) \geq g_i(x^{(i)})$,

$$\sum_{i=1}^n \left(g(x) - g(x^{(i)}) \right)^2 \leq v.$$

We term such a function *v-regular*. If $X = (X_1, \dots, X_n) \in \mathcal{X}^n$ is a vector of independent \mathcal{X} -valued random variables, then by Theorem 6.7, for all $t > 0$,

$$P \{ g(X) \geq E g(X) + t \} \leq e^{-t^2/(2v)}.$$

Even though Theorem 6.7 provides an exponential inequality for the lower tail, it fails to give an analogous sub-Gaussian bound for $P \{ g(X) \leq E g(X) - t \}$. Here we show how Theorem 6.20 may be used to derive lower-tail bounds under an additional bounded-differences condition for the *square* of g .

Corollary 6.24 Let $g : \mathcal{X}^n \rightarrow \mathbb{R}$ be a v -regular function such that for all $x \in \mathcal{X}^n$ and $i = 1, \dots, n$, $g(x)^2 - g_i(x^{(i)})^2 \leq 1$. Then for all $t \geq 0$,

$$P \{ g(X)^2 \leq E [g(X)^2] - t \} \leq \exp \left(\frac{-t^2}{8vE [g(X)^2] + t(4v - 1/3)} \right).$$

In particular, if g is nonnegative and $v \geq 1/12$, then for all $0 \leq t \leq E g(X)$,

$$P \{ g(X) \leq E g(X) - t \} \leq e^{-t^2/(8v)}.$$

Proof Introduce $f(x) = g(x)^2$ and $f_i(x^{(i)}) = g_i(x^{(i)})^2$. Then

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1.$$

Moreover,

$$\begin{aligned} \sum_{i=1}^n \left(f(x) - f_i(x^{(i)}) \right)^2 &= \sum_{i=1}^n \left(g(x) - g_i(x^{(i)}) \right)^2 \left(g(x) + g_i(x^{(i)}) \right)^2 \\ &= 4g(x)^2 \sum_{i=1}^n \left(g(x) - g_i(x^{(i)}) \right)^2 \\ &\leq 4vf(x) \end{aligned}$$

and therefore f is weakly $(4v, 0)$ -self-bounding. This means that Theorem 6.20 is applicable and this is how the first inequality is obtained.

The second inequality follows from the first by noting that

$$\begin{aligned} P\{g(X) \leq Eg(X) - t\} &\leq P\left\{g(X)\sqrt{E[g(X)^2]} \leq E[g(X)^2] - t\sqrt{E[g(X)^2]}\right\} \\ &\leq P\left\{g(X)^2 \leq E[g(X)^2] - t\sqrt{E[g(X)^2]}\right\}, \end{aligned}$$

and now the first inequality may be applied. \square

For a more concrete class of applications, consider a nonnegative separately convex Lipschitz function g defined on $[0, 1]^n$. If $X = (X_1, \dots, X_n)$ are independent random variables taking values in $[0, 1]$, then by Theorem 6.10,

$$P\{g(X) - Eg(X) > t\} \leq e^{-t^2/2}.$$

Now we may derive a lower-tail inequality for g , under the additional assumption that g^2 takes its values in an interval of length 1. Indeed, without loss of generality we may assume that g is differentiable on $[0, 1]^n$ because otherwise one may approximate g by a smooth function in a standard way. Then, denoting

$$g_i(x^{(i)}) = \inf_{x'_i \in \mathcal{X}} g(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n),$$

by separate convexity,

$$g(x) - g_i(x^{(i)}) \leq \left| \frac{\partial g}{\partial x_i}(x) \right|.$$

Thus, for every $x \in [0, 1]^n$,

$$\sum_{i=1}^n \left(g(x) - g_i(x^{(i)}) \right)^2 \leq 1.$$

We return to the this problem in Section 7.5 where we will be able to drop the extra assumptions on the range of g^2 .

For a concrete example, consider the ℓ_p norm $\|x\|_p$ for some $p \geq 2$. Then $g(x) = \|x\|_p$ is convex and Lipschitz, so we obtain that if $X = (X_1, \dots, X_n)$ is a vector of independent random variables taking values in an interval of length 1, then for all $t > 0$,

$$P \left\{ \|X\|_p^2 \leq E\|X\|_p^2 - t \right\} \leq e^{-t^2/(8E\|X\|_p^2)}$$

and

$$P \left\{ \|X\|_p \leq E\|X\|_p - t \right\} \leq e^{-t^2/8}.$$

6.12 Proof of Lemma 6.22

The key to the success of the entropy method is that the differential inequalities for the logarithmic moment-generating function of Z can be solved in many interesting cases. The cases considered so far were all easily solvable by lucky coincidences. Here we try to extract the essence of these circumstances and generalize them so that a large family of solvable differential inequalities can be dealt with. The next lemma establishes some simple sufficient conditions. Then Lemma 6.26 will allow us to use Lemma 6.25 to cope with more difficult cases, and this will lead to the proof of Lemma 6.22.

Lemma 6.25 *Let f be a nondecreasing continuously differentiable function on some interval I containing 0 such that $f(0) = 0$, $f'(0) > 0$ and $f(x) \neq 0$ for every $x \neq 0$. Let g be a continuous function on I and consider an infinitely many times differentiable function G on I such that $G(0) = G'(0) = 0$ and for every $\lambda \in I$,*

$$f(\lambda)G'(\lambda) - f'(\lambda)G(\lambda) \leq f^2(\lambda)g(\lambda).$$

Then, for every $\lambda \in I$, $G(\lambda) \leq f(\lambda) \int_0^\lambda g(x)dx$.

Note the special case when $f(\lambda) = \lambda$, and $g(\lambda) = L^2/2$ is the differential inequality obtained, for example, in Theorems 5.3 and 6.7 and is used to obtain sub-Gaussian concentration inequalities. If we choose $f(\lambda) = e^\lambda - 1$ and $g(\lambda) = -d(\lambda/e^\lambda - 1)/d\lambda$, we recover the differential inequality seen in the proof of Theorem 6.12.

Proof Define $\rho(\lambda) = G(\lambda)/f(\lambda)$ for every $\lambda \neq 0$ and $\rho(0) = 0$. Using the assumptions on G and f , we see that ρ is continuously differentiable on I with

$$\rho'(\lambda) = \frac{f(\lambda)G'(\lambda) - f'(\lambda)G(\lambda)}{f^2(\lambda)} \quad \text{for } \lambda \neq 0 \quad \text{and} \quad \rho'(0) = \frac{G''(0)}{2f'(0)}.$$

Hence $f(\lambda)G'(\lambda) - f'(\lambda)G(\lambda) \leq f^2(\lambda)g(\lambda)$ implies that

$$\rho'(\lambda) \leq g(\lambda)$$

and therefore that the function $\Delta(\lambda) = \int_0^\lambda g(x)dx - \rho(\lambda)$ is nondecreasing on I . Since $\Delta(0) = 0$, Δ and f have the same sign on I , which means that $\Delta(\lambda)f(\lambda) \geq 0$ for $\lambda \in I$ and the result follows. \square

Except when $a = 1$, the differential inequality (6.4) cannot be solved exactly. A roundabout is provided by the following lemma that compares the solutions of a possibly difficult differential inequality with solutions of a differential equation.

Lemma 6.26 *Let I be an interval containing 0 and let ρ be continuous on I . Let $a \geq 0$ and $v > 0$. Let $H : I \rightarrow \mathbb{R}$, be an infinitely many times differentiable function satisfying*

$$\lambda H'(\lambda) - H(\lambda) \leq \rho(\lambda) (aH'(\lambda) + v)$$

with

$$aH'(\lambda) + v > 0 \quad \text{for every } \lambda \in I \text{ and } H'(0) = H(0) = 0.$$

Let $\rho_0 : I \rightarrow \mathbb{R}$ be a function. Assume that $G_0 : I \rightarrow \mathbb{R}$ is infinitely many times differentiable such that for every $\lambda \in I$,

$$aG_0'(\lambda) + 1 > 0 \quad \text{and } G_0'(0) = G_0(0) = 0 \text{ and } G_0''(0) = 1.$$

Assume also that G_0 solves the differential equation

$$\lambda G_0'(\lambda) - G_0(\lambda) = \rho_0(\lambda) (aG_0'(\lambda) + 1).$$

If $\rho(\lambda) \leq \rho_0(\lambda)$ for every $\lambda \in I$, then $H \leq vG_0$.

Proof Let $I, \rho, a, v, H, G_0, \rho_0$ be defined as in the statement of the lemma. Combining the assumptions on H, ρ_0, ρ and G_0 ,

$$\lambda H'(\lambda) - H(\lambda) \leq \frac{(\lambda G_0'(\lambda) - G_0(\lambda)) (aH'(\lambda) + v)}{aG_0'(\lambda) + 1}$$

for every $\lambda \in I$, or equivalently,

$$(\lambda + aG_0(\lambda)) H'(\lambda) - (1 + aG_0'(\lambda)) H(\lambda) \leq v (\lambda G_0'(\lambda) - G_0(\lambda)).$$

Setting $f(\lambda) = \lambda + aG_0(\lambda)$ for every $\lambda \in I$ and defining $g : I \rightarrow \mathbb{R}$ by

$$g(\lambda) = \frac{v (\lambda G_0'(\lambda) - G_0(\lambda))}{(\lambda + aG_0(\lambda))^2} \quad \text{if } \lambda \neq 0 \text{ and } g(0) = \frac{v}{2},$$

our assumptions on G_0 imply that g is continuous on the whole interval I so that we may apply Lemma 6.25. Hence, for every $\lambda \in I$

$$H(\lambda) \leq f(\lambda) \int_0^\lambda g(x)dx = vf(\lambda) \int_0^\lambda \left(\frac{G_0(x)}{f(x)} \right)' dx$$

and the conclusion follows since $\lim_{x \rightarrow 0} G_0(x)/f(x) = 0$. \square

Observe that the differential inequality in the statement of Lemma 6.22 has the same form as the inequalities considered in Lemma 6.26 where ϕ replaces ρ . Note also that for any $\gamma \geq 0$,

$$2G_\gamma(\lambda) = \frac{\lambda^2}{1 - \gamma\lambda}$$

solves the differential inequality

$$\lambda H'(\lambda) - H(\lambda) = \lambda^2(\gamma H'(\lambda) + 1). \quad (6.5)$$

So by choosing $\gamma = a$ and recalling that for $\lambda \geq 0$, $\phi(-\lambda) \leq \lambda^2/2$, it follows immediately from Lemma 6.26, that

$$G(\lambda) \leq \frac{\lambda^2 v}{2(1 - a\lambda)} \quad \text{for } \lambda \in (0, 1/a).$$

As G is the logarithmic moment-generating function of $Z - EZ$, this can be used to derive a Bernstein-type inequality for the left tail of Z . However, the obtained constants are not optimal, so proving that Lemma 6.22 requires some more care.

Proof of Lemma 6.22. The function $2G_\gamma$ may be the unique solution of equation (6.5) but this is not the only equation for which G_γ is the solution. Define

$$\rho_\gamma(\lambda) = \frac{\lambda G'_\gamma(\lambda) - G_\gamma(\lambda)}{1 + aG'_\gamma(\lambda)}.$$

Then, on some interval I , G_γ is the solution of the differential equation

$$\lambda H'(\lambda) - H(\lambda) = \rho_\gamma(\lambda)(1 + aH'(\lambda)),$$

provided $1 + aG'_\gamma$ remains positive on I .

Thus, we have to look for the smallest $\gamma \geq 0$ such that, on the relevant interval I (with $0 \in I$), we have both $\phi(-\lambda) \leq \rho_\gamma(\lambda)$ and $1 + aG'_\gamma(\lambda) > 0$ for $\lambda \in I$.

Introduce

$$\begin{aligned} D_\gamma(\lambda) &= (1 - \gamma\lambda)^2(1 + aG'_\gamma(\lambda)) = (1 - \gamma\lambda)^2 + a\lambda \left(1 - \frac{\gamma\lambda}{2}\right) \\ &= 1 + 2(a/2 - \gamma)\lambda - \gamma(a/2 - \gamma)\lambda^2. \end{aligned}$$

Observe that $\rho_\gamma(\lambda) = \lambda^2/(2D_\gamma(\lambda))$.

For any interval I , $1 + aG'_\gamma(\lambda) > 0$ for $\lambda \in I$ holds if and only if $D_\gamma(\lambda) > 0$ for $\lambda \in I$. Hence, if $D_\gamma(\lambda) > 0$ and $\phi(-\lambda) \leq \rho_\gamma(\lambda)$, then it follows from Lemma 6.26 that for every $\lambda \in I$, we have $G(\lambda) \leq vG_\gamma(\lambda)$.

We first deal with intervals of the form $I = [0, c_+^{-1})$ (with $c_+^{-1} = \infty$ when $c_+ = 0$). If $a \leq 1/3$, that is, $c_+ = 0$, $D_{c_+}(\lambda) = 1 + a\lambda > 0$ and $\rho_{c_+}(\lambda) \geq \lambda^2/(2(1 + \lambda/3)) \geq \phi(-\lambda)$ for $\lambda \in I = [0, +\infty)$.

If $a > 1/3$, then $D_{c_+}(\lambda) = 1 + \lambda/3 - c_+\lambda^2/6$ satisfies $0 < 1 + \lambda/6 \leq D_{c_+}(\lambda) \leq 1 + \lambda/3$ on an interval I containing $[0, c_+^{-1})$, and therefore $\rho_{c_+}(\lambda) \geq \phi(-\lambda)$ on I .

Next we deal with intervals of the form $I = (-\theta, 0]$ where $\theta = a^{-1}$ if $c_- = 0$, and $\theta = c_-^{-1}(1 - \sqrt{1 - 6c_-})$ otherwise. Recall that for any $\lambda \in (-3, 0]$, $\phi(-\lambda) \leq \lambda^2/(2(1 + \lambda/3))$.

If $a \geq 1/3$, that is, $c_- = 0$, $D_{-c_-}(\lambda) = 1 + a\lambda > 0$ for $\lambda \in (a^{-1}, 0]$,

$$\rho_{-c_-}(\lambda) = \frac{\lambda^2}{2(1 + a\lambda)} \geq \frac{\lambda^2}{2(1 + \lambda/3)}.$$

For $a \in (0, 1/3)$, note first that $0 < c_- \leq 1/6$, and that

$$0 < D_{-c_-}(\lambda) \leq 1 + \frac{\lambda}{3} + \frac{\lambda^2}{36} \leq \left(1 + \frac{\lambda}{6}\right)^2$$

for every $\lambda \in (-\theta, 0]$. This also entails that $\rho_{-c_-}(\lambda) \geq \phi(-\lambda)$ for $\lambda \in (-\theta, 0]$. \square

6.13 Some Variations

Next we present a few inequalities that are based on slight variations of the entropy method. These versions differ in the assumptions on how V^+ or V^- are controlled by different functions of Z . These inequalities demonstrate the flexibility of the method, but our aim is not to give an exhaustive list of concentration inequalities that can be obtained this way. The message of this section is that by simple modifications of the main argument one may exploit many special properties of the function f .

We start with inequalities that use negative association between increasing and decreasing functions of Z .

Theorem 6.27 Assume that for some nondecreasing function $g : \mathbb{R} \rightarrow \mathbb{R}$,

$$V^- \leq g(Z).$$

Then for all $t > 0$,

$$P\{Z < EZ - t\} \leq e^{-t^2/(4Eg(Z))}.$$

Proof In order to prove lower-tail inequalities, it suffices to derive suitable upper bounds for the moment-generating function $F(\lambda) = Ee^{\lambda Z}$ for negative values of λ . By the third inequality of Theorem 6.15,

$$\begin{aligned}
& \lambda E [Ze^{\lambda Z}] - E [e^{\lambda Z}] \log E [e^{\lambda Z}] \\
& \leq \sum_{i=1}^n E [e^{\lambda Z} \tau(\lambda(Z'_i - Z)_+)] \\
& \leq \sum_{i=1}^n E [e^{\lambda Z} \lambda^2 (Z'_i - Z)_+^2] \\
& \quad (\text{using } \lambda < 0 \text{ and that } \tau(-x) \leq x^2 \text{ for } x > 0) \\
& = \lambda^2 E [e^{\lambda Z} V^-] \\
& \leq \lambda^2 E [e^{\lambda Z} g(Z)].
\end{aligned}$$

Since $g(Z)$ is a nondecreasing and $e^{\lambda Z}$ is a decreasing function of Z , Chebyshev's association inequality (Theorem 2.14) implies that

$$E [e^{\lambda Z} g(Z)] \leq E [e^{\lambda Z}] E [g(Z)].$$

The inequality obtained has the same form as the differential inequality we saw in the proof of Theorem 6.2 (with $Eg(Z)$ in place of $\nu/2$) and it can be solved in an analogous way to obtain the announced lower-tail inequality. \square

Often it is more natural to bound V^+ by an increasing function of Z than to bound V^- . In such situations one can still say something about lower tail probabilities of Z but we need the additional guarantee that $|Z - Z'_i|$ remains bounded and that the inequality only applies in a restricted range of the values of t .

Theorem 6.28 *Assume that there exists a nondecreasing function g such that $V^+ \leq g(Z)$ and for any value of $X = (X_1, \dots, X_n)$ and X'_i , $|Z - Z'_i| \leq 1$. Then for all $K > 0$, if $\lambda \in [0, 1/K)$, then*

$$\log E e^{-\lambda(Z-EZ)} \leq \lambda^2 \frac{\tau(K)}{K^2} Eg(Z).$$

Moreover, for all $0 < t \leq (e-1)Eg(Z)$, we have

$$P \{Z < EZ - t\} \leq \exp \left(-\frac{t^2}{4(e-1)Eg(Z)} \right).$$

Proof The key observation is that the function $\tau(x)/x^2 = (e^x - 1)/x$ is increasing if $x > 0$. Choose $K > 0$. Thus, for $\lambda \in (-1/K, 0)$, the second inequality of Theorem 6.15 implies that

$$\begin{aligned}
\lambda E[Ze^{\lambda Z}] - E[e^{\lambda Z}] \log E[e^{\lambda Z}] &\leq \sum_{i=1}^n E[e^{\lambda Z} \tau(-\lambda(Z - Z'_i)_+)] \\
&\leq \lambda^2 \frac{\tau(K)}{K^2} E[e^{\lambda Z} V^+] \\
&\leq \lambda^2 \frac{\tau(K)}{K^2} E[g(Z) e^{\lambda Z}],
\end{aligned}$$

where at the last step we used the assumption of the theorem.

As in the proof of Theorem 6.27, we bound $E[g(Z)e^{\lambda Z}]$ by $E[g(Z)]E[e^{\lambda Z}]$. The rest of the proof is identical to that of Theorem 6.27. Here, we took $K = 1$. \square

Our last general result deals with a frequently faced situation. In these cases V^+ may be bounded by the product of Z and another random variable W with well-behaved moment-generating function. The following theorem provides a way to deal with such functionals efficiently and painlessly.

Theorem 6.29 Assume that f is nonnegative and that there exists a random variable W , such that

$$V^+ \leq WZ.$$

Then for all $\theta > 0$ and $\lambda \in (0, 1/\theta)$,

$$\log Ee^{\lambda(\sqrt{Z} - E\sqrt{Z})} \leq \frac{\lambda\theta}{1 - \lambda\theta} \log Ee^{\lambda W/\theta}.$$

Note that this theorem only bounds the moment-generating function of \sqrt{Z} . However, one may easily obtain bounds for the upper-tail probability of Z by observing that, since $\sqrt{EZ} \geq E\sqrt{Z}$, and by writing $x = \sqrt{EZ} + t - \sqrt{EZ}$, we have, for $\lambda > 0$,

$$P\{Z > EZ + t\} \leq P\left\{\sqrt{Z} > E\sqrt{Z} + x\right\} \leq Ee^{\lambda(\sqrt{Z} - E\sqrt{Z})} e^{-\lambda x}$$

by Markov's inequality.

Proof Introduce $Y = \sqrt{Z}$ and $Y^{(i)} = \sqrt{Z^{(i)}}$. Then

$$\begin{aligned}
E' \left[\sum_{i=1}^n (Y - Y^{(i)})_+^2 \right] &= E' \left[\sum_{i=1}^n \left(\sqrt{Z} - \sqrt{Z^{(i)}} \right)_+^2 \right] \\
&\leq E' \left[\sum_{i=1}^n \left(\frac{(Z - Z^{(i)})_+}{\sqrt{Z}} \right)^2 \right] \\
&\leq \frac{1}{Z} E' \left[\sum_{i=1}^n (Z - Z^{(i)})_+^2 \right] \\
&\leq W.
\end{aligned}$$

Thus, applying Theorem 6.16 for Y proves the statement. \square

Example 6.30 (TRIANGLES IN A RANDOM GRAPH) Consider the Erdős–Rényi $G(n, p)$ model of a random graph. Recall that such a graph has n vertices and for each pair (u, v) of vertices an edge is inserted between u and v with probability p , independently. We write $m = \binom{n}{2}$, and denote the indicator variables of the m edges by X_1, \dots, X_m (i.e. $X_i = 1$ if edge $i = (u, v)$ is present in the random graph and $X_i = 0$ otherwise). Three edges form a *triangle* if there are vertices u, v, w such that the edges are of the form (u, v) , (v, w) , and (w, u) . Concentration properties of the number of triangles in a random graph have received a great deal of attention and sharp bounds have been derived by various sophisticated methods for different ranges of the parameter p of the random graph (see the bibliographical remarks at the end of the chapter). Interestingly, the left tail is substantially easier to handle, as Janson’s inequality, presented in the next section, offers sharp estimates. However, proving sharp inequalities for the upper tail was much more challenging. Here we only show some sub-optimal versions that are easy to obtain from the general results of this chapter.

Let $Z = f(X_1, \dots, X_m)$ denote the number of triangles in a random graph. Note that

$$EZ = \frac{n(n-1)(n-2)}{6} p^3 \approx \frac{n^3 p^3}{6}$$

and

$$\text{Var}(Z) = \binom{n}{3} (p^3 - p^6) + \binom{n}{4} \binom{4}{2} (p^5 - p^6).$$

To obtain exponential upper-tail inequalities, we estimate the random variable

$$V^+ = \sum_{i=1}^n E'(Z - Z'_i)_+^2.$$

If v and u denote the extremities of edge i ($1 \leq i \leq m$), then we denote by B_i the number of vertices w such that both edges (u, w) and (v, w) exist in the random graph. Then

$$V^+ = \sum_{i=1}^m X_i (1-p) B_i^2.$$

Since $\sum_{i=1}^m X_i B_i = 3Z$, we have

$$\begin{aligned} V^+ &\leq (1-p) \sum_{i=1}^m X_i \left(\max_{j=1, \dots, m} B_j \right) B_i \\ &= (1-p) \left(\max_{j=1, \dots, m} B_j \right) \sum_{i=1}^m X_i B_i \\ &= 3(1-p) \left(\max_{j=1, \dots, m} B_j \right) Z. \end{aligned}$$

By bounding $\max_{j=1, \dots, m} B_j$ trivially by n , we have $V^+ \leq 3(1-p)nZ$. Define $f_i(X^{(i)})$ as the number of triangles when we force the i -th edge to be absent in the graph. Then

clearly $\sum_{i=1}^n (f(X) - f_i(X^{(i)}))^2 = V^+/(1-p)$ and therefore, using the terminology of Section 6.11, f is weakly $(3n, 0)$ -self-bounding. Thus, by Theorem 6.19,

$$P\{Z \geq EZ + t\} \leq \exp\left(-\frac{t^2}{n^4 p^3 + 3nt}\right).$$

It is clear that in the argument above a lot is lost by bounding $W \stackrel{\text{def}}{=} 3 \max_{j=1, \dots, m} B_j$ by n . Indeed, one may achieve a significant improvement by using Theorem 6.29. In order to do so, we need to bound the moment-generating function of W . This may be done by another application of Theorem 6.19. Let $W^{(i)}$ denote the value of W when edge i is deleted from the random graph (if the graph contained that edge). Then $W^{(i)} \leq W$ and

$$\sum_{i=1}^n \left(W - W^{(i)}\right)^2 \leq 18W,$$

so W is weakly $(18, 0)$ -self-bounding. Hence, by Theorem 6.19,

$$\log Ee^{\lambda(W-EW)} \leq \frac{9\lambda^2 EW}{1-9\lambda}.$$

Denoting $Y = \sqrt{Z}$, Theorem 6.29 leads to

$$\log Ee^{\lambda(Y-EY)} \leq \frac{\lambda}{1-\lambda} \left(\frac{9\lambda^2 EW}{1-9\lambda} + \lambda EW \right) \leq \frac{\lambda^2 EW}{1-10\lambda}.$$

This is a sub-gamma bound for the moment-generating function of Y , and the computations of Sections 2.4 and 2.8 imply

$$P\{Y > EY + t\} \leq \exp\left(-\frac{t^2}{4EW + 20t}\right).$$

Now it remains to bound the expected value of W . Note that $W/3$ is the maximum of $m = \binom{n}{2}$ binomial random variables with parameters (n, p^2) . In order to obtain a quick upper bound for $EW/3$, it is convenient to use the technique presented in Section 2.5 as follows: let S_i with $i \leq m$ denote a sequence of binomially distributed random variables with parameters n and p^2 . By Jensen's inequality,

$$\begin{aligned} EW/3 &\leq \log \left(E \max_{i=1, \dots, m} e^{S_i} \right) \\ &\leq \log (E [me^{S_1}]) \\ &= \log m + \log (Ee^{S_1}) \\ &\leq \log m + (e-1)np^2 \\ &\leq 2 \log n + 2np^2. \end{aligned}$$

Arguably, the most interesting values for p are those when p is at most of the order of $n^{-1/2}$ and in this case, the dominating term in the above expression is $2 \log n$. Hence, we obtain the following bound for the tail of $Y = \sqrt{Z}$

$$P\{Y \geq EY + t\} \leq \exp\left(-\frac{t^2}{24(np^2 + \log n) + 20t}\right).$$

It is now easy to get tail bounds for the number Z of triangles. We spare the reader from the straightforward details (see the exercises).

6.14 Janson's Inequality

As we saw in the examples of Section 6.13, in many cases the special structure of the function of independent random variables can be used to deduce concentration inequalities. In this section we present another general result, a celebrated exponential lower-tail inequality for Boolean polynomials.

More precisely, consider independent binary random variables X_1, \dots, X_n such that $P\{X_i = 1\} = 1 - P\{X_i = 0\} = p_i$ for some $p_1, \dots, p_n \in [0, 1]$. To simplify notation, we identify every binary vector $\alpha \in \{0, 1\}^n$ with the subset of $\{1, \dots, n\}$ defined by the non-zero components of α . For example, for $i \in \{1, \dots, n\}$, we write $i \in \alpha$ to denote that the i -th component of α equals 1. Then for each $\alpha \in \{0, 1\}^n$, we introduce the binary random variable

$$Y_\alpha = \prod_{i \in \alpha} X_i.$$

Given a collection \mathcal{I} of subsets of the binary hypercube $\{0, 1\}^n$, we may define

$$Z = \sum_{\alpha \in \mathcal{I}} Y_\alpha,$$

which is a polynomial of the binary vector $X = (X_1, \dots, X_n)$.

Boolean polynomials of this type are common in many applications of the probabilistic method in discrete mathematics and also in the theory of random graphs, and their concentration properties have been the subject of intensive study. Note that for any $\alpha, \beta \in \mathcal{I}$ with $\alpha \cap \beta = \emptyset$ (i.e. if $\alpha_i \beta_i = 0$ for all $i = 1, \dots, n$), $EY_\alpha Y_\beta = EY_\alpha EY_\beta$ and therefore the variance of Z equals

$$\begin{aligned} \text{Var}(Z) &= EZ^2 - (EZ)^2 = \sum_{\alpha, \beta \in \mathcal{I}} EY_\alpha Y_\beta - \sum_{\alpha, \beta \in \mathcal{I}} EY_\alpha EY_\beta \\ &= \sum_{\alpha, \beta \in \mathcal{I}: \alpha \cap \beta \neq \emptyset} (EY_\alpha Y_\beta - EY_\alpha EY_\beta) \\ &\leq \sum_{\alpha, \beta \in \mathcal{I}: \alpha \cap \beta \neq \emptyset} EY_\alpha Y_\beta \\ &\stackrel{\text{def}}{=} \Delta. \end{aligned}$$

Thus, by Chebyshev's inequality,

$$P\{|Z - EZ| > t\} \leq \frac{\Delta}{t^2}.$$

The next theorem shows the surprising fact that, at least for the lower tail, there is always an exponential version of this inequality.

Theorem 6.31 (JANSON'S INEQUALITY) *Let \mathcal{I} denote a collection of subsets of $\{0, 1\}^n$ and define Z and Δ as above. Then for all $\lambda \leq 0$,*

$$\log Ee^{\lambda(Z-EZ)} \leq \phi\left(\frac{\lambda\Delta}{EZ}\right) \frac{(EZ)^2}{\Delta}$$

where $\phi(x) = e^x - x - 1$. In particular, for all $0 \leq t \leq EZ$,

$$P\{Z \leq EZ - t\} \leq e^{-t^2/(2\Delta)}.$$

The proof of Janson's inequality shown here shows certain similarities with the entropy method. In particular, the proof is based on bounding the derivative of the logarithmic moment-generating function of Z . However, sub-additivity inequalities can be avoided because of a positive association property that can be exploited by an appropriate use of Harris' inequality (Theorem 2.15).

Proof Denote the logarithmic moment generating function of $Z - EZ$ by $G(\lambda) = \log Ee^{\lambda(Z-EZ)}$. Then the derivative of G equals

$$G'(\lambda) = \frac{E[Ze^{\lambda Z}]}{Ee^{\lambda Z}} - EZ = \sum_{\alpha \in \mathcal{I}} \frac{E[Y_\alpha e^{\lambda Z}]}{Ee^{\lambda Z}} - EZ.$$

In the following, we derive an upper bound for each term $E[Y_\alpha e^{\lambda Z}]$ of the sum on the right-hand side.

Fix an $\alpha \in \mathcal{I}$ and introduce $U_\alpha = \sum_{\beta: \beta \cap \alpha \neq \emptyset} Y_\beta$ and $Z_\alpha = \sum_{\beta: \beta \cap \alpha = \emptyset} Y_\beta$. Clearly, regardless of what α is, $Z = U_\alpha + Z_\alpha$. Since

$$E[Y_\alpha e^{\lambda Z}] = E[e^{\lambda Z} | Y_\alpha = 1] EY_\alpha,$$

it suffices to bound the conditional expectation. The key observation is that since $\lambda \leq 0$, both $\exp(\lambda U_\alpha)$ and $\exp(\lambda Z_\alpha)$ are decreasing functions of X_1, \dots, X_n .

$$\begin{aligned}
& E[e^{\lambda Z} \mid Y_\alpha = 1] \\
&= E[e^{\lambda U_\alpha} e^{\lambda Z_\alpha} \mid Y_\alpha = 1] \\
&\geq E[e^{\lambda U_\alpha} \mid Y_\alpha = 1] E[e^{\lambda Z_\alpha} \mid Y_\alpha = 1] \quad (\text{by Harris' inequality}) \\
&= E[e^{\lambda U_\alpha} \mid Y_\alpha = 1] E e^{\lambda Z_\alpha} \quad (\text{since } Z_\alpha \text{ and } Y_\alpha \text{ are independent}) \\
&\geq E[e^{\lambda U_\alpha} \mid Y_\alpha = 1] E e^{\lambda Z} \quad (\text{as } Z_\alpha \leq Z) \\
&\geq e^{\lambda E[U_\alpha \mid Y_\alpha = 1]} E e^{\lambda Z} \quad (\text{by Jensen's inequality}).
\end{aligned}$$

Note that we apply Harris' inequality above conditionally, given $Y_\alpha = 1$. This condition simply forces $X_i = 1$ for all $i \in \alpha$, so both U_α and Z_α are increasing functions of the independent random variables $X_i, i \notin \alpha$ and Harris' inequality is used legally. Thus, we obtain

$$\begin{aligned}
& \frac{E[Z e^{\lambda Z}]}{EZ} \\
&\geq E e^{\lambda Z} \sum_{\alpha \in \mathcal{I}} \frac{E Y_\alpha}{EZ} e^{E[\lambda U_\alpha \mid Y_\alpha = 1]} \\
&\geq E e^{\lambda Z} \exp \left(\sum_{\alpha \in \mathcal{I}} \frac{E Y_\alpha}{EZ} E[\lambda U_\alpha \mid Y_\alpha = 1] \right) \quad (\text{by Jensen's inequality}) \\
&= E e^{\lambda Z} \exp \left(\lambda \frac{\Delta}{EZ} \right)
\end{aligned}$$

where we use the fact that

$$\Delta = \sum_{\alpha \in \mathcal{I}} E[Y_\alpha U_\alpha].$$

Summarizing, we have, for all $\lambda \leq 0$,

$$G'(\lambda) \geq EZ (e^{\lambda \Delta / EZ} - 1).$$

Thus, integrating this inequality between λ and 0 and using $G(0) = 0$, we find that for $\lambda \leq 0$,

$$G(\lambda) \leq -EZ \int_{\lambda}^0 \left(e^{u \frac{\Delta}{EZ}} - 1 \right) du = \phi \left(\frac{\lambda \Delta}{EZ} \right) \frac{(EZ)^2}{\Delta}$$

as desired. The second inequality follows from the simple fact that for $x > 0$, $\phi(-x) \leq x^2/2$. \square

Remark 6.6 (PROBABILITY OF NON-EXISTENCE) In many applications of Janson's inequality, one wishes to show that in a random draw of the vector $X = (X_1, \dots, X_n)$,

with high probability, there exists at least one element $\alpha \in \mathcal{I}$ for which $Y_\alpha = 1$. In other words, the goal is to show that $Z > 0$ with high probability. To this end, one may write

$$P\{Z = 0\} = P\{Z \leq EZ - EZ\} \leq \exp\left(-\frac{(EZ)^2}{2\Delta}\right),$$

which is guaranteed to be exponentially small whenever $\sqrt{\Delta}$ is small compared to EZ .

Example 6.32 (TRIANGLES IN A RANDOM GRAPH) A prototypical application of Janson's inequality is the case of the number of triangles in an Erdős–Rényi random graph $G(n, p)$, discussed in Example 6.30 in the previous section. If Z denotes the number of triangles in $G(n, p)$, then recall that

$$EZ = \binom{n}{3}p^3 \quad \text{and} \quad \text{Var}(Z) = \binom{n}{3}p^3(1-p^3) + 2\binom{n}{4}\binom{4}{2}p^5(1-p).$$

The value of Δ may also be computed in a straightforward way. One obtains

$$\Delta = \binom{n}{3}p^3 + 2\binom{n}{4}\binom{4}{2}p^5$$

which is only slightly larger than $\text{Var}(Z)$. For the probability that the random graph does not contain any triangle, we may use Janson's inequality with $t = EZ$:

$$P\{Z = 0\} \leq \exp\left(-\frac{\binom{n}{3}^2 p^6}{2\left(\binom{n}{3}p^3 + 2\binom{n}{4}\binom{4}{2}p^5\right)}\right) \leq \exp\left(-\frac{\binom{n}{3}p^2}{2(1+2np^2)}\right).$$

6.15 Bibliographical Remarks

The key principles of the entropy method rely on the ideas of proving Gaussian concentration inequalities based on logarithmic Sobolev inequalities. These are summarized in Chapter 5, where we also give some of the main references. It was Michel Ledoux (1997) who realized that these ideas may be used as an alternative route to some of Talagrand's exponential concentration inequalities for empirical processes and Rademacher chaos. Ledoux's ideas were taken further by Massart (2000a), Bousquet (2002a), Klein (2002), Rio (2001), and Klein and Rio (2005), while the core of the material of this chapter is based on Boucheron, Lugosi, and Massart (2000, 2003, 2009).

Different versions of the modified logarithmic Sobolev inequalities used in this chapter are due to Ledoux (1997, 1999, 2001) and Massart (2000a).

The bounded differences inequality is perhaps the simplest and most widely used exponential concentration inequality. The basic idea of writing a function of independent random variables as a sum of martingale differences, and using exponential inequalities for martingales, was first used in various applications by mathematicians including Yurinskii (1976), Maurey (1979), Milman and Schechtman (1986), and Shamir and

Spencer (1987). The inequality was first laid down explicitly and illustrated by a wide variety of applications in an excellent survey paper by McDiarmid (1989), and the result itself has often been referred to as McDiarmid's inequality. Martingale methods have served as a flexible and versatile tool for proving concentration inequalities (see the more recent surveys of McDiarmid (1998), Chung and Lu (2006b), and Dubhashi and Panconesi (2009)).

The exponential tail inequality for sums of independent Hilbert-space valued random variables derived in Example 6.3 is just a simple example. There is a vast literature dealing with tails of sums of vector-valued random variables. It is outside the scope of this book to derive the sharpest and most general results. Here we merely try to make the point that general concentration inequalities prove to be a versatile tool in such applications. In fact, applications of this type motivated some of the most significant advances in the theory of concentration inequalities. In Chapters 11, 12, and 13 we discuss many of the principal modern tools for analyzing the tails of sums of independent vector-valued random variables and empirical processes. For some of the classical references, the interested reader is referred to Yurinskii (1976, 1995), Ledoux and Talagrand (1991), and Pinelis (1995).

The inequality described in Exercise 6.4 was proved independently by Guntuboyina and Leeb (2009) and Bordenave, Caputo, and Chafaï (2011).

Theorem 6.5 is due to McDiarmid (1998) who proved it using martingale methods. The proof presented here is due to Andreas Maurer who kindly permitted us to reproduce his elegant work.

The exponential inequality for the largest eigenvalue of a random symmetric matrix described in Example 6.8 was proved by Alon, Krivelevich, and Vu (2002) who used Talagrand's convex distance inequality. Maurer (2006) obtained a better exponent with a more careful analysis. Alon, Krivelevich, and Vu (2002) show, with a simple extension of the argument, that for the k -th largest (or k -th smallest) eigenvalue the upper bounds become $e^{-t^2/(16k^2)}$, though it is not clear whether the factor k^{-2} in the exponent is necessary.

Theorem 6.9 appears in Maurer (2006). Theorem 6.10 was first established by Talagrand (1996c) who also proves a corresponding lower tail inequality which is presented in Section 7.5. The proof given here is due to Ledoux (1997).

Self-bounding functions were introduced by Boucheron, Lugosi, and Massart (2000) who prove Theorem 6.12 building on techniques developed by Massart (2000a). Various generalizations of the self-bounding property were considered by Boucheron, Lugosi, and Massart (2003, 2009), Boucheron *et al.* (2005b), Devroye (2002), Maurer (2006), and McDiarmid and Reed (2006). In particular, McDiarmid and Reed (2006) considered what we call strongly (a, b) -self-bounding functions and proved results that are only slightly weaker than those presented in Section 6.11. The weak self-bounding property was first considered by Maurer (2006), and Theorem 6.19 is due to him. Theorems 6.21 and 6.20 appear in Boucheron, Lugosi, Massart (2009).

We note here that the inequality linking the expected and annealed VC entropies answers, in a positive way, a question raised by Vapnik (1995, pp. 53–54): the empirical risk minimization procedure is *non-trivially consistent* and *rapidly convergent* if and only if the annealed entropy rate $(1/n) \log_2 ET(X)$ converges to zero. For the definitions and discussion we refer to Vapnik (1995).

The material of Sections 6.8, 6.9, and 6.13 is based on Boucheron, Lugosi, and Massart (2003).

Klaassen (1985) showed that Poisson distributions satisfy the “modified Poincaré” inequality

$$\text{Var}(f(Z)) \leq EZ \times E[|Df(Z)|^2]$$

(see Exercise 3.21).

The search for modified logarithmic Sobolev inequalities, that is, functional inequalities which capture the tail behavior of distributions that are less concentrated than the Gaussian distribution, was initiated by Bobkov and Ledoux (1997). Their aim was to recover some results of Talagrand concerning concentration properties of the exponential distribution. Bobkov and Ledoux (1997, 1998) pointed out that the Poisson distribution cannot satisfy an analog of the Gaussian logarithmic Sobolev inequality. They establish the second inequality of Theorem 6.17. The first inequality of Theorem 6.17 is due to Wu (2000). Other modified logarithmic Sobolev inequalities have been investigated by Ané and Ledoux (2000), Chafai (2006), Bobkov and Tetali (2006), and others.

Janson’s inequality (Theorem 6.31) was first established by Janson (1990). This inequality has since become one of the basic standard tools of the probabilistic method of discrete mathematics and random graph theory, and many variations, refinements, and alternative proofs are now known. We refer the reader to the monographs of Alon and Spencer (1992), and Janson, Łuczak, and Ruciński (2000) for surveys and further references.

The number of triangles, and more generally, the number of copies of a fixed subgraph, in a random graph $G(n, p)$ has been a subject of intensive study. For the lower-tail probabilities, Janson’s inequality, shown in Section 6.14, gives an essentially tight bound. However, obtaining sharp bounds for the upper tail has been an important non-trivial challenge. For such upper-tail inequalities we refer the interested reader to the papers Kim and Vu (2000, 2004), Vu (2000, 2001), Janson and Ruciński (2004, 2002), Janson, Oleszkiewicz, and Ruciński (2004), Bolthausen, Comets, and Dembo (2009), Döring and Eichelsbacher (2009), Chatterjee and Dey (2010), Chatterjee (2010), DeMarco and Kahn (2010), and Schudy and Sviridenko (2012).

The inequalities derived in Example 6.30 are not the best possible.

6.16 EXERCISES

- 6.1. Relax the condition of Theorem 6.7 in the following way. Show that if $X = (X_1, \dots, X_n)$ and

$$E \left[\sum_{i=1}^n (Z - Z'_i)_+^2 \middle| X \right] \leq \nu$$

then for all $t > 0$,

$$P\{Z > EZ + t\} \leq e^{-t^2/(2v)}$$

and if

$$E\left[\sum_{i=1}^n (Z - Z'_i)^2 \middle| X\right] \leq v,$$

then

$$P\{Z < EZ - t\} \leq e^{-t^2/(2v)}.$$

- 6.2. (THE CAUCHY INTERLACING THEOREM) Let A be an $n \times n$ Hermitian matrix with eigenvalues $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n$. Denote by R_A the Rayleigh quotient defined, for every $x \in \mathbb{C}^n \setminus \{0\}$, by

$$R_A(x) = \frac{x^* A x}{x^* x}.$$

Prove the min-max formulas

$$\alpha_k = \max \left\{ \min \{R_A(x) : x \in U \text{ and } x \neq 0\} : \dim(U) = n - k + 1 \right\}$$

and

$$\alpha_k = \min \left\{ \max \{R_A(x) : x \in U \text{ and } x \neq 0\} : \dim(U) = k \right\}.$$

Let P be an orthogonal projection matrix with rank m and define the Hermitian matrix $B = PAP$. Denoting by $\beta_1 \leq \beta_2 \leq \dots \leq \beta_m$ the eigenvalues of B , using the minmax formulas, show that the eigenvalues of A and B interlace, that is, for all $j \leq m$, $\alpha_j \leq \beta_j \leq \alpha_{n-m+j}$. (See Bai and Silverstein (2010).)

- 6.3. (RANK INEQUALITY FOR SPECTRAL MEASURES) Let A and B be $n \times n$ Hermitian matrices and denote by F_A and F_B the distribution functions related to the spectral measures L_A and L_B of A and B , respectively. Setting $k = \text{rank}(A - B)$, prove the rank inequality

$$\|F_A - F_B\|_\infty \leq \frac{k}{n}.$$

Hint: show that one can always assume that

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \text{ and } B = \begin{bmatrix} B_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where the order of A_{22} is $n - k \times n - k$. Use the Cauchy interlacing theorem (see Exercise 6.2 above) for the pairs of Hermitian matrices A and A_{22} on the one hand and B and A_{22} on the other hand. (See Bai and Silverstein (2010).)

- 6.4. Show that the convexity assumption is essential in Theorem 6.10, by considering the following example: let n be an even positive integer and define $A = \{x \in [0, 1]^n : \sum_{i=1}^n x_i \leq n/2\}$. Let $f(x) = \inf_{y \in A} \|x - y\|$. Then clearly f is Lipschitz but not convex. Let the components of $X = (X_1, \dots, X_n)$ be i.i.d. with $P\{X_i = 0\} = P\{X_i = 1\} = 1/2$. Show that there exists a constant $c > 0$ such that $P\{f(X) > Mf(X) + cn^{1/4}\} \geq 1/4$ for all sufficiently large n . (This example is taken from Ledoux and Talagrand (1991, p. 17).)
- 6.5. Prove the following generalization of Theorem 6.10. Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex compact set with diameter B . Let X_1, \dots, X_n be independent random variables taking values in \mathcal{X} and assume that $f : \mathcal{X}^n \rightarrow \mathbb{R}$ is *separately convex* and Lipschitz, that is, $|f(x) - f(y)| \leq \|x - y\|$ for all $x, y \in \mathcal{X}^n \subset \mathbb{R}^{dn}$. Then $Z = f(X_1, \dots, X_n)$ satisfies, for all $t > 0$,

$$P\{Z > EZ + t\} \leq e^{-t^2/(2B^2)}.$$

- 6.6. Let X_1, \dots, X_n be independent vector-valued random variables taking values in a compact convex set $\mathcal{X} \subset \mathbb{R}^d$ with diameter B . Let A denote the $d \times n$ matrix whose columns are X_1, \dots, X_n and let Z denote the largest singular value of A . Show that

$$P\{Z > EZ + t\} \leq e^{-t^2/(2B^2)}.$$

Compare the result with Example 6.11.

- 6.7. Assume that $Z = f(X) = f(X_1, \dots, X_n)$ where X_1, \dots, X_n are independent real-valued random variables and f is a nondecreasing function of each variable. Suppose that there exists another nondecreasing function $g : \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$\sum_{i=1}^n (Z - Z'_i)^2 \leq g(X).$$

Show that for all $t > 0$,

$$P\{Z < EZ - t\} \leq e^{-t^2/(4Eg(X))}.$$

Hint: use Harris' inequality (Theorem 2.15).

- 6.8. (ALMOST BOUNDED DIFFERENCES) Assume that $Z = f(X) = f(X_1, \dots, X_n)$ where X_1, \dots, X_n are independent real-valued random variables. Assume there exists a monotone set $A \subset \mathbb{R}^n$ and constants $\nu, C > 0$ such that for $x = (x_1, \dots, x_n) \in A$, $\sum_{i=1}^n (f(x) - \inf_{x'_i} f(x_1, \dots, x'_i, \dots, x_n))^2 \leq \nu$ and for all $x \notin A$, $\sum_{i=1}^n (f(x) - \inf_{x'_i} f(x_1, \dots, x'_i, \dots, x_n))^2 \leq C$. (A monotone set is such that if $x \in A$ and $y \geq x$ (component-wise) then $y \in A$.) Show that for all $t > 0$,

$$P\{Z > EZ + t\} \leq \exp\left(\frac{-t^2}{2(v + CP\{X \notin A\})}\right).$$

Hint: use Harris' inequality (Theorem 2.15).

- 6.9. (RADEMACHER CHAOS OF ORDER TWO) Let \mathcal{T} be a finite set of $n \times n$ symmetric matrices with zero diagonal entries. Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ be a vector of independent Rademacher variables. Let

$$Z = \max_{M \in \mathcal{T}} \sum_{i=1}^n \sum_{j=1}^n M_{ij} \varepsilon_i \varepsilon_j$$

and

$$Y = \max_{M \in \mathcal{T}} \left(\sum_{i=1}^n \left(\sum_{j=1}^n \varepsilon_j M_{ij} \right)^2 \right)^{1/2}.$$

Let $B = \max_{M \in \mathcal{T}} \|M\|^2$ where $\|M\|$ denotes the (operator) norm of matrix M . Prove that

$$\begin{aligned} \text{Var}(Z) &\leq 8E[Y^2] \\ \text{Var}(Y^2) &\leq 8BE[Y^2] \\ \log Ee^{\lambda(Y^2 - EY^2)} &\leq \frac{\lambda^2}{(1 - 8B\lambda)} 8BE[Y^2] \\ \log Ee^{\lambda(Z - EZ)} &\leq \frac{16\lambda^2}{2(1 - 64B\lambda)} E[Y^2], \end{aligned}$$

where $\lambda \geq 0$. *Hint:* use Theorem 6.16 twice. Show that $8Y^2$ upper bounds an Efron–Stein estimate of the variance of Z . Then use the fact that Y may be represented as the supremum of a Rademacher process, and prove that Y^2 is $(16B, 0)$ -weakly self-bounding. Note that

$$E[Y^2] = E \left[\sup_{M \in \mathcal{T}} \sum_{i,j=1}^n \varepsilon_i \varepsilon_j M_{ij}^2 \right].$$

See Talagrand (1996b), Ledoux (1997), and Boucheron, Lugosi, and Massart (2003).

- 6.10. Prove Theorem 6.17. *Hint:* use Lemma 6.18 and the so-called “law of rare events,” that is, the convergence of the binomial distribution to a Poisson.
- 6.11. (A LOGARITHMIC SOBOLEV INEQUALITY FOR THE EXPONENTIAL DISTRIBUTION). Assume X is exponentially distributed, that is, it has density $\exp(-x)$ for $x > 0$. Prove that if $f : [0, \infty) \rightarrow \mathbb{R}$ is differentiable, then

$$\text{Ent}((f(X))^2) \leq 4E[X(f'(X))^2].$$

Hint: use the fact that if X_1 and X_2 are independent standard Gaussian random variables, $(X_1^2 + X_2^2)/2$ is exponentially distributed, and use the Gaussian logarithmic Sobolev inequality.

- 6.12. (SQUARE ROOT OF A POISSON RANDOM VARIABLE) Let X be a Poisson random variable. Prove that for $0 \leq \lambda < 1/2$,

$$\log Ee^{\lambda(\sqrt{X}-E\sqrt{X})} \leq \frac{\lambda^2}{1-2\lambda}.$$

Show that

$$\log Ee^{\lambda(\sqrt{X}-E\sqrt{X})} \leq \nu\lambda(e^\lambda - 1)$$

where $\nu = (EX)E[1/(4X+1)]$. Use Markov's inequality to show that

$$P\left\{\sqrt{X} \geq E\sqrt{X} + t\right\} \leq \exp\left(-\frac{t}{2} \log\left(1 + \frac{t}{2\nu}\right)\right).$$

Hint: the first inequality may be derived from Theorem 6.29. The second inequality may be derived from Theorem 6.17.

- 6.13. (ENTROPIC VERSION OF THE LAW OF RARE EVENTS) Let X be a random variable taking nonnegative integer values and define $p(k) = P\{X = k\}$ for $k = 0, 1, 2, \dots$. The *scaled Fisher information* of X is defined by

$$K(X) = (EX)E\left[\left(\frac{(X+1)p(X+1)}{(EX)p(X)} - 1\right)^2\right].$$

Let $\mu = EX$. Use Theorem 6.17 to prove that the Kullback–Leibler divergence of X and a $\text{Poisson}(\mu)$ random variable is at most $K(X)$.

Let S be the sum of the independent integer-valued random variables X_1, \dots, X_n with $EX_i = p_i$. Let $\mu = \sum_{i=1}^n p_i$. Prove that

$$K(S) \leq \sum_{i=1}^n \frac{p_i}{\mu} K(X_i).$$

From this sub-additivity property, prove that the Kullback–Leibler divergence of S and a $\text{Poisson}(\mu)$ random variable is at most $(1/\mu) \sum_{i=1}^n p_i^3 / (1 - p_i)$. (See Kontoyiannis, Harremoës, and Johnson (2005).)

- 6.14. Consider the maximal degree D of any vertex in a random $G(n, p)$ graph defined as in Example 6.13. Show that for any sequence $a_n \rightarrow \infty$, with probability tending to 1 as $n \rightarrow \infty$,

$$\left| D - np - \sqrt{2p(1-p)n \log n} \right| \leq a_n \sqrt{\frac{p(1-p)n}{\log n}}$$

(see Bollobás (2001, Corollary 3.14)). What do you obtain if you combine Lemma 2.4 with Theorem 6.12?

- 6.15. (LOWER BOUND FOR TRIANGLES) Let Z denote the number of triangles in a random graph $G(n, p)$ where $p \geq 1/n$. Show that for every $a > 0$ there exists a constant $c = c(a)$ such that

$$P\{Z > EZ + an^3p^3\} \geq e^{-cp^2n^2 \log(1/p)}.$$

Hint: the lower bound is the probability that a fixed clique of size proportional to np exists in $G(n, p)$. (Vu (2001).)

- 6.16. Let Z be as in the previous exercise. Use the inequality for \sqrt{Z} shown in the text to prove that for any $K > 1$, if $t \leq (K^2 - 1)EZ$, then

$$P\{Z > EZ + t\} \leq \exp \left(-\frac{t^2}{(K+1)^2 EZ \left(24np^2 + 24 \log n + \frac{20t}{(K+1)\sqrt{EZ}} \right)} \right).$$