Entropy and Concentration

Andreas Maurer

1 Introduction

Concentration inequalities bound the probabilities that random quantities deviate from their average, median, or otherwise typical values. If this deviation is small with high probability, then a repeated experiment or observation will likely produce a similar result. In this way concentration inequalities can give quantitative guarantees of reproducibility, a concept at the heart of empirical science [25].

In this chapter we limit ourselves to study quantities whose randomness arises through the dependence on many independent random variables. Suppose that (Ω_i, Σ_i) are measurable spaces for $i \in \{1, ..., n\}$ and that f is real valued function defined on the product space $\Omega = \prod_{i=1}^n \Omega_i$,

$$f : \mathbf{x} = (x_1, ..., x_n) \in \Omega \mapsto f(\mathbf{x}) \in \mathbb{R}.$$

Now let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent random variables, where X_i is distributed as μ_i in Ω_i. For $t > 0$ and $\mathbf{X}'$ iid to $\mathbf{X}$ we then want to give bounds on the upwards deviation probability

$$\Pr_{\mathbf{X}} \left\{ f(\mathbf{X}) - E\left[f\left(\mathbf{X}'\right)\right] > t \right\}$$

in terms of the deviation t, the measures μ_i and properties of the function f. Downward deviation bounds are then obtained by replacing f with $-f$. Usually we will just write $\Pr\{f - Ef > t\}$ for the deviation probability above.

The first bounds of this type were given by Chebychev and Bienaimé [11] in the late 19th century for additive functions of the form

A. Maurer (✉)
Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy
e-mail: am@andreas-maurer.eu

F. De Mari and E. De Vito (eds.), *Harmonic and Applied Analysis*, Applied and Numerical Harmonic Analysis, https://doi.org/10.1007/978-3-030-86664-8_2

$$f(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i). \tag{1}$$

The subject has since been developed by Bernstein, Chernoff, Bennett, Hoeffding, and many others [4, 16], and results were extended from sums to more general and complicated nonlinear functions. During the past decades research activity has been stimulated by the contributions of Michel Talagrand [27, 28] and by the relevance of concentration phenomena to the rapidly growing field of computer science. Some concentration inequalities, like the well known bounded difference inequality, have become standard tools in the analysis of algorithms [23].

One of the more recent methods to derive concentration inequalities, the so-called *entropy method*, is rooted in the early investigations of Boltzmann [5] and Gibbs [12] into the foundations of statistical mechanics. While the modern entropy method evolved along a complicated historical path via quantum field theory and the logarithmic Sobolev-inequality of Leonard Gross [14], its hidden simplicity was understood and emphasized by Michel Ledoux, who also recognized the key role which the subadditivity of entropy can play in the derivation of concentration inequalities [18]. The method has been refined by Bobkov, Massart [20], Bousquet [9], and Boucheron et al. [7]. Recently Boucheron et al. [8] showed that the entropy method is sufficiently strong to derive a form of Talagrand's convex distance inequality.

In this chapter we present a variation of the entropy method in a compact and simplified form, closely tied to its origins in statistical mechanics. We give an exposition of the method in Sect. 2 and compress it into a toolbox to derive concentration inequalities.

In Sect. 3 we will then use this method to prove two classical concentration inequalities, the bounded difference inequality and a generalization of Bennett's inequality. As example applications we treat vector-valued concentration and generalization in empirical risk minimization, a standard problem in machine learning theory.

In Sect. 4 we address more difficult problems. The bounded difference inequality is used to prove the famous Gaussian concentration inequality. We also give some more recent inequalities which we apply to analyze the concentration of convex Lipschitz functions on $[0, 1]^n$, or of the spectral norm of a random matrix.

In Sect. 5 we describe some of the more advanced techniques, self-boundedness, and decoupling. As examples we give sub-Gaussian lower tail bounds for convex Lipschitz functions and a version of the Hanson-Wright inequality for bounded random variables and we derive an exponential inequality for the suprema of empirical processes. We conclude with another version of Bernstein's inequality and its application to U-statistics.

We limit ourselves to exponential deviation bounds from the mean. For moment bounds and other advanced methods to establish concentration inequalities, such as the transportation method or an in-depth treatment of logarithmic Sobolev inequalities, we recommend the monographs by Ledoux [18] and Boucheron, Lugosi, and Massart [6], and the overview article by McDiarmid [23]. Another important recent development not covered is the method of exchangeable pairs proposed by Chatterjee [10] .

We fix some conventions and notation:

If (Ω, Σ) is any measurable space $\mathcal{A}(\Omega)$ will denote the algebra of bounded, measurable real valued functions on Ω. When there is no ambiguity we often just write $\mathcal{A}$ for $\mathcal{A}(\Omega)$. Although we give some results for unbounded functions, most functions for which we will prove concentration inequalities are assumed to be measurable and bounded, that is $f \in \mathcal{A}$. This assumption simplifies the statement of our results, because it guarantees the existence of algebraic and exponential moments and makes our arguments more transparent.

If (Ω, Σ, μ) is a probability space we write $\Pr F = \mu(F)$ for $F \in \Sigma$, and $E[f] = \int_\Omega f d\mu$ for $f \in L_1[\mu]$ and $\sigma^2[f] = E\left[(f - E[f])^2\right]$ for $f \in L_2[\mu]$. Wherever we use Pr, E or σ^2, we assume that there is an underlying probability space (Ω, Σ, μ). If we refer to other measures than μ, then we identify them with corresponding subscripts.

If $\mathcal{X}$ is any set and $n \in \mathbb{N}$, then for $y \in \mathcal{X}$ and $k \in \{1, ..., n\}$ the substitution operator $S_y^k : \mathcal{X}^n \to \mathcal{X}^n$ is defined as

$$S_y^k x = (x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n) \text{ for } x = (x_1, ..., x_n) \in \mathcal{X}^n.$$

This and other notation which we introduce along the way is also summarized in a final section in tabular form.

2 The Entropy Method

In this section we develop the entropy method and package it into a toolbox to prove concentration inequalities.

2.1 Markov's Inequality and Exponential Moment Method

The most important tool in the proof of deviation bounds is Markov's inequality, which we now introduce along with two corollaries, Chebychev's inequality and the exponential moment method.

Theorem 1 (Markov inequality) *Let* $f \in L_1[\mu]$, $f \geq 0$ *and* $t > 0$. *Then*

$$\Pr\{f > t\} \leq \frac{E[f]}{t}$$

Proof Since $f \geq 0$ and $t > 0$ we have $1_{f>t} \leq f/t$ and therefore

$$\Pr\{f > t\} = E\left[1_{f>t}\right] \leq E[f/t] = \frac{E[f]}{t}.$$

□

Corollary 2 (Chebychev inequality) *Let* $f \in L_2[\mu]$ *and* $t > 0$. *Then*

$$\Pr\{|f - E[f]| > t\} = \Pr\left\{(f - E[f])^2 > t^2\right\} \leq \frac{E\left[(f - E[f])^2\right]}{t^2} = \frac{\sigma^2(f)}{t^2}.$$

To use Chebychev's inequality we need to bound the variance $\sigma^2(f)$. If f is a sum of independent variables, the variance of f is just the sum of the variances of the individual variables, but this doesn't work for general functions. In Sect. 3.1, however, we give the Efron–Stein inequality, which asserts that for functions of independent variables the variance is bounded by the expected sum of conditional variances.

The idea of Chebychev's inequality obviously extends to other even centered moments $E\left[(f - E[f])^{2p}\right]$. Bounding higher moments of functions of independent variables is an important technique discussed, for example, in [6].

Here the most important corollary of Markov's inequality is the *exponential moment method*, an idea apparently due to Bernstein [4].

Corollary 3 (exponential moment method) *Let* $f \in \mathcal{A}$, $\beta \geq 0$ *and* $t > 0$. *Then*

$$\Pr\{f > t\} = \Pr\left\{e^{\beta f} > e^{\beta t}\right\} \leq e^{-\beta t} E\left[e^{\beta f}\right].$$

To use this we need to bound the quantity $E\left[e^{\beta f}\right]$ and to optimize the right-hand side above over β. We call $E\left[e^{\beta f}\right]$ the *partition function*, denoted $Z_{\beta f} = E\left[e^{\beta f}\right]$. Bounding the partition function (or its logarithm) is the principal problem in the derivation of exponential tail bounds.

If f is a sum of independent components (as in (1)), then the partition function is the product of the partition functions corresponding to these components, and its logarithm (called the *moment generating function*) is additive. This is a convenient basis to obtain deviation bounds for sums, but it does not immediately extend to general non-additive functions. The approach is taken here, the entropy method, balances simplicity, and generality.

2.2 *Entropy and Concentration*

For the remainder of this section we take the function $f \in \mathcal{A}$ as fixed. We could interpret the points $x \in \Omega$ as possible states of a physical system and f as the negative energy (or Hamiltonian) function, so that $-f(x)$ is the system's energy in the state x. The measure μ then models an a priori probability distribution of states in the absence of any constraining information. We will define another probability measure on Ω, with specified expected energy, but with otherwise minimal assumptions.

If ρ is a function on Ω, $\rho \geq 0$ and $E[\rho] = 1$, the Kullback–Leibler divergence $KL(\rho d\mu, d\mu)$ of $\rho d\mu$ to $d\mu$ is

$$KL(\rho d\mu, d\mu) = E[\rho \ln \rho].$$

$KL(\rho d\mu, d\mu)$ can be interpreted as the information we gain, if we are told that the probability measure is $\rho d\mu$ instead of the a priori measure $d\mu$.

Theorem 4 *For all $f \in \mathcal{A}$, $\beta \in \mathbb{R}$*

$$\sup_{\rho} \beta E[\rho f] - E[\rho \ln \rho] = \ln E\left[e^{\beta f}\right],$$

where the supremum is over all nonnegative measurable functions ρ on Ω satisfying $E[\rho] = 1$.

The supremum is attained for the density

$$\rho_{\beta f} = e^{\beta f} / E\left[e^{\beta f}\right].$$

Proof We can assume $\beta = 1$ by absorbing it in f. Let $\rho \geq 0$ satisfy $E[\rho] = 1$, so that $\rho d\mu$ is a probability measure and $g \in \mathcal{A} \mapsto E_\rho[g] := E[\rho g]$ an expectation functional. Let $\phi(x) = 1/\rho(x)$ if $\rho(x) > 0$ and $\phi(x) = 0$ if $\rho(x) = 0$. Then $E[\rho \ln \rho] = -E[\rho \ln \phi] = -E_\rho[\ln \phi]$ and with Jensen's inequality

$$\begin{aligned} E[\rho f] - E[\rho \ln \rho] &= E_\rho[f + \ln \phi] = \ln \exp\left(E_\rho[f + \ln \phi]\right) \\ &\leq \ln E_\rho\left[\exp(f + \ln \phi)\right] = \ln E_\rho\left[\phi e^f\right] \\ &= \ln E\left[\rho \phi e^f\right] = \ln E\left[1_{\rho > 0} e^f\right] \\ &\leq \ln E\left[e^f\right]. \end{aligned}$$

On the other hand

$$E\left[\rho_f f\right] - E\left[\rho_f \ln \rho_f\right] = \frac{E\left[f e^f\right]}{E\left[e^f\right]} - \frac{E\left[e^f \ln\left(e^f / E\left[e^f\right]\right)\right]}{E\left[e^f\right]} = \ln E\left[e^f\right].$$

□

In statistical physics the maximizing probability measure $d\mu_{\beta f} = \rho_{\beta f} d\mu = e^{\beta f} d\mu / E\left[e^{\beta f}\right]$ is called the *thermal measure*, sometimes also the *canonical ensemble*. It is used to describe a system in thermal equilibrium with a heat reservoir at temperature $T \approx 1/\beta$. The corresponding expectation functional

$$E_{\beta f}[g] = \frac{E\left[g e^{\beta f}\right]}{E\left[e^{\beta f}\right]} = Z_{\beta f}^{-1} E\left[g e^{\beta f}\right], \text{ for } g \in \mathcal{A}$$

is called the *thermal expectation*. The normalizing quantity $Z_{\beta f} = E\left[e^{\beta f}\right]$ is the *partition function* already introduced above. Notice that for any constant c we have $E_{\beta(f+c)}[g] = E_{\beta f}[g]$.

The value of the function $\rho \mapsto E[\rho \ln \rho]$ at the thermal density $\rho_{\beta f} = Z_{\beta f}^{-1} e^{\beta f}$ is called the *canonical entropy* or simply entropy,

$$\mathrm{Ent}_f(\beta) = E\left[\rho_{\beta f} \ln \rho_{\beta f}\right] = \beta E_{\beta f}[f] - \ln Z_{\beta f}. \tag{2}$$

Note that $\mathrm{Ent}_{-f}(\beta) = \mathrm{Ent}_f(-\beta)$, a simple but very useful fact.

Suppose that ρ is any probability density on Ω, which gives the same expected value for the energy as $\rho_{\beta f}$, so that $E[\rho f] = E_{\beta f}[f]$. Then

$$\begin{aligned} 0 &\leq KL\left(\rho d\mu, Z_{\beta f}^{-1} e^{\beta f} d\mu\right) \\ &= E[\rho \ln \rho] - \beta E[\rho f] + \ln Z_{\beta f} \\ &= E[\rho \ln \rho] - \beta E_{\beta f}[f] + \ln Z_{\beta f} \\ &= KL(\rho d\mu, d\mu) - KL\left(\rho_{\beta f} d\mu, d\mu\right). \end{aligned}$$

The thermal measure $d\mu_{\beta f} = \rho_{\beta f} d\mu$ therefore minimizes the information gain relative to the a priori measure $d\mu$, given the expected value $-E_{\beta f}[f]$ of the internal energy.

For $g \in \mathcal{A}$ and $\rho = Z_{\beta f}^{-1} e^{\beta f}$ Theorem 4 gives

$$E_{\beta f}[g] \leq \mathrm{Ent}_f(\beta) + \ln E\left[e^g\right],$$

which allows to decouple g from f. This plays an important role later on in this chapter.

For $\beta \neq 0$ define a function

$$A_f(\beta) = \frac{1}{\beta} \ln Z_{\beta f} = \frac{1}{\beta} \ln E\left[e^{\beta f}\right]. \tag{3}$$

By l'Hospital's rule we have $\lim_{\beta \to 0} A_f(\beta) = E[f]$, so A_f extends continuously to $\mathbb{R}$ by setting $A_f(0) = E[f]$. In statistical physics the quantity $A_f(\beta)$ so defined is called the *free energy* corresponding to the Hamiltonian (energy function) $H = -f$ and temperature $T \approx \beta^{-1}$. Theorem 4 exhibits the free energy and the canonical entropy as a pair of convex conjugates. Dividing (2) by β and writing $U = E_{\beta f}[f]$, we recover the classical thermodynamic relation

$$A = U - T\ \mathrm{Ent},$$

which describes the macroscopically available energy A as the difference between the internal energy U and an energy portion T Ent, which is inaccessible due to ignorance of the microscopic state.

The following theorem establishes the connection of entropy, the exponential moment method and concentration inequalities.

Theorem 5 *For $f \in \mathcal{A}$ and any $\beta \geq 0$ we have*

$$\ln E\left[e^{\beta(f-Ef)}\right] = \beta \int_0^\beta \frac{Ent_f(\gamma)}{\gamma^2} d\gamma$$

and, for $t \geq 0$,

$$\Pr\{f - Ef > t\} \leq \inf_{\beta \geq 0} \exp\left(\beta \int_0^\beta \frac{Ent_f(\gamma)}{\gamma^2} d\gamma - \beta t\right).$$

Proof Differentiating the free energy with respect to β we find

$$A_f'(\beta) = \frac{1}{\beta} E_{\beta f}[f] - \frac{1}{\beta^2} \ln Z_{\beta f} = \beta^{-2} \mathrm{Ent}_f(\beta).$$

By the fundamental theorem of calculus

$$\begin{aligned} \ln E\left[e^{\beta(f-Ef)}\right] &= \ln Z_{\beta f} - \beta E[f] = \beta\left(A_f(\beta) - A_f(0)\right) \\ &= \beta \int_0^\beta A_f'(\gamma)\, d\gamma = \beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma, \end{aligned}$$

which is the first inequality. Then by Markov's inequality

$$\begin{aligned} \Pr\{f - Ef > t\} &\leq e^{-\beta t} E\left[e^{\beta(f-Ef)}\right] \\ &\leq \exp\left(\beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma - \beta t\right). \end{aligned}$$

□

Our strategy to establish concentration results will therefore be the search for appropriate bounds on the entropy.

2.3 Entropy and Energy Fluctuations

The *thermal variance* of a function $g \in \mathcal{A}$ is just the variance of g relative to the thermal expectation. It is denoted $\sigma_{\beta f}^2(g)$ and defined by

$$\sigma_{\beta f}^2(g) = E_{\beta f}\left[\left(g - E_{\beta f}[g]\right)^2\right] = E_{\beta f}\left[g^2\right] - \left(E_{\beta f}[g]\right)^2.$$

For constant c we have $\sigma_{\beta(f+c)}^2[g] = \sigma_{\beta f}^2[g]$.

The proof of the following lemma consists of straightforward calculations, an easy exercise to familiarize oneself with thermal measure, expectation and variance.

Lemma 6 *The following formulas hold for* $f \in \mathcal{A}$
1. $\frac{d}{d\beta}\left(\ln Z_{\beta f}\right) = E_{\beta f}[f]$.
2. If $h : \beta \mapsto h(\beta) \in \mathcal{A}$ *is differentiable and* $(d/d\beta)\, h(\beta) \in \mathcal{A}$ *then*

$$\frac{d}{d\beta} E_{\beta f}[h(\beta)] = E_{\beta f}[h(\beta) f] - E_{\beta f}[h(\beta)]\, E_{\beta f}[f] + E_{\beta f}\left[\frac{d}{d\beta} h(\beta)\right].$$

3. $\frac{d}{d\beta} E_{\beta f}\left[f^k\right] = E_{\beta f}\left[f^{k+1}\right] - E_{\beta f}\left[f^k\right] E_{\beta f}[f]$.
4. $\frac{d^2}{d\beta^2}\left(\ln Z_{\beta f}\right) = \frac{d}{d\beta} E_{\beta f}[f] = \sigma^2_{\beta f}[f]$.

Proof 1. is immediate and 2. a straightforward computation. 3. and 4. are immediate consequences of 1. and 2. □

Since the members of $\mathcal{A}$ are bounded it follows from 2. that for $f, g \in \mathcal{A}$ the functions $\beta \mapsto E_{\beta f}[g]$ and $\beta \mapsto \sigma^2_{\beta f}[g]$ are C_∞.

The thermal variance of f itself corresponds to energy fluctuations. The next theorem represents entropy as a double integral of such fluctuations. The utility of this representation to derive concentration results has been noted by David McAllester [22].

Theorem 7 *We have for* $\beta > 0$

$$Ent_f(\beta) = \int_0^\beta \int_t^\beta \sigma^2_{sf}[f]\, ds\, dt.$$

Proof Using the previous lemma and the fundamental theorem of calculus we obtain the formulas

$$\beta E_{\beta f}[f] = \int_0^\beta E_{\beta f}[f]\, dt = \int_0^\beta \left(\int_0^\beta \sigma^2_{sf}[f]\, ds + E[f]\right) dt$$

and

$$\ln Z_{\beta f} = \int_0^\beta E_{tf}[f]\, dt = \int_0^\beta \left(\int_0^t \sigma^2_{sf}[f]\, ds + E[f]\right) dt,$$

which we subtract to obtain

$$\begin{aligned} \mathrm{Ent}_f(\beta) = \beta E_{\beta f}[f] - \ln Z_{\beta f} &= \int_0^\beta \left(\int_0^\beta \sigma^2_{sf}[f]\, ds - \int_0^t \sigma^2_{sf}[f]\, ds\right) dt \\ &= \int_0^\beta \left(\int_t^\beta \sigma^2_{sf}[f]\, ds\right) dt. \end{aligned}$$

□

Since bounding $\sigma^2_{\beta f}[f]$ is central to our method, it is worth mentioning an interpretation in terms of heat capacity, or specific heat. Recall that $-E_{\beta f}[f]$ is the expected internal energy. The rate of change of this quantity with temperature T is the heat capacity. By conclusion 4 of Lemma 6 we have

$$\frac{d}{dT}\left(-E_{\beta f}[f]\right) = \frac{1}{T^2}\sigma^2_{\beta f}[f],$$

which exhibits the proportionality of heat capacity and energy fluctuations.

2.4 Product Spaces and Conditional Operations

We now set $\Omega = \prod_{k=1}^n \Omega_k$ and $d\mu = \prod_{k=1}^n d\mu_k$, where each μ_k is the probability measure representing the distribution of some variable X_k in the space Ω_k, so that the X_k are assumed to be independent.

With $\mathcal{A}_k$ we denote the subalgebra of those functions $f \in \mathcal{A}$, which are independent of the k-th argument. To efficiently deal with operations performed on individual arguments of functions in $\mathcal{A}$ we need some special notation.

Now let $k \in \{1, ..., n\}$ and $y \in \Omega_k$. If Ξ is any set and F is any function $F : \Omega \to \Xi$, we extend the definition of the *substitution operator* S^k_y to F by $S^k_y(F) = F \circ S^k_y$. This means

$$S^k_y(F)(x_1, ..., x_n) = F(x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n),$$

so the k-th argument is simply replaced by y. Since for $f \in \mathcal{A}$ the function $S^k_y f$ is independent of x_k (which had been replaced by y) we see that S^k_y is a homomorphic (linear and multiplication-preserving) projection of $\mathcal{A}$ onto $\mathcal{A}_k$.

For $k \in \{1, ..., n\}$ and $y, y' \in \Omega_k$ we define the difference operator $D^k_{y,y'} : \mathcal{A} \to \mathcal{A}_k$ by

$$D^k_{y,y'} f = S^k_y f - S^k_{y'} f \text{ for } f \in \mathcal{A}.$$

Clearly $D^k_{y,y'}$ annihilates $\mathcal{A}_k$. The operator $r_k : \mathcal{A} \to \mathcal{A}_k$, defined by $r_k f = \sup_{y,y' \in \Omega_k} D^k_{y,y'} f$ is called the k-th *conditional range*. We also use the abbreviations $\inf_k f = \inf_{y \in \Omega_k} S^k_y f$ and $\sup_k f = \sup_{y \in \Omega_k} S^k_y f$ for the conditional infimum and supremum.

Given the measures μ_k and $k \in \{1, ..., n\}$ we the operator $E_k : \mathcal{A} \to \mathcal{A}_k$ by

$$E_k f = E_{y \sim \mu_k}\left[S^k_y f\right] = \int_{\Omega_k} S^k_y f \, d\mu_k(y).$$

The operator $E_k[.] = E\left[.|X_1, ..., X_{k-1}, X_{k+1}, ..., X_n\right]$ is the expectation conditional to all variables with indices different to k. E_k is a linear projection onto $\mathcal{A}_k$. Also the E_k commute among each other, and for $h \in \mathcal{A}$ and $g \in \mathcal{A}_k$ we have

$$E\left[\left[E_k h\right] g\right] = E\left[E_k\left[hg\right]\right] = E\left[hg\right]. \tag{4}$$

Replacing the operator E by E_k leads to the definition of conditional thermodynamic quantities, all of which are now members of the algebra $\mathcal{A}_k$:

- The conditional partition function $Z_{k,\beta f} = E_k\left[e^{\beta f}\right]$,
- The conditional thermal expectation $E_{k,\beta f}[g] = Z_{k,\beta f}^{-1} E_k\left[g e^{\beta f}\right]$ for $g \in \mathcal{A}$,
- The conditional entropy $\mathrm{Ent}_{k,f}(\beta) = \beta E_{k,\beta f}[f] - \ln Z_{k,\beta f}$,
- The conditional thermal variance $\sigma_{k,\beta f}^2[g] = E_{k,\beta f}\left[\left(g - E_{k,\beta f}[g]\right)^2\right]$ for $g \in \mathcal{A}$. As $\beta \to 0$ this becomes
- The conditional variance $\sigma_k^2[g] = E_k\left[(g - E_k[g])^2\right]$ for $g \in \mathcal{A}$.

The previously established relations hold also for the corresponding conditional quantities. Of particular importance for our method is the conditional version of Theorem 7

$$\mathrm{Ent}_{k,f}(\beta) = \int_0^\beta \int_t^\beta \sigma_{k,sf}^2[f]\, ds\, dt.$$

The following lemma, which states that the conditional thermal expectation just behaves like a conditional expectation, will also be used frequently.

Lemma 8 *For any $f, g \in \mathcal{A}$, $k \in \{1, ..., n\}$, $\beta \in \mathbb{R}$*

$$E_{\beta f}\left[E_{k,\beta f}[g]\right] = E_{\beta f}[g].$$

Proof Using $E\left[E_k[h] g\right] = E\left[h E_k[g]\right]$

$$\begin{aligned} E_{\beta f}\left[E_{k,\beta f}[g]\right] &= Z_{\beta f}^{-1} E\left[E_k\left[g e^{\beta f}\right] \frac{e^{\beta f}}{E_k\left[e^{\beta f}\right]}\right] \\ &= Z_{\beta f}^{-1} E\left[g e^{\beta f} E_k\left[\left(\frac{e^{\beta f}}{E_k\left[e^{\beta f}\right]}\right)\right]\right] \\ &= Z_{\beta f}^{-1} E\left[g e^{\beta f}\right] \\ &= E_{\beta f}[g]. \end{aligned}$$

□

2.5 The Subadditivity of Entropy

In the non-interacting case, when the energy function f is a sum, $f = \sum f_k$, it is easily verified that $\mathrm{Ent}_{k,f}(\beta)(\mathbf{x}) = \mathrm{Ent}_{k,f}(\beta)$ is independent of $\mathbf{x}$ and that

$$\mathrm{Ent}_f(\beta) = \sum_{k=1}^{n} \mathrm{Ent}_{k,f}(\beta).$$

In statistical physics it is said that entropy is an extensive quantity: the entropy of non-interacting systems is equal to the sum of the individual entropies.

Equality no longer holds in the interacting, nonlinear case, but there is a subadditivity property which is sufficient for the purpose of concentration inequalities:

The total entropy is no greater than the thermal average of the sum of the conditional entropies.

Theorem 9 *For $f \in \mathcal{A}$ and $\beta > 0$*

$$Ent_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^{n} Ent_{k,f}(\beta)\right] \tag{5}$$

In 1975 Elliott Lieb [19] gave a proof of this result, which was probably known some time before, at least in the classical setting relevant to our arguments. Together with Theorem 5 and Theorem 7 it completes our basic toolbox to prove concentration inequalities. For the proof we need two auxiliary results.

Lemma 10 *Let $h, g > 0$ be bounded measurable functions on Ω. Then for any expectation E*

$$E[h]\ln\frac{E[h]}{E[g]} \leq E\left[h\ln\frac{h}{g}\right].$$

Proof Define an expectation functional E_g by $E_g[h] = E[gh]/E[g]$. The function $\Phi(t) = t\ln t$ is convex for positive t, since $\Phi'' = 1/t > 0$. Then

$$\Phi\left(E_g\left[\frac{h}{g}\right]\right) = \frac{E[h]}{E[g]}\ln\frac{E[h]}{E[g]}.$$

Thus, by Jensen's inequality,

$$\begin{aligned} E[h]\ln\frac{E[h]}{E[g]} &= E[g]\,E_g\left[\frac{h}{g}\right]\ln E_g\left[\frac{h}{g}\right] = E[g]\,\Phi\left(E_g\left[\frac{h}{g}\right]\right) \\ &\leq E[g]\,E_g\left[\Phi\left(\frac{h}{g}\right)\right] = E\left[h\ln\frac{h}{g}\right]. \end{aligned}$$

□

Next we prove (5) for general positive functions.

Lemma 11 *Let $\rho \in \mathcal{A}$, $\rho > 0$. Then*

$$E\left[\rho \ln \frac{\rho}{E[\rho]}\right] \leq \sum_k E\left[\rho \ln \frac{\rho}{E_k[\rho]}\right].$$

Proof Write $E^k[.] = E_1 E_2 ... E_k[.]$ with E^0 being the identity map on $\mathcal{A}$. The innocuous looking identity $E\left[E^k[.]\right] = E[.]$ is an obvious consequence of the fact that we work with product probabilities. Without independence it would not hold, and the following simple argument would break down. Note that $E^n = E$. We expand

$$\frac{\rho}{E[\rho]} = \frac{E^0[\rho]}{E^1[\rho]} \frac{E^1[\rho]}{E_2[\rho]} \cdots \frac{E^{n-1}[\rho]}{E^n[\rho]} = \prod_{k=1}^n \frac{E^{k-1}[\rho]}{E^{k-1}[E_k[\rho]]}.$$

We get from Lemma 10, using $E\left[E^{k-1}[.]\right] = E[.]$,

$$\begin{aligned} E\left[\rho \ln \frac{\rho}{E[\rho]}\right] &= \sum_k E\left[E^{k-1}[\rho] \ln \frac{E^{k-1}[\rho]}{E^{k-1}[E_k[\rho]]}\right] \\ &\leq \sum_k E\left[E^{k-1}\left[\rho \ln \frac{\rho}{E_k[\rho]}\right]\right] = \sum_k E\left[\rho \ln \frac{\rho}{E_k[\rho]}\right]. \end{aligned}$$

□

Finally we specialize to the canonical entropy.

Proof of Theorem 9 9 Set $\rho = e^{\beta f}$ in Lemma 11 to get

$$\begin{aligned} \mathrm{Ent}_f(\beta) &= Z_{\beta f}^{-1} E\left[e^{\beta f} \ln \frac{e^{\beta f}}{E\left[e^{\beta f}\right]}\right] \\ &\leq Z_{\beta f}^{-1} \sum_k E\left[e^{\beta f} \ln \frac{e^{\beta f}}{E_k\left[e^{\beta f}\right]}\right] \\ &= \sum_k E_{\beta f}\left[\beta f - \ln E_k\left[e^{\beta f}\right]\right] \\ &= E_{\beta f}\left[\sum_k \mathrm{Ent}_{k,f}(\beta)\right], \end{aligned}$$

where we used Lemma 8 in the last identity. □

2.6 Summary of Results

The exponential moment method, Corollary 3, and Theorems 5, 7, and 9 provide us with the tools to prove several useful concentration inequalities. Here is a summary:

Theorem 12 *For $f \in A$ and $\beta > 0$ we have*

$$\Pr\{f - Ef > t\} \leq E\left[e^{\beta(f-Ef)}\right]e^{-\beta t} \tag{6}$$

$$\ln E\left[e^{\beta(f-Ef)}\right] = \beta\int_0^\beta \frac{Ent_f(\gamma)}{\gamma^2}d\gamma \tag{7}$$

$$Ent_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^n Ent_{k,f}(\beta)\right] \tag{8}$$

$$Ent_f(\beta) = \int_0^\beta\int_t^\beta \sigma_{sf}^2[f]\,ds\,dt \tag{9}$$

$$Ent_{k,f}(\beta) = \int_0^\beta\int_t^\beta \sigma_{k,sf}^2[f]\,ds\,dt \tag{10}$$

Concatenating the exponential moment bound (6), the entropy representation of the moment generating function (7), the subadditivity of entropy (8) and the fluctuation representation of the conditional entropy (10), we obtain the following generic concentration inequality.

$$\Pr\{f - Ef > t\} \leq \inf_{\beta>0}\exp\left(\beta\int_0^\beta \gamma^{-2}E_{\gamma f}\left[\sum_{k=1}^n\int_0^\gamma\int_t^\gamma \sigma_{k,sf}^2[f]\,ds\,dt\right]d\gamma - \beta t\right).$$

This is the template for the results given in the next section.

3 First Applications of the Entropy Method

We now develop some first consequences of the method, beginning with the Efron–Stein inequality, a general bound on the variance. Then we continue with the derivation of the bounded difference inequality, a simple and perhaps the most useful concentration inequality, for which we give two illustrating applications. Then we give a Bennett-Bernstein type inequality which we apply to the concentration of vector-valued random variables.

3.1 The Efron–Stein Inequality

Combining the fluctuation representations (9) and (10) with the subadditivity (8) of entropy and dividing by β^2 we obtain

$$\frac{1}{\beta^2}\int_0^\beta \int_t^\beta \sigma_{sf}^2 [f]\, ds\, dt \leq E_{\beta f}\left[\sum_{k=1}^n \frac{1}{\beta^2}\int_0^\beta \int_t^\beta \sigma_{k,sf}^2 [f]\, ds\, dt.\right]$$

Using the continuity properties of $\beta \mapsto E_{\beta f}[g]$ and $\beta \mapsto \sigma_{\beta f}^2 [f]$, which follow from Lemma 6 we can take the limit as $\beta \to 0$ and multiply by 2 to obtain

$$\sigma^2 [f] \leq E\left[\sum_k \sigma_k^2 [f]\right] = E\left[\Sigma^2 (f)\right], \tag{11}$$

where we introduced the notation $\Sigma^2 (f) = \sum_k \sigma_k^2 [f]$ for the sum of conditional variances.

Equation (11) is the famous Efron–Stein–Steele inequality [26]. It is an easy exercise to provide the details of the above limit process and to extend the inequality to general functions $f \in L_2 [\mu]$ by approximation with a sequence of truncations.

3.2 The Bounded Difference Inequality

The variance of a bounded real random variable is never greater than a quarter of the square of its range.

Lemma 13 *If $f \in \mathcal{A}$ satisfies $a \leq f \leq b$ then $\sigma^2 [f] \leq (b-a)^2 /4$.*

Proof

$$\begin{aligned}\sigma^2 (f) &= E[(f - E[f])\, f] = E[(f - E[f])(f - a)] \\ &\leq E[(b - E[f])(f - a)] = (b - E[f])(E[f] - a) \\ &\leq \frac{(b-a)^2}{4}.\end{aligned}$$

To see the last inequality use calculus to find the maximal value of the function $t \to (b-t)(t-a)$. □.

The bounded difference inequality bounds the deviation of a function from its mean in terms of the *sum of squared conditional ranges*, which is the operator $R^2 : \mathcal{A} \to \mathcal{A}$ defined by

$$R^2(f) = \sum_{k=1}^{n} r_k(f)^2 = \sum_{k=1}^{n} \sup_{y,y' \in \Omega_k} \left(D^k_{y,y'} f\right)^2.$$

Theorem 14 (Bounded difference inequality) *For $f \in \mathcal{A}$ and $t > 0$*

$$\Pr\{f - Ef > t\} \le \exp\left(\frac{-2t^2}{\sup_{\mathbf{x} \in \Omega} R^2(f)(\mathbf{x})}\right).$$

Proof Applied to the conditional thermal variance Lemma 13 gives

$$\sigma^2_{k,sf}[f] \le \frac{1}{4} \sup_{y,y' \in \Omega_k} \left(D^k_{y,y'} f\right)^2 = \frac{1}{4} r_k(f)^2,$$

so combining the subadditivity of entropy (8) and the fluctuation representation (10) gives

$$\begin{aligned}
\mathrm{Ent}_f(\gamma) &\le E_{\gamma f}\left[\sum_{k=1}^{n} \mathrm{Ent}_{k,f}(\gamma)\right] = E_{\gamma f}\left[\sum_{k=1}^{n} \int_0^{\gamma} \int_t^{\gamma} \sigma^2_{k,sf}[f]\, ds\, dt\right] \\
&\le \frac{1}{4} E_{\gamma f}\left[\int_0^{\gamma} \int_t^{\gamma} \sum_{k=1}^{n} r_k(f)^2\right] ds\, dt \\
&= \frac{\gamma^2}{8} E_{\gamma f}\left[R^2(f)\right]. \qquad (12)
\end{aligned}$$

Bounding the thermal expectation $E_{\gamma f}$ by the supremum over $\mathbf{x} \in \Omega$ we obtain from Theorem 12 (7)

$$\ln E\left[e^{\beta(f - Ef)}\right] = \beta \int_0^{\beta} \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma \le \frac{\beta^2}{8} \sup_{\mathbf{x} \in \Omega} R^2(f)(\mathbf{x}),$$

and the tail bound (6) gives for all $\beta > 0$

$$\Pr\{f - Ef > t\} \le \exp\left(\frac{\beta^2}{8} \sup_{\mathbf{x} \in \Omega} R^2(f)(\mathbf{x}) - \beta t\right).$$

Substitution of the minimizing value $\beta = 4t / \left(\sup_{\mathbf{x} \in \Omega} R^2(f)(\mathbf{x})\right)$ completes the proof. □

Notice that the conditional range $r_k(f)$ is a function in $\mathcal{A}_k$ and may depend on all x_i except x_k. The sum $R^2(f) = \sum_{k=1}^{n} r_k(f)^2$ may thus depend on all the x_i. It is therefore a very pleasant feature that the supremum over $\mathbf{x}$ is taken *outside the sum*. In the sequel this will allow us to derive the Gaussian concentration inequality from Theorem 14. The bound (12) will be re-used in Sect. 5.4 to prove a version of the Hanson-Wright inequality for quadratic forms.

In the literature one often sees the following weaker version of Theorem 14.

Corollary 15 *For $f \in \mathcal{A}$ and $t > 0$*

$$\Pr\left\{f - Ef > t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{k=1}^{n} \sup_{\mathbf{x}\in\Omega} r_k\left(f\right)^2\left(\mathbf{x}\right)}\right).$$

If f is a sum $f = \sum_k X_k$, then r_k^2 is independent of $\mathbf{x}$ and the two results are equivalent. In this case we obtain the well known Hoeffding inequality [16].

Corollary 16 (Hoeffding's inequality) *Let X_k be real random variables $a_k \leq X_k \leq b_k$. Then*

$$\Pr\left\{\sum_k \left(X_k - E\left[X_k\right]\right) > t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{k=1}^{n}\left(b_k - a_k\right)^2}\right).$$

In returning to the general case of non-additive functions, it is remarkable that for many applications the following "little bounded difference inequality", which is yet weaker than Corollary 15, seems to be sufficient.

Corollary 17 *For $f \in \mathcal{A}$ and $t > 0$*

$$\Pr\left\{f - Ef > t\right\} \leq \exp\left(\frac{-2t^2}{nc^2}\right),$$

where

$$c = \max_k \sup_{\mathbf{x}\in\Omega, y, y'\in\Omega_k} D^k_{y,y'} f\left(\mathbf{x}\right).$$

3.3 Vector-Valued Concentration

Suppose the X_i are independent random variables with values in a normed space $\mathcal{B}$ such that $EX_i = 0$ and $\|X_i\| \leq c_i$. Let $\Omega_i = \{y \in \mathcal{B} : \|y\| \leq c_i\}$ and define $f : \prod_{i=1}^{n} \Omega_i \to \mathbb{R}$ by

$$f\left(\mathbf{x}\right) = \left\|\sum_i x_i\right\|.$$

Then by the triangle inequality, for y, y' with $\|y\|, \left\|y'\right\| \leq c_k$

$$\begin{aligned} D^k_{y,y'} f\left(\mathbf{x}\right) &= \left\| \sum_i S^k_y\left(\mathbf{x}\right)_i \right\| - \left\| \sum_i S^k_{y'}\left(\mathbf{x}\right)_i \right\| \\ &\le \left\| \sum_i S^k_y\left(x\right)_i - \sum_i S^k_{y'}\left(x\right)_i \right\| = \left\| y - y' \right\| \\ &\le 2c_k, \end{aligned}$$

so $R^2\left(f\right)\left(\mathbf{x}\right) \le 4\sum_i c_i^2$. It follows from Corollary 15 that

$$\Pr\left\{f - E\left[f\right] > t\right\} \le \exp\left(\frac{-t^2}{2\sum_i c_i^2}\right),$$

or that for $\delta > 0$ with probability at least $1 - \delta$ in $\left(X_1, ..., X_n\right)$

$$\left\| \sum_i X_i \right\| \le E\left\| \sum_i X_i \right\| + \sqrt{2\sum_i c_i^2 \ln\left(1/\delta\right)}. \tag{13}$$

If $\mathcal{B}$ is a Hilbert space we can bound $E\left\|\sum_i X_i\right\| \le \sqrt{\sum_i E\left[\left\|X_i\right\|^2\right]}$ by Jensen's inequality and if all the X_i are iid we get with probability at least $1 - \delta$

$$\left\| \frac{1}{n}\sum_i X_i \right\| \le \sqrt{\frac{E\left[\left\|X_1\right\|^2\right]}{n}} + c_1\sqrt{\frac{2\ln\left(1/\delta\right)}{n}} \tag{14}$$

3.4 Rademacher Complexities and Generalization

Now let $\mathcal{X}$ be any measurable space and $\mathcal{F}$ a countable class of functions $f : \mathcal{X} \to [0, 1]$ and $\mathbf{X} = \left(X_1, ..., X_n\right)$ be a vector of iid random variables with values in $\mathcal{X}$.

Empirical risk minimization really wants to find $f \in \mathcal{F}$ with minimal risk $E\left[f\left(X\right)\right]$, but, as the true distribution of X is unknown, it has to be content with minimizing the empirical surrogate

$$\frac{1}{n}\sum_i f\left(X_i\right).$$

One way to justify this method is by giving a bound on the uniform estimation error

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_i f\left(X_i\right) - E\left[f\left(X\right)\right]\right|.$$

The vector space

$$\mathcal{B} = \left\{ g : \mathcal{F} \to \mathbb{R} : \sup_{f \in \mathcal{F}} |g(f)| < \infty \right\}$$

becomes a normed space with norm $\|g\| = \sup_{f \in \mathcal{F}} |g(f)|$. For each X_i define $\hat{X}_i \in \mathcal{B}$ by $\hat{X}_i(f) = f(X_i) - E[f(X_i)]$. Then the $\hat{X}_i$ are zero mean random variables in $\mathcal{B}$ satisfying $\left\| \hat{X}_i \right\| \leq 1$, and (13) of the preceding section gives with probability at least $1 - \delta$

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i f(X_i) - E[f(X_i)] \right| \leq \frac{1}{n} E \sup_{f \in \mathcal{F}} \left| \sum_i f(X_i) - E[f(X_i)] \right| + \sqrt{\frac{2 \ln(1/\delta)}{n}}.$$

The expectation term on the right-hand side can be bounded in terms of *Rademacher complexity* [3]. This is the function $\mathcal{R} : \mathcal{F} \times \mathcal{X}^n \to \mathbb{R}$ defined as

$$\mathcal{R}(\mathcal{F}, \mathbf{x}) = \frac{2}{n} E_\epsilon \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i f(x_i) \right|,$$

where the $\epsilon = (\epsilon_1, ..., \epsilon_n)$ are vectors of independent Rademacher variables which are uniformly distributed on $\{-1, 1\}$. We have, with X_i' iid to X_i

$$\begin{aligned} \frac{1}{n} E \sup_{f \in \mathcal{F}} \left| \sum_i f(X_i) - E[f(X_i)] \right| &\leq \frac{1}{n} E_{\mathbf{X}\mathbf{X}'} \sup_{f \in \mathcal{F}} \left| \sum_i f(X_i) - f\left(X_i'\right) \right| \\ &= \frac{1}{n} E_{\mathbf{X}\mathbf{X}'} \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i \left(f(X_i) - f\left(X_i'\right) \right) \right|, \end{aligned}$$

for any $\epsilon \in \{-1, 1\}^n$, because the expectation is invariant under the interchange of X_i and X_i' on an arbitrary subset of indices. Passing to the expectation in ϵ and using the triangle inequality gives

$$\begin{aligned} \frac{1}{n} E \sup_{f \in \mathcal{F}} \left| \sum_i f(X_i) - E[f(X_i)] \right| &\leq \frac{1}{n} E_{\mathbf{X}\mathbf{X}'} E_\epsilon \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i \left(f(X_i) - f\left(X_i'\right) \right) \right| \\ &\leq \frac{2}{n} E_{\mathbf{X}} E_\epsilon \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i f(X_i) \right| \\ &= E_{\mathbf{X}} \mathcal{R}(\mathcal{F}, \mathbf{X}). \end{aligned}$$

Now we use the bounded difference inequality again to bound the deviation of $\mathcal{R}(\mathcal{F}, .)$ from its expectation. We have, again using the triangle inequality,

$$D^k_{y,y'}\mathcal{R}(\mathcal{F},\mathbf{x}) = \frac{2}{n}E_\epsilon\left[\sup_{f\in\mathcal{F}}\left|\sum_i \epsilon_i S^k_y f(x_i)\right| - \sup_{f\in\mathcal{F}}\left|\sum_i \epsilon_i S^k_{y'} f(x_i)\right|\right]$$
$$\leq \frac{2}{n}E_\epsilon\left[\sup_{f\in\mathcal{F}}\left|\epsilon_i\left(f(y) - f\left(y'\right)\right)\right|\right] \leq \frac{2}{n}$$

and thus obtain

$$\Pr\left\{E\left[\mathcal{R}(\mathcal{F},.)\right] > \mathcal{R}(\mathcal{F},.) + t\right\} \leq e^{-nt^2/2},$$

or, for every $\delta > 0$ with probability at least $1-\delta$

$$E\left[\mathcal{R}(\mathcal{F},\mathbf{X})\right] \leq \mathcal{R}(\mathcal{F},\mathbf{X}) + \sqrt{\frac{2\ln(1/\delta)}{n}}. \tag{15}$$

By a union bound we conclude that with probability at least $1-\delta$

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_i f(X_i) - E[f(X_i)]\right| \leq \mathcal{R}(\mathcal{F},\mathbf{X}) + 2\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

3.5 *The Bennett and Bernstein Inequalities*

The proof of the bounded difference inequality relied on bounding the thermal variance $\sigma^2_{k,\beta f}(f)$ uniformly in β, using the constraints on the conditional ranges of f. We now consider the case, where we only use one constraint on the ranges, say $f - E_k[f] \leq 1$, but we use information on the conditional variances. This leads to a Bennett type inequality as in [23]. Recall the notation for the sum of conditional variances $\Sigma^2(f) := \sum \sigma^2_k(f)$. Again we start with a bound on the thermal variance.

Lemma 18 *Assume $f - Ef \leq 1$. Then for $\beta > 0$*

$$\sigma^2_{\beta f}(f) \leq e^\beta \sigma^2(f)$$

Proof

$$\sigma^2_{\beta f}(f) = \sigma^2_{\beta(f-Ef)}(f-Ef) = E_{\beta(f-Ef)}\left[(f-Ef)^2\right] - \left(E_{\beta(f-Ef)}[f-Ef]\right)^2$$
$$\leq E_{\beta(f-Ef)}\left[(f-Ef)^2\right] = \frac{E\left[(f-Ef)^2 e^{\beta(f-Ef)}\right]}{E\left[e^{\beta(f-Ef)}\right]}$$
$$\leq E\left[(f-Ef)^2 e^{\beta(f-Ef)}\right] \text{ (by Jensen's inequality)}$$
$$\leq e^\beta E\left[(f-Ef)^2\right] \text{ (now using } f - Ef \leq 1\text{).}$$

□

Next we bound the entropy $\mathrm{Ent}_f(\beta)$.

Lemma 19 *Assume that $f - E_k f \leq 1$ for all $k \in \{1, ..., n\}$. Then for $\beta > 0$*

$$Ent_f(\beta) \leq \left(\beta e^{\beta} - e^{\beta} + 1\right) \; E_{\beta f}\left[\Sigma^2(f)\right].$$

Proof From Theorem 12 and the previous lemma we get

$$\mathrm{Ent}_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^{n} \int_0^{\beta} \int_t^{\beta} \sigma^2_{k,sf}[f]\, ds\, dt\right] \leq \int_0^{\beta} \int_t^{\beta} e^s ds\, dt\; E_{\beta f}\left[\Sigma^2(f)\right].$$

The conclusion follows from the formula

$$\int_0^{\beta} \int_t^{\beta} e^s ds\, dt = \int_0^{\beta} \left(e^{\beta} - e^t\right) dt = \beta e^{\beta} - e^{\beta} + 1.$$

□

We need one more technical Lemma.

Lemma 20 *For $x \geq 0$*

$$(1 + x) \ln(1 + x) - x \geq 3x^2/(6 + 2x).$$

Proof We have to show that

$$f_1(x) := \left(6 + 8x + 2x^2\right) \ln(1 + x) - 6x - 5x^2 \geq 0.$$

Since $f_1(0) = 0$ and $f_1'(x) = 4f_2(x)$ with $f_2(x) := (2 + x) \ln(1 + x) - 2x$, it is enough to show that $f_2(x) \geq 0$. But $f_2(0) = 0$ and $f_2'(x) = (1 + x)^{-1} + \ln(1 + x) - 1$, so $f_2'(0) = 0$, but $f_2''(x) = x(1 + x)^{-2} \geq 0$, so $f_2(x) \geq 0$. □

Now we can prove our version of Bennett's inequality.

Theorem 21 *Assume $f - E_k f \leq 1, \forall k$. Let $t > 0$ and denote $V = \sup_{\mathbf{x}\in\Omega} \Sigma^2(f)(\mathbf{x})$. Then*

$$\begin{aligned}\Pr\{f - E[f] > t\} &\leq \exp\left(-V\left(\left(1 + tV^{-1}\right) \ln\left(1 + tV^{-1}\right) - tV^{-1}\right)\right) \\ &\leq \exp\left(\frac{-t^2}{2V + 2t/3}\right).\end{aligned}$$

Proof Fix $\beta > 0$. We define the real function

$$\psi(t) = e^t - t - 1, \tag{16}$$

which arises from deleting the first two terms in the power series expansion of the exponential function and observe that

$$\int_0^\beta \frac{\gamma e^\gamma - e^\gamma + 1}{\gamma^2} d\gamma = \beta^{-1}\left(e^\beta - \beta - 1\right) = \beta^{-1}\psi(\beta),$$

because $(d/d\gamma)\left(\gamma^{-1}(e^\gamma - 1)\right) = \gamma^{-2}(\gamma e^\gamma - e^\gamma + 1)$ and $\lim_{\gamma \to 0} \gamma^{-1}(e^\gamma - 1) = 1$. Theorem 12 and Lemma 19 combined with a uniform bound then give

$$\begin{aligned}\ln E e^{\beta(f - Ef)} &= \beta \int_0^\beta \frac{\text{Ent}_f(\gamma)\, d\gamma}{\gamma^2} \\ &\leq \beta \left(\int_0^\beta \frac{\gamma e^\gamma - e^\gamma + 1}{\gamma^2} d\gamma\right) \sup_{\mathbf{x}\in\Omega} \Sigma^2(f)(\mathbf{x}) = \psi(\beta)\, V.\end{aligned}$$

It now follows from Theorem 12 that $\Pr\{f - E[f] > t\} \leq \exp(\psi(\beta)V - \beta t)$ for any $\beta > 0$. Substitution of $\beta = \ln\left(1 + tV^{-1}\right)$ gives the first inequality, the second follows from Lemma 20. □

Observe that f is assumed bounded above by the assumptions of the theorem. The existence of exponential moments $E\left[e^{\beta f}\right]$ is needed only for $\beta \geq 0$, so the assumption $f \in \mathcal{A}$ can be dropped in this case.

If f is additive the theorem reduces to the familiar Bennett and Bernstein inequalities [16].

Corollary 22 *Let X_k be real random variables $X_k \leq E[X_k] + 1$ and let $V = \sum_k \sigma^2(X_k)$. Then*

$$\begin{aligned}\Pr\left\{\sum_k (X_k - E[X_k]) > t\right\} &\leq \exp\left(-V\left(\left(1 + tV^{-1}\right)\ln\left(1 + tV^{-1}\right) - tV^{-1}\right)\right) \\ &\leq \exp\left(\frac{-t^2}{2V + 2t/3}\right).\end{aligned}$$

Theorem 21 and its corollary can be applied to functions satisfying $f - E_k[f] < b$ by a simple rescaling argument. Then Bernstein's inequality becomes

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega} \Sigma^2(f)(\mathbf{x}) + 2bt/3}\right).$$

Inequalities of this kind exhibit two types of tails, depending in which of the two terms in the denominator $A + Bt$ of the exponent is dominant. In the sub-Gaussian regime $A >> Bt$ the tail decays as $e^{-t^2/A}$. This is the way the bounded difference inequality behaves globally, but with a very crude approximation for A, while Bernstein's inequality uses variance information. But for larger deviations, when $A << Bt$, the tail only decays as $e^{-t/A}$. This subexponential behavior is absent in the bounded difference inequality and the price paid for the fine-tuning in Bernstein's inequality.

3.6 Vector-Valued Concentration Revisited

We look again at the situation of Sect. 3.3. Suppose again that the X_i are independent zero mean random variables with values in normed space, which we now assume to be a Hilbert space H, but that now we have a uniform bound $\|X_i\| \leq c$. Again we define $f : \{y \in H : \|y\| \leq c\}^n \to \mathbb{R}$ by $f(\mathbf{x}) = \left\|\sum_i x_i\right\|$ and observe that for $y, y' \in H$, $D^k_{y,y'} f(\mathbf{x}) \leq \left\|y - y'\right\|$. This implies that $f - E_k[f] \leq 2c$ and also

$$\sigma_k^2(f) = \frac{1}{2} E_{(y,y')\sim\mu_k^2} \left(D^k_{y,y'} f(\mathbf{x})\right)^2 \leq \frac{1}{2} E_{(y,y')\sim\mu_k^2} \left\|y - y'\right\|^2 = E\|X_k\|^2 .$$

Thus $\Sigma^2(f) \leq \sum_i E\|X_i\|^2$ and by Bernstein's inequality, Theorem 21,

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\sum_i E\|X_i\|^2 + 4ct/3}\right),$$

or that for $\delta > 0$ with probability at least $1 - \delta$ in $(X_1, ..., X_n)$

$$\left\|\sum_i X_i\right\| \leq \sqrt{\sum_i E\left[\|X_i\|^2\right]} + \sqrt{2\sum_i E\|X_i\|^2 \ln(1/\delta)} + 4c\ln(1/\delta)/3,$$

where we again used Jensen's inequality to bound $E\left\|\sum_i X_i\right\|$. If all the X_i are iid we get with probability at least $1 - \delta$

$$\left\|\frac{1}{n}\sum_i X_i\right\| \leq \sqrt{\frac{E\left[\|X_1\|^2\right]}{n}}\left(1 + \sqrt{2\ln(1/\delta)}\right) + \frac{4c\ln(1/\delta)}{2n}.$$

If the variance $E\left[\|X_1\|^2\right]$ is small and n is large, this is much better than the bound (14), which we got from the bounded difference inequality.

4 Inequalities for Lipschitz Functions and Dimension Free Bounds

We now prove some more advanced concentration inequalities. First we will use the bounded difference inequality to prove a famous sub-gaussian bound for Lipschitz functions of independent standard normal variables. We then derive an exponential Efron–Stein inequality which allows to prove a similar result for convex Lipschitz functions on $[0, 1]^n$. We also obtain a concentration inequality for the operator norm of a random matrix, with deviations independent of the size of the matrix.

4.1 Gaussian Concentration

The advantage of the bounded difference inequality, Theorem 14, over its simplified Corollary 15 is the supremum over $\mathbf{x}$ outside the sum over k. This allows us to prove the following powerful Gaussian concentration inequality (Tsirelson-Ibragimov–Sudakov inequality, Theorem 5.6 in [6]). We assume $\Omega_k = \mathbb{R}$ and μ_k to be the distribution of a standard normal variable, and we require f to be an L-Lipschitz function, which means that for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$

$$f(\mathbf{x}) - f\left(\mathbf{x}'\right) \leq L \left\|\mathbf{x} - \mathbf{x}'\right\|,$$

where $\|.\|$ is the Euclidean norm on $\mathbb{R}^n$.

Theorem 23 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz and let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent standard normal variables. Then for any $s > 0$*

$$\Pr\{f(\mathbf{X}) > \mathbb{E}f(\mathbf{X}) + s\} \leq e^{-s^2/2L^2}.$$

Note that the function f is not assumed to be bounded on $\mathbb{R}^n$.

Proof The idea of the proof is to use the central limit theorem to approximate the X_i by appropriately scaled Rademacher sums $h_K(\epsilon_i)$ and to apply the bounded difference inequality to $f(h_K(\epsilon_1), ..., h_K(\epsilon_n))$.

By an approximation argument using convolution with Gaussian kernels of decreasing width it suffices to prove the result if f is C^2 with $\left|\left(\partial^2/x_i^2\right) f(\mathbf{x})\right| \leq B$ for all $\mathbf{x} \in \mathbb{R}^n$ and $i \in \{1, ..., n\}$, where B is a finite, but arbitrarily large constant. For $K \in \mathbb{N}$ define a function $h_K : \{-1, 1\}^K \to \mathbb{R}$, a vector-valued function $\mathbf{h}_K : \{-1, 1\}^{Kn} \to \mathbb{R}^n$ and a function $G_K : \{-1, 1\}^{Kn} \to \mathbb{R}$ by

$$\begin{aligned} h_K(\epsilon) &= \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \epsilon_k, \text{ for } \epsilon \in \{-1, 1\}^K \\ \mathbf{h}_K(\epsilon) &= (h_K(\epsilon_1), ..., h_K(\epsilon_n)) \text{ for } \epsilon = (\epsilon_1, ..., \epsilon_n) \in \{-1, 1\}^{Kn} \\ G_K &= f(\mathbf{h}_K(\epsilon)) \text{ for } \epsilon \in \{-1, 1\}^{Kn}. \end{aligned}$$

We will use Theorem 14 on the function G_K applied to independent Rademacher variables ϵ.

Fix an arbitrary configuration $\epsilon \in \{-1, 1\}^{Kn}$ and let $\mathbf{x} = (x_1, ..., x_n) = \mathbf{h}_K(\epsilon)$. For each $i \in \{1, ..., n\}$ we introduce the real function $f_i(t) = S_t^i f(\mathbf{x})$, so that we replace the i-th argument x_i by t, leaving all the other x_j fixed. Since f is C^2 we have for any $t \in \mathbb{R}$

$$f_i(x + t) - f_i(x) = t f_i'(x) + \frac{t^2}{2} f_i''(s)$$

for some $s \in \mathbb{R}$, and by the Lipschitz condition and the bound on $\left|f_i''\right|$

$$\begin{aligned}(f_i(x+t) - f_i(x))^2 &= t^2\left(f_i'(x)\right)^2 + t^3 f_i'(x) f_i''(s) + \frac{t^4}{4}\left(f_i''(s)\right)^2 \\ &\leq t^2\left(f_i'(x)\right)^2 + |t|^3 LB + \frac{t^4}{4}B^2.\end{aligned}$$

Now fix a pair of indices (i, k) with $i \in \{1, ..., n\}$ and $k \in \{1, ..., K\}$ and arbitrary values $y, y' \in \{-1, 1\}$ with $y \neq y'$. We want to bound $\left(D_{y,y'}^{(i,k)} G_K(\epsilon)\right)^2$. Now either one of y or y' is equal to ϵ_{ik}, so either $S_y^{(i,k)} G_K(\epsilon)$ or $S_{y'}^{(i,k)} G_K(\epsilon)$ is equal to $G_K(\epsilon)$. Without loss of generality we assume the second. Furthermore $S_y^k h_K(\epsilon_i)$ and $h_K(\epsilon_i)$ differ by at most $2/\sqrt{K}$, so

$$\begin{aligned}\left(D_{y,y'}^{(i,k)} G_K(\epsilon)\right)^2 &= \left(f\left(x_1, ..., S_y^k h_K(\epsilon_i), ..., x_n\right) - f(x_1, ..., h_K(\epsilon_i), ..., x_n)\right)^2 \\ &= \left(f_i\left(h_K(\epsilon_i) \pm \frac{2}{\sqrt{K}}\right) - f_i(h_K(\epsilon_i))\right)^2 \\ &\leq \frac{4 f_i'(h_K(\epsilon_i))^2}{K} + \frac{8LB}{K^{3/2}} + \frac{4B^2}{K^2}.\end{aligned}$$

Now $f_i'(h_K(\epsilon_i))$ is just equal to $(\partial/\partial x_i) f(\mathbf{x})$, so

$$\sum_i f_i'(h_K(\epsilon_i))^2 \leq \sup_{\mathbf{x}\in\mathbb{R}^n} \|\nabla f(\mathbf{x})\|^2 \leq L^2.$$

Since ϵ was arbitrary we have

$$\sup_\epsilon \sum_{k,i} \sup_{y,y'} \left(D_{y,y'}^{(i,k)} G_K(\epsilon)\right)^2 \leq 4L^2 + \frac{8nLB}{K^{1/2}} + \frac{4nB^2}{K}.$$

From Theorem 14 we conclude from $f(\mathbf{h}_K(\epsilon)) = G_K(\epsilon)$ that

$$\Pr\left\{f(\mathbf{h}_K(\epsilon)) - \mathbb{E} f\left(\mathbf{h}_K\left(\epsilon'\right)\right) > s\right\} \leq \exp\left(\frac{-s^2}{2L^2 + 4nLB/K^{1/2} + 2nB^2/K}\right).$$

The conclusion now follows from the central limit theorem since $h_K(\epsilon) \to \mathbf{X}$ weakly as $K \to \infty$. □

4.2 Exponential Efron Stein Inequalities

We will now use the entropy method to derive some other "dimension free" bounds of this type. We need the following very useful result.

Lemma 24 (Chebychev's association inequality) *Let g and h be real functions, X a real random variable.*

If g and h are either both nondecreasing or both nonincreasing then

$$E\left[g\left(X\right)h\left(X\right)\right]\geq E\left[g\left(X\right)\right]E\left[h\left(X\right)\right].$$

If either one of g or h is nondecreasing and the other nonincreasing then

$$E\left[g\left(X\right)h\left(X\right)\right]\leq E\left[g\left(X\right)\right]E\left[h\left(X\right)\right].$$

Proof Let X' be a random variable iid to X. Then

$$E\left[g\left(X\right)h\left(X\right)\right]-E\left[g\left(X\right)\right]E\left[h\left(X\right)\right]=\frac{1}{2}E\left[\left(g\left(X\right)-g\left(X'\right)\right)\left(h\left(X\right)-h\left(X'\right)\right)\right].$$

Now if g and h are either both nondecreasing or both nonincreasing then

$$\left(g\left(X\right)-g\left(X'\right)\right)\left(h\left(X\right)-h\left(X'\right)\right)$$

is always nonnegative, because both factors always have the same sign, in the other case it is always nonpositive. □

We use this inequality to prove a bound on the thermal variance. First recall that for two iid random variables X and X' we have

$$\begin{aligned}\sigma^2\left(X\right)&=\frac{1}{2}E_{XX'}\left[\left(X-X'\right)^2\right]\\&=\frac{1}{2}E_{XX'}\left[\left(X-X'\right)^2 1_{X>X'}\right]+\frac{1}{2}E_{XX'}\left[\left(X-X'\right)^2 1_{X<X'}\right]\\&=E_{XX'}\left[\left(X-X'\right)_+^2\right].\end{aligned}$$

Lemma 25 *Let* $0\leq s\leq\beta$. *Then*

$$\sigma_{sf}^2\left(f\right)\leq E_{x\sim\mu_{\beta f}}\left[E_{x'\sim\mu}\left[\left(f\left(x\right)-f\left(x'\right)\right)_+^2\right]\right].$$

Proof Let ψ be any real function. Lemma 6 (2) gives

$$\frac{d}{d\beta}E_{\beta f}\left[\psi\left(f\right)\right]=E_{\beta f}\left[\psi\left(f\right)f\right]-E_{\beta f}\left[\psi\left(f\right)\right]E_{\beta f}\left[f\right]. \tag{17}$$

By Chebychev's association inequality $E_{\beta f}\left[\psi\left(f\right)\right]$ is nonincreasing (nondecreasing) in β if ψ is nonincreasing (nondecreasing). Now define $g:\mathbb{R}^2\rightarrow\mathbb{R}$ by

$$g\left(s,t\right)=E_{x\sim\mu_{sf}}\left[E_{x'\sim\mu_{tf}}\left[\left(f\left(x\right)-f\left(x'\right)\right)^2 1_{f(x)\geq f(x')}\right]\right],$$

so that

$$\sigma^2_{sf}\left(f\right)=\frac{1}{2}E_{x\sim\mu_{sf}}\left[E_{x'\sim\mu_{sf}}\left[\left(f\left(x\right)-f\left(x'\right)\right)^2\right]\right]=g\left(s,s\right).$$

Now for fixed x the function $\left(f\left(x\right)-f\left(x'\right)\right)^2 1_{f(x)\geq f(x')}$ is nonincreasing in $f\left(x'\right)$, so $g\left(s,t\right)$ is nonincreasing in t. On the other hand, for fixed x', $\left(f\left(x\right)-f\left(x'\right)\right)^2$ $1_{f(x)\geq f(x')}$ is nondecreasing in $f\left(x\right)$, so $g\left(s,t\right)$ is nondecreasing in s (this involves exchanging the two expectations in the definition of $g\left(s,t\right)$). So, since $\mu_{0f}=\mu$, we get from $0\leq s\leq\beta$ that

$$\sigma^2_{sf}\left(f\right)=g\left(s,s\right)\leq g\left(\beta,0\right)=E_{x\sim\mu_{\beta f}}\left[E_{x'\sim\mu}\left[\left(f\left(x\right)-f\left(x'\right)\right)^2_+\right]\right].$$

□

Here is another way to write the conclusion: let $h\in\mathcal{A}$ be defined by $h\left(x\right)=E_{x'\sim\mu}\left[\left(f\left(x\right)-f\left(x'\right)\right)^2_+\right]$. Then $\sigma^2_{sf}\left(f\right)\leq E_{\beta f}\left[h\right]$.

Define two operators $D^2:\mathcal{A}\rightarrow\mathcal{A}$ and $V^2_+:\mathcal{A}\rightarrow\mathcal{A}$ by

$$D^2 f=\sum_k\left(f-\inf_{y\in\Omega_k}S^k_y f\right)^2$$

$$\text{and } V^2_+ f=\sum_k E_{y\sim\mu_k}\left[\left(\left(f-S^k_y f\right)_+\right)^2\right].$$

Clearly $V^2_+ f\leq D^2 f$ as $D^2 f$ is obtained by bounding the expectations in the definition of V^2_+ by their suprema.

Lemma 26 *For $\beta>0$ and $f\in\mathcal{A}$*

$$Ent_f\left(\beta\right)\leq\frac{\beta^2}{2}E_{\beta f}\left[V^+\left(f\right)\right].$$

Proof For $k\in\{1,...,n\}$ write $h_k=E_{y\sim\mu_k}\left[\left(f-S^k_y f\right)^2_+\right]$, so that $V^+\left(f\right)=\sum_k h_k$. The conditional version of Lemma 25 then reads for $0\leq s\leq\beta$ and $k\in\{1,...,n\}$

$$\sigma^2_{k,sf}\left(f\right)\leq E_{k,\beta f}\left[h_k\right].$$

Theorem 12 gives

$$\begin{aligned}\mathrm{Ent}_f(\beta) &\leq \int_0^\beta \int_t^\beta \sum_k E_{\beta f}\left[\sigma_{k,sf}^2(f)\right] ds dt\\ &\leq \int_0^\beta \int_t^\beta \sum_k E_{\beta f}\left[E_{k,\beta f}[h_k]\right] ds dt\\ &= \int_0^\beta \int_t^\beta \sum_k E_{\beta f}[h_k] ds dt\\ &= \frac{\beta^2}{2} E_{\beta f}\left[V^+(f)\right],\end{aligned}$$

where we used the identity $E_{\beta f}\left[E_{k,\beta f}[h]\right] = E_{\beta f}[h]$ for $h \in \mathcal{A}$. □

The usual arguments involving Theorem 12 and an optimization in β now immediately lead to

Theorem 27 *With $t > 0$*

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega} V_+^2 f(\mathbf{x})}\right) \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega} D^2 f(\mathbf{x})}\right).$$

We get a corresponding lower tail bound only for D^2 and we have to use an estimate similar to what was used in the proof of Bennett's inequality.

Lemma 28 *If $f - \inf_k f \leq 1, \forall k$ then for $\beta > 0$*

$$Ent_{-f}(\beta) \leq \psi(\beta) E_{-\beta f}\left[D^2 f\right],$$

with $\psi(t) = e^t - t - 1$ defined as in (16).

Proof Let $k \in \{1, ..., n\}$. We write $h_k := f - \inf_k f$. Then $h_k \in [0, 1]$ and for $s \leq \beta$ we have $1 \leq e^{(\beta-s)h_k} \leq e^{\beta-s}$, so

$$E_{k,-sh_k}\left[h_k^2\right] = \frac{E_k\left[h_k^2 e^{-\beta h_k} e^{(\beta-s)h_k}\right]}{E_k\left[e^{-\beta h_k} e^{(\beta-s)h_k}\right]} \leq e^{(\beta-s)} \frac{E_k\left[h_k^2 e^{-\beta h_k}\right]}{E_k\left[e^{-\beta h_k}\right]} = e^{(\beta-s)} E_{k,-\beta h_k}\left[h_k^2\right].$$

We therefore have

$$\begin{aligned}\int_0^\beta \int_t^\beta E_{k,-sf}\left[h_k^2\right] ds\, dt &= \int_0^\beta \int_t^\beta E_{k,-sh_k}\left[h_k^2\right] ds\, dt\\ &\leq \left(\int_0^\beta \int_t^\beta e^{\beta-s} ds\, dt\right) E_{k,-\beta h_k}\left[h_k^2\right] = \psi(\beta) E_{k,-\beta f}\left[h_k^2\right],\end{aligned}$$

where we used the formula

$$\int_0^\beta \int_t^\beta e^{-s} ds\, dt = 1 - e^{-\beta} - \beta e^{-\beta}.$$

Thus, using Theorem 12 and the identity $E_{-\beta f}E_{k,-\beta f}=E_{-\beta f}$,

$$\begin{aligned}\mathrm{Ent}_{-f}(\beta) &\leq E_{-\beta f}\left[\sum_k\int_0^\beta\int_t^\beta \sigma^2_{k,-sf}[f]\,ds\,dt\right]\leq E_{-\beta f}\left[\sum_k\int_0^\beta\int_t^\beta E_{k,-sf}\left[h_k^2\right]ds\,dt\right]\\ &\leq \psi(\beta)\,E_{-\beta f}\left[\sum_k E_{k,-\beta f}\left[h_k^2\right]\right]=\psi(\beta)\,E_{-\beta f}\left[D^2 f\right].\end{aligned}$$

□

Lemmas 26 and 28 together with (7) imply the inequalities

$$\ln E\left[e^{\beta(f-E[f])}\right]\leq\frac{\beta}{2}\int_0^\beta E_{\gamma f}\left[V_+^2 f\right]d\gamma. \tag{18}$$

and, if $f-\inf_k f\leq 1$ for all k, then

$$\ln E\left[e^{\beta(E[f]-f)}\right]\leq\frac{\psi(\beta)}{\beta}\int_0^\beta E_{-\gamma f}\left[D^2 f\right]d\gamma, \tag{19}$$

where in the last inequality we also used the fact that $\gamma\mapsto\psi(\gamma)/\gamma^2$ is nondecreasing. Bounding the thermal expectation with the uniform norm and substitution of $\beta=\ln\left(1+t\left\|D^2 f\right\|_\infty^{-1}\right)$ gives the following lower tail bound as in the proof of the Bennett-Bernstein inequalities.

Theorem 29 *If $f-\inf_k f\leq 1$ for all k, then for $t>0$ and with $\Delta:=\sup_{\mathbf{x}\in\Omega}D^2 f(\mathbf{x})$*

$$\begin{aligned}\Pr\{Ef-f>t\}&\leq\exp\left(-\Delta\left(\left(1+\frac{t}{\Delta}\right)\ln\left(1+\frac{t}{\Delta}\right)-\frac{t}{\Delta}\right)\right)\\ &\leq\exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega}D^2 f(\mathbf{x})+2t/3}\right).\end{aligned}$$

4.3 Convex Lipschitz Functions

In Sect. 4.1 we gave a sub-gaussian bound for Lipschitz functions of independent standard normal variables. Now we prove the same upper tail bound under different hypotheses. Instead of assuming μ_k to be standard normal we require $\Omega_k=[0,1]$ and let μ_k be perfectly arbitrary. On the other hand, in addition to being an L-Lipschitz function, we require f to be convex (actually only separately convex in each argument).

Theorem 30 *Let $\Omega_k=\mathcal{I}$, an interval of unit diameter, and let $f\in\mathcal{A}$ be C^1, L-Lipschitz and such that $y\in[0,1]\mapsto S_y^k f(\mathbf{x})$ is convex for all k and all $\mathbf{x}$. Then*

$$\Pr\{f - Ef > t\} \leq e^{-t^2/2L^2}.$$

Proof By an approximation argument we can assume f to be differentiable. Let $\mathbf{x} \in [0,1]^n$, $k \in \{1, ..., n\}$ and $y \in [0,1]$ such that $S_y^k f(\mathbf{x}) \leq f(\mathbf{x})$. Then, using separate convexity,

$$f(\mathbf{x}) - S_y^k f(\mathbf{x}) \leq \left\langle \mathbf{x} - S_y^k \mathbf{x}, \partial f(\mathbf{x}) \right\rangle_{\mathbb{R}^n} = (x_k - y) \frac{\partial}{\partial x_k} f(\mathbf{x}) \leq \left| \frac{\partial}{\partial x_k} f(\mathbf{x}) \right|.$$

We therefore have $f(\mathbf{x}) - \inf_y S_y^k f(\mathbf{x}) \leq |(\partial/\partial x_k) f(\mathbf{x})|$ and

$$D^2 f(\mathbf{x}) = \sum_{k=1}^n \left(f(\mathbf{x}) - \inf_y S_y^k f(\mathbf{x}) \right)^2 \leq \|\nabla f(\mathbf{x})\|_{\mathbb{R}^n}^2 \leq L^2.$$

Theorem 27 then gives the conclusion. □

For future reference we record the following fact: if Ω_k is an interval of unit diameter and A an $m \times n$-matrix then $\mathbf{x} \mapsto \|A\mathbf{x}\|$ is a convex function with Lipschitz constant $\|A\|$ and thus

$$D^2(\|A\mathbf{x}\|) \leq \|A\|^2. \tag{20}$$

4.4 The Operator Norm of a Random Matrix

For $\mathbf{x} \in [-1,1]^{n^2}$ let $M(\mathbf{x})$ be the $n \times n$ matrix whose entries are given by the components of $\mathbf{x}$. We are interested in the concentration properties of the operator norm of $M(\mathbf{X})$, when $\mathbf{X}$ is a vector with independent, but possibly not identically distributed components chosen from $[-1,1]$. The function in question is then $f : [-1,1]^{n^2} \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \|M(\mathbf{x})\| = \sup_{\|w\|, \|v\|=1} \langle M(\mathbf{x}) v, w \rangle,$$

where $\langle ., . \rangle$ and $\|.\|$ refer to inner product and norm in $\mathbb{R}^n$.

To bound $D^2 f(\mathbf{x})$ first let $\mathbf{x} \in [-1,1]^{n^2}$ be arbitrary but fixed, and let v and w be unit vectors witnessing the supremum in the definition of $f(\mathbf{x})$.

Now let (k,l) be any index to a matrix entry and choose any $y \in [-1,1]$ such that $S_y^{(k,l)} f(\mathbf{x}) \leq f(\mathbf{x})$. Then

$$\begin{aligned} f(\mathbf{x}) - S_y^{(k,l)} f(\mathbf{x}) &= \langle M(\mathbf{x}) v, w \rangle - \sup_{\|w'\|, \|v'\|=1} \left\langle S_y^{(k,l)} M(\mathbf{x}) v', w' \right\rangle \\ &\leq \left\langle \left(M(\mathbf{x}) - S_y^{(k,l)} M(\mathbf{x}) \right) v, w \right\rangle = (x_{kl} - y) v_k w_l \\ &\leq 2 |v_k| |w_l|. \end{aligned}$$

Note that $f - \inf_k f \leq 2$. Also

$$D^2 f(\mathbf{x}) = \sum_{k,l} \left(f(\mathbf{x}) - \inf_{y \in [-1,1]} S_y^{(k,l)} f(\mathbf{x}) \right)^2$$
$$\leq 4 \sum_{k,l} |v_k|^2 |w_l|^2 = 4.$$

The results of the previous section (rescaling for the lower tail to get $f - \inf_k f \leq 1$) then lead to a concentration inequality independent of the size of the random matrix.

Theorem 31 *Let* $\mathbf{X} = \left(X_{ij}\right)_{1 \leq i,j \leq n}$ *be a vector of* n^2 *independent random variables with values in* $[-1, 1]$*, and* $\mathbf{X}'$ *iid to* $\mathbf{X}$. *Then for* $t > 0$.

$$\Pr\left\{\|M(\mathbf{X})\| - E\left[\left\|M\left(\mathbf{X}'\right)\right\|\right] \geq t\right\} \leq \exp\left(\frac{-t^2}{8}\right)$$

and

$$\Pr\left\{E\left[\left\|M\left(\mathbf{X}'\right)\right\|\right] - \|M(\mathbf{X})\| \geq t\right\} \leq \exp\left(\frac{-t^2}{8 + 4t/3}\right).$$

Observe that the argument depends on the fact that the unit vectors v and w could be fixed independent of k and l. This would not have been possible with the bounded difference inequality. Also note that square matrices were chosen for notational convenience only. The same proof would work for rectangular matrices.

5 Beyond Uniform Bounds

All of the above applications of the entropy method to derive upper tail bounds involved an inequality of the form

$$\mathrm{Ent}_f(\gamma) \leq \xi(\gamma) E_{\gamma f}[G(f)],$$

where ξ is some nonnegative real function and G is some operator $G : \mathcal{A} \to \mathcal{A}$, which is positively homogeneous of order two. For the bounded difference inequality $\xi(\gamma) = \gamma^2/8$ and $G = R^2$, for the Bennett inequality $\xi(\gamma) = \gamma e^\gamma - e^\gamma + 1$ and $G = \Sigma^2$, for Theorem 27 we had $\xi(\gamma) = \gamma^2/2$ and $G = V_+^2$. Theorem 12 is then used to conclude that

$$\ln E e^{\beta(f - Ef)} \leq \beta \int_0^\beta \frac{\xi(\gamma)}{\gamma^2} E_{\gamma f}[G(f)] \, d\gamma \leq \beta \sup_{\mathbf{x}} G(f)(\mathbf{x}) \int_0^\beta \frac{\xi(\gamma) \, d\gamma}{\gamma^2}. \tag{21}$$

An analogous strategy was employed for the various lower tail bounds.

The uniform estimate $E_{\gamma f}\left[G\left(f\right)\right] \leq \sup_{\mathbf{x}} G\left(f\right)\left(\mathbf{x}\right)$ in (21), while being very simple, is somewhat loose and can sometimes be avoided by exploiting special properties of the thermal expectation and the function in question.

5.1 Self-boundedness

The first possibility we consider is that the function $G\left(f\right)$ can be bounded in terms of the function f itself, a property referred to as *self-boundedness* [8]. For example, if simply $G\left(f\right) \leq f$, then $E_{\gamma f}\left[G\left(f\right)\right] \leq E_{\gamma f}\left[f\right] = \left(d/d\gamma\right)\ln Z_{\gamma f}$, and if the function ξ has some reasonable behavior, then the first integral in (21) above can be bounded by partial integration or even more easily. As an example we apply this idea in the setting of Theorems 27 and 29.

Lemma 32 *Suppose that for* $f \in \mathcal{A}$ *there are nonnegative numbers* a, b *such that*
(i) $V_+^2 f \leq af + b$. *Then for* $0 \leq \beta < 2/a$

$$\ln E\left[e^{\beta(f-E[f])}\right] \leq \frac{\beta^2\left(aEf + b\right)}{2 - a\beta},$$

(ii) $D^2 f \leq af + b$. *If in addition* $f - \inf_k f \leq 1$ *for all k, then for* $\beta < 0$ *and* $a \geq 1$

$$\ln E\left[e^{\beta(E[f]-f)}\right] \leq \frac{\beta^2\left(aE\left[f\right] + b\right)}{2}.$$

Proof (i) We use (18) and get

$$\begin{aligned}\ln E\left[e^{\beta(f-E[f])}\right] &\leq \frac{\beta}{2}\int_0^\beta E_{\gamma f}\left[V_+^2 f\right]d\gamma \leq \frac{a\beta}{2}\int_0^\beta E_{\gamma f}\left[f\right]d\gamma + \frac{b\beta^2}{2}\\ &= \frac{a\beta}{2}\ln Z_{\beta f} + \frac{b\beta^2}{2},\end{aligned}$$

where the last identity follows from the fact that $E_{\gamma f}\left[f\right] = \left(d/d\gamma\right)\ln Z_{\gamma f}$. Thus

$$\ln E\left[e^{\beta(f-E[f])}\right] \leq \frac{a\beta}{2}\ln Ee^{\beta(f-E[f])} + \frac{a\beta^2}{2}Ef + \frac{b\beta^2}{2},$$

and rearranging this inequality for $\beta \in (0, 2/a)$ establishes the claim.

(ii) We use (19)

$$\begin{aligned}\ln E\left[e^{\beta(E[f]-f)}\right] &\leq \frac{\psi(\beta)}{\beta}\int_0^\beta E_{-\gamma f}\left[D^2 f\right]d\gamma \\ &\leq \frac{a\psi(\beta)}{\beta}\int_0^\beta E_{-\gamma f}[f]\,d\gamma + b\psi(\beta) = \frac{-a\psi(\beta)}{\beta}\ln Z_{-\beta f} + b\psi(\beta) \\ &= \frac{-a\psi(\beta)}{\beta}\ln E\left[e^{\beta(E[f]-f)}\right] + \psi(\beta)\left(aE[f]+b\right).\end{aligned}$$

Rearranging gives

$$\ln E\left[e^{\beta(E[f]-f)}\right] \leq \frac{\psi(\beta)}{1+a\beta^{-1}\psi(\beta)}\left(aE[f]+b\right) \leq \frac{\beta^2\left(aE[f]+b\right)}{2},$$

where one verifies that for $\beta > 0$ and $a \geq 1$ we have $\psi(\beta)\left(1+a\beta^{-1}\psi(\beta)\right)^{-1} \leq \beta^2/2$. □

The bound in part (i) requires an upper bound on β. To proceed we need the following optimization lemma, which will be used several times in the sequel and leads to tail bounds with both sub-Gaussian and subexponential regimes, similar to Bernstein's inequality.

Lemma 33 *Let C and b denote two positive real numbers, $t > 0$. Then*

$$\inf_{\beta\in[0,1/b)}\left(-\beta t + \frac{C\beta^2}{1-b\beta}\right) \leq \frac{-t^2}{2(2C+bt)}. \tag{22}$$

Proof Let $h(t) = 1+t-\sqrt{1+2t}$. Then use

$$\begin{aligned}2h(t)(1+t) &= 2(1+t)^2 - 2(1+t)\sqrt{1+2t} \\ &= (1+t)^2 - 2(1+t)\sqrt{1+2t} + (1+2t) + t^2 \\ &= \left(1+t-\sqrt{1+2t}\right)^2 + t^2 \\ &\geq t^2,\end{aligned}$$

so that

$$h(t) \geq \frac{t^2}{2(1+t)}. \tag{23}$$

Substituting

$$\beta = \frac{1}{b}\left(1-\left(1+\frac{bt}{C}\right)^{-1/2}\right)$$

in the left side of (22) we obtain

$$\inf_{\beta\in[0,1/b)}\left(-\beta t+\frac{C\beta^2}{1-b\beta}\right)\leq-\frac{2C}{b^2}h\left(\frac{bt}{2C}\right)\leq\frac{-t^2}{2\left(2C+bt\right)},$$

where we have used (23). □

Theorem 34 *Suppose for $f\in A$ there are nonnegative numbers a, b such that*
(i) $V_+^2 f\leq af+b$. Then for $t>0$ we have

$$\Pr\left\{f-E\left[f\right]>t\right\}\leq\exp\left(\frac{-t^2}{2\left(aE\left[f\right]+b+at/2\right)}\right).$$

(ii) $D^2 f\leq af+b$. If in addition, $a\geq 1$ and $f-\inf_k f\leq 1, \forall k\in\{1,...,n\}$, then

$$\Pr\left\{E\left[f\right]-f>t\right\}\leq\exp\left(\frac{-t^2}{2\left(aE\left[f\right]+b\right)}\right).$$

Proof Part (i) follows from Lemmas 32 (i) and Lemma 33). Part (ii) is immediate from Lemma 32 (ii). □

Boucheron et al. [8] have given a refined version for the lower tail, where the condition $a\geq 1$ is relaxed to $a\geq 1/3$ for the lower tail. There they also show that Theorems 34 and 27 together suffice to derive a version of the convex distance inequality which differs from Talagrand's original result only in that it has an inferior constant in the exponent.

5.2 Convex Lipschitz Functions Revisited

In Sect. 4.3 we gave a sub-Gaussian bound for the upper tail of separately convex Lipschitz functions on $[0,1]^n$. Now we use self-boundedness to complement this with a sub-Gaussian lower bound, using an elegant trick of Boucheron et al. [6] where the lower bound in Theorem 34 is applied to the square of the Lipschitz function f. The essence of the trick is the following simple lemma.

Lemma 35 *If $f\geq 0$ then $D^2\left(f^2\right)\leq 4D^2\left(f\right)f^2$.*

Proof Since $f\geq 0$ we have $\inf_k\left(f^2\right)=\left(\inf_k f\right)^2$, so, using $(a+b)^2\leq 2a^2+2b^2$,

$$\begin{aligned}D^2\left(f^2\right)&=\sum_k\left(f^2-\inf_k f^2\right)^2=\sum_k\left(f-\inf_k f\right)^2\left(f+\inf_k f\right)^2\\&\leq 4f^2\sum_k\left(f-\inf_k f\right)^2=4D^2\left(f\right)f^2.\end{aligned}$$

□

For the sub-Gaussian lower bound we need the additional assumption that f^2 takes values in an interval of length at most one.

Theorem 36 *Let $\Omega_k = [0,1]$ and let $f \in \mathcal{A}$ be L-Lipschitz, nonnegative and such that $y \in [0,1] \mapsto S_y^k f(\mathbf{x})$ is convex for all k and all $\mathbf{x}$, and suppose in addition, that f^2 takes values in an interval of length at most one. Then for all $t \in [0, E[f]]$*

$$\Pr\{E[f] - f > t\} \leq e^{-t^2/8L^2}.$$

Proof The trick is to study the function f^2 instead of f. Let $\mathbf{x} \in [0,1]^n$. Using separate convexity as in the proof of Theorem 30 we have $D^2 f \leq L^2$, so by the previous lemma $D^2(f)^2 \leq 4L^2 f^2$. For any k we have $f^2(\mathbf{x}) - \inf f_k^2(\mathbf{x}) \leq 1$, so by the lower tail bound of Theorem 34 we get a lower tail bound for f^2

$$\Pr\left\{E\left[f^2\right] - f^2 > t\right\} \leq \exp\left(\frac{-t^2}{8L^2 E\left[f^2\right]}\right).$$

Thus

$$\begin{aligned}
\Pr\{E[f] - f > t\} &= \Pr\left\{\sqrt{E\left[f^2\right]}(E[f] - f) > \sqrt{E\left[f^2\right]}t\right\} \\
&\leq \Pr\left\{\left(\sqrt{E\left[f^2\right]} + f\right)\left(\sqrt{E\left[f^2\right]} - f\right) > \sqrt{E\left[f^2\right]}t\right\} \\
&= \Pr\left\{E\left[f^2\right] - f^2 > \sqrt{E\left[f^2\right]}t\right\} \\
&\leq \exp\left(\frac{-t^2}{8L^2}\right).
\end{aligned}$$

Here we used $E[f] \leq \sqrt{E\left[f^2\right]}$ and the assumption that f is nonnegative in the first inequality. □

5.3 Decoupling

A second method to avoid the uniform bound on the thermal expectation uses decoupling. By the duality formula of Theorem 4 we have for any $f, g \in \mathcal{A}$ and $\beta \in \mathbb{R}$

$$E_{\beta f}[g] \leq \text{Ent}_f(\beta) + \ln E\left[e^g\right]. \tag{24}$$

Recall the discussion at the beginning of Sect. 5, where we had a general bound of the form $\text{Ent}_f(\beta) \leq \xi(\beta) E_{\beta f}[G(f)]$. Using (24) we can now obtain for any $\lambda > 0$

$$\mathrm{Ent}_f(\beta) \leq \xi(\beta)\,\lambda^{-1} E_{\beta f}[\lambda G(f)] \leq \xi(\beta)\,\lambda^{-1}\left(\mathrm{Ent}_f(\beta) + \ln E\left[\exp(\lambda G(f))\right]\right),$$

and for values of β and λ where $\lambda > \xi(\beta)$ we obtain

$$\begin{aligned} \mathrm{Ent}_f(\beta) &\leq \frac{\xi(\beta)}{\lambda - \xi(\beta)} \ln E\left[\exp(\lambda G(f))\right] \\ &= \frac{\xi(\beta)}{\lambda - \xi(\beta)} \left(\ln E\left[e^{\lambda(G(f)-E[G(f)])}\right] + \lambda E[G(f)]\right). \end{aligned} \tag{25}$$

Hence, if we can control the moment generating function of $G(f)$ (or some suitable bound thereof), we obtain concentration inequalities for f, effectively passing from the thermal measure $\mu_{\beta f}$ to the thermal measure $\mu_{\lambda G(f)}$. The second line shows that in this way the supremum of $G(f)$ can possibly be replaced by an expectation. The $\lambda - \xi(\beta)$ in the denominator makes some constraint on β necessary, so the improvement comes at the price of an extra or enlarged subexponential term in the resulting concentration inequality. We conclude this chapter with three applications of this trick, which has been proposed in [7].

5.4 Quadratic Forms

As a first illustration we give a version of the Hanson-Wright inequality (Theorem 6.2.1 in [29]) for bounded variables. Let A be a symmetric $n \times n$-matrix, which is zero on the diagonal, that is $A_{ii} = 0$ for all i, and suppose that $X_1, ..., X_n$ are independent random variables with values in an interval $\mathcal{I}$ of unit diameter. We study the random variable $f(\mathbf{X})$, where

$$f(\mathbf{x}) = \sum_{i,j} x_i A_{ij} x_j.$$

As operator G we use R^2, the sum of squared conditional ranges which appears in the bounded difference inequality. For the function in question we have

$$D^k_{y,y'} f(\mathbf{x}) = 2\left(y - y'\right)\sum_i A_{ki} x_i = 2\left(y - y'\right)(A\mathbf{x})_k,$$

and, since $\mathcal{I}$ has unit diameter

$$R^2(f)(\mathbf{x}) = \sum_k \sup_{y,y' \in \mathcal{I}} \left(D^k_{y,y'} f(\mathbf{x})\right)^2 \leq 4\sum_k (A\mathbf{x})_k^2 = 4\,\|A\mathbf{x}\|^2.$$

We can therefore conclude from (12) in the proof of the bounded difference inequality (Theorem 14), that $\mathrm{Ent}_f(\gamma) \leq \left(\gamma^2/8\right) E_{\gamma f}\left[R^2(f)\right] \leq \left(\gamma^2/2\right) E_{\gamma f}\left[\|A\mathbf{X}\|^2\right]$. But instead of bounding the last thermal expectation by a supremum, as we did before, we now look for concentration properties of the function $\mathbf{x} \mapsto \|A\mathbf{x}\|^2$.

By (20) and Lemma 35 we have the self-bounding inequality $D^2\left(\|A\mathbf{x}\|^2\right) \leq 4\|A\|^2\|A\mathbf{x}\|^2$ and Lemma 32 gives for $0 \leq \lambda < 1/\left(2\|A\|^2\right)$

$$\ln E\left[e^{\lambda\|A\mathbf{x}\|^2}\right] \leq \frac{\lambda E\left[\|A\mathbf{x}\|^2\right]}{1-2\|A\|^2\lambda}.$$

Now Let $0 < \gamma < 1/\|A\|$ and set $\lambda := \gamma/(2\|A\|) < 1/\left(2\|A\|^2\right)$. Using the above bound on $\text{Ent}_f(\gamma)$ and the decoupling inequality (24) we get

$$\begin{aligned}\lambda\text{Ent}_f(\gamma) &\leq \frac{\gamma^2}{2}E_{\gamma f}\left[\lambda\|Ax\|^2\right] \leq \frac{\gamma^2}{2}\left(\text{Ent}_f(\gamma)+\ln E\left[e^{\lambda\|Ax\|^2}\right]\right)\\ &\leq \frac{\gamma^2}{2}\text{Ent}_f(\gamma)+\frac{\gamma^2}{2}\frac{\lambda E\left[\|Ax\|^2\right]}{1-2\|A\|^2\lambda}.\end{aligned}$$

Collect terms in $\text{Ent}_f(\gamma)$, divide by $\lambda-\gamma^2/2$ (which is positive by the constraint on γ and the choice of λ) and substitute the value of λ to get

$$\text{Ent}_f(\gamma) \leq \frac{\gamma^2}{(1-\|A\|\gamma)^2}\frac{E\left[\|Ax\|^2\right]}{2}.$$

From Theorem 12 we conclude that for $\beta < 1/\|A\|$

$$\begin{aligned}\Pr\{f-Ef\} &\leq \exp\left(\beta\int_0^\beta\frac{\text{Ent}_f(\gamma)}{\gamma^2}d\gamma-\beta t\right)\\ &\leq \exp\left(\frac{\beta^2}{1-\|A\|\beta}\frac{E\left[\|Ax\|^2\right]}{2}-\beta t\right),\end{aligned}$$

and using Lemma 33 to minimize the last expression in $\beta \in (0, 1/\|A\|)$ gives our version of the Hanson-Wright inequality for bounded variables.

Theorem 37 *Let A be a symmetric $n \times n$-matrix, zero on the diagonal, and $\mathbf{X} = (X_1, \ldots, X_n)$ a vector of independent random variables with values in an interval $\mathcal{I}$ of unit diameter. Let $f : \mathcal{X}^n \to \mathbb{R}$ be defined by $f(x) = \sum_{ij} x_i A_{ij} x_j$. Then for $t > 0$*

$$\Pr\{f-Ef > t\} \leq \exp\left(\frac{-t^2}{2E\left[\|A\mathbf{X}\|^2\right]+2\|A\|t}\right).$$

5.5 *The Supremum of an Empirical Process*

We will now apply the decoupling trick to the upwards tail of the supremum of an empirical process, sharpening the bound obtained in Sect. 3.4.

Theorem 38 *Let $X_1, \ldots, X_n$ be independent with values in some space $\mathcal{X}$ with X_i distributed as μ_i, and let $\mathcal{F}$ be an at most countable class of functions $f : \mathcal{X} \to [-1, 1]$ with $E[f(X_i)] = 0$. Define $F : \mathcal{X}^n \to \mathbb{R}$ and $W : \mathcal{X}^n \to \mathbb{R}$ by*

$$F(\mathbf{x}) = \sup_{f \in \mathcal{F}} \sum_i f(x_i) \ \textit{and}$$
$$W(\mathbf{x}) = \sup_{f \in \mathcal{F}} \sum_i \left(f^2(x_i) + E\left[f^2(X_i) \right] \right).$$

Then for $t > 0$

$$\Pr\{F - E[F] > t\} \le \exp\left(\frac{-t^2}{2E[W] + t} \right).$$

This inequality improves over Theorem 12.2 in [6], since by the triangle inequality $E[W] \le \Sigma^2 + \sigma^2$ and the constants in the denominator of the exponent are better by a factor of two, and optimal for the variance term.

Proof Let $0 < \gamma \le \beta < 2$. Using Theorem 26 and (24) we get

$$\mathrm{Ent}_F(\gamma) \le \frac{\gamma}{2} E_{\gamma F}\left[\gamma V_+^2(F) \right] \le \frac{\gamma}{2} \left(\mathrm{Ent}_F(\gamma) + \ln E e^{\gamma V_+^2(F)} \right).$$

Rearranging gives

$$\mathrm{Ent}_F(\gamma) \le \frac{\gamma}{2 - \gamma} \ln E e^{\gamma V_+^2(F)}. \tag{26}$$

Fix some $\mathbf{x} \in \mathcal{X}^n$ and let $\hat{f} \in \mathcal{F}$ witness the maximum in the definition of $F(\mathbf{x})$. For $y \in \mathcal{X}$ we have $\left(F - S_y^k F \right)_+ \le \left(\hat{f}(x_i) - \hat{f}(y) \right)_+$ and by the zero mean assumption

$$\begin{aligned} V_+^2(F)(\mathbf{x}) &= \sum_k E_{y \sim \mu_k}\left[\left(F(\mathbf{x}) - S_y^k F(\mathbf{x}) \right)_+^2 \right] \\ &\le \sum_k E_{y \sim \mu_k} \left(\hat{f}(x_k) - \hat{f}(y) \right)_+^2 \\ &\le \sum_k E_{y \sim \mu_k} \left(\hat{f}(x_k) - \hat{f}(y) \right)^2 \\ &= \sum_k \left(\hat{f}^2(x_k) + E\left[\hat{f}^2(X_k) \right] \right) \\ &\le W(\mathbf{x}). \end{aligned}$$

So $V_+^2(F) \le W$. It follows from (26) that

$$\mathrm{Ent}_F(\gamma) \le \frac{\gamma}{2 - \gamma} \ln E e^{\gamma V^+(F)} \le \frac{\gamma}{2 - \gamma} \ln E\left[e^{\gamma W} \right]. \tag{27}$$

Next we establish self-boundedness of W. Let $\hat{f} \in \mathcal{F}$ (different from the previous $\hat{f}$, which we don't need any more) witness the maximum in the definition of $W(\mathbf{x})$. Then

$$\begin{aligned} V_+^2(W)(\mathbf{x}) &= \sum_k E_{y\sim\mu_k}\left(W(\mathbf{x}) - S_y^k W(\mathbf{x})\right)_+^2 \\ &\leq \sum_k E_{y\sim\mu_k}\left[\left(\hat{f}^2(x_k) - \hat{f}^2(y)\right)_+^2\right] \\ &\leq \sum_k \hat{f}^2(x_k) \\ &\leq W. \end{aligned}$$

It therefore follows from the self-bounding lemma, Lemma 32, that

$$\ln E\left[e^{\gamma W}\right] \leq \frac{\gamma^2 E[W]}{2-\gamma} + \gamma E[W] = \frac{\gamma E[W]}{1-\gamma/2}.$$

Combining this with (27) gives

$$\mathrm{Ent}_F(\gamma) \leq \frac{\gamma}{2-\gamma}\left(\frac{\gamma E[W]}{1-\gamma/2}\right) = \frac{\gamma^2}{(1-\gamma/2)^2}\frac{E[W]}{2}.$$

From (6) in Theorem 12 we conclude that

$$\begin{aligned} \ln Ee^{\beta(F-EF)} &= \beta\int_0^\beta \frac{\mathrm{Ent}_F(\gamma)}{\gamma^2}d\gamma \leq \beta\int_0^\beta \frac{1}{(1-\gamma/2)^2}d\gamma\ \frac{E[W]}{2} \\ &= \frac{\beta^2}{1-\beta/2}\frac{E[W]}{2}. \end{aligned}$$

Using Lemma 33 it follows that

$$\begin{aligned} \Pr\{F - E[F] > t\} &\leq \inf_{\beta\in(0,2)} \exp\left(-\beta t + \frac{\beta^2}{1-\beta/2}\frac{E[W]}{2}\right) \\ &\leq \exp\left(\frac{-t^2}{2E[W]+t}\right). \end{aligned}$$

□

5.6 Another Version of Bernstein's Inequality

A potential weakness of Theorem 21 is the occurrence of the supremum in the definition of the variance parameter $V = \sup_{\mathbf{x}\in\Omega}\Sigma^2(f)(\mathbf{x})$. If the supremum could be replaced by an expectation, the variance parameter would become the Efron–Stein upper bound $E\left[\Sigma^2(f)\right]$ on the variance $\sigma^2(f)$, making the inequality considerably stronger. Such a modification is possible at the expense of enlarging the subexponential term in Bernstein's inequality. Define the interaction functional

$$J(f) = 2\left(\sup_{\mathbf{x},\mathbf{z}\in\Omega} \sum_{k,l:k\neq l} \sigma_k^2\left(f - S_{z_l}^l f\right)(\mathbf{x})\right)^{1/2}.$$

The following theorem is given in [21]

Theorem 39 *Suppose $f \in \mathcal{A}(\Omega)$ satisfies $f - E_k f \leq b$ for all k. Then for all $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{2E\left[\Sigma^2(f)\right] + (2b/3 + J(f))\,t}\right).$$

Here we will use the tools introduced above to prove a slight strengthening of this result, removing the boundedness conditions above.

Let $f : \Omega = \prod_{i=1}^n \Omega_i \to \mathbb{R}$ and consider the three conditions

$$(A) = ((f - E_k f) \leq b \text{ for all } k)$$
$$(B) = \left(E_k\left[(f - E_k f)^m\right] \leq \frac{1}{2} m!\sigma_k^2(f)\, b^{m-2} \text{ for } m \geq 2 \text{ and all } k\right)$$
$$(C) = \left(\sum_{k=1}^n E_k\left[(f - E_k f)^m\right] \leq \frac{\Sigma^2(f)}{2} m! b^{m-2} \text{ for } m \geq 2\right).$$

Then $(A) \implies (B) \implies (C)$. The last condition (sometimes called "Bernstein condition" in the literature) is sufficient for the following version of Bernstein's inequality, which extends Theorem 2.10 in [6] from sums to general functions and replaces the one-sided boundedness requirement of Theorem 39 by the Bernstein condition.

Theorem 40 *Let $f : \Omega = \prod_{i=1}^n \Omega_i \to \mathbb{R}$ be measurable and suppose that (C) holds. Then for $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{2E\left[\Sigma^2(f)\right] + (2b + J(f))\,t}\right).$$

The first step is to bound the entropy of f under the condition (C), thus replacing Lemma 19 in the proof of Theorem 21.

Lemma 41 *Suppose (C) holds with $b = 1$. Then for all $\beta \in [0, 1)$*

$$Ent_f(\beta) \leq \frac{\beta^2 E_{\beta f}\left[\Sigma^2(f)\right]}{2(1-\beta)^2}.$$

Proof First we get from the variational property of variance, that

$$\begin{aligned}\sigma_{k,\beta f}^{2}(f) &\leq E_{k,\beta f}\left[(f-E_{k}(f))^{2}\right]=\frac{E_{k}\left[(f-E_{k}(f))^{2}e^{\beta(f-E_{k}f)}\right]}{E_{k}\left[e^{\beta(f-E_{k}f)}\right]}\\ &\leq E_{k}\left[(f-E_{k}(f))^{2}e^{\beta(f-E_{k}f)}\right],\end{aligned}$$

where we used Jensen's inequality to get $E_k\left[\exp\left(\beta\left(f-E_k f\right)\right)\right]\geq 1$ for the second inequality. From monotone convergence and (C) we then get

$$\begin{aligned}\sum_{k=1}^{n}\sigma_{k,\beta f}^{2}(f) &\leq \sum_{k=1}^{n}E_{k}\left[(f-E_{k}f)^{2}e^{\beta(f-E_{k}f)}\right]=\sum_{m=0}^{\infty}\sum_{k=1}^{n}\frac{\beta^{m}}{m!}E_{k}\left[(f-E_{k}f)^{m+2}\right]\\ &\leq \frac{\Sigma^{2}(f)}{2}\sum_{m=0}^{\infty}(m+1)(m+2)\beta^{m}.\end{aligned}$$

Thus from Theorem 12

$$\begin{aligned}\mathrm{Ent}_{f}(\beta) &\leq E_{\beta f}\left[\int_{0}^{\beta}\int_{t}^{\beta}\sum_{k=1}^{n}\sigma_{k,sf}^{2}(f)\ ds\ dt\right]\\ &\leq \frac{E_{\beta f}\left[\Sigma^{2}(f)\right]}{2}\sum_{m=0}^{\infty}(m+1)(m+2)\int_{0}^{\beta}\int_{t}^{\beta}s^{m}dsdt\\ &= \frac{E_{\beta f}\left[\Sigma^{2}(f)\right]}{2}\beta^{2}\sum_{m=0}^{\infty}(m+1)\beta^{m}=\frac{\beta^{2}E_{\beta f}\left[\Sigma^{2}(f)\right]}{2(1-\beta)^{2}}.\end{aligned}$$

□

At this point we could bound the thermal expectation $E_{\beta f}\left[\Sigma^2(f)\right]$ by a supremum and proceed along the usual path to obtain a version of Theorem 21 under condition (C), which, for sums of independent variables, would reduce to Theorem 2.10 in [6]. Instead we wish to exploit the decoupling idea and look for concentration properties of $\Sigma^2(f)$.

The crucial property of the interaction functional J is, that J^2 is a self-bound for $\Sigma^2(f)$. The following Lemma is also the key to the proof of Theorem 39.

Lemma 42 *We have* $D^2\left(\Sigma^2(f)\right)\leq J(f)^2\ \Sigma^2(f)$ *for any* $f\in\mathcal{A}(\Omega)$.

Proof Fix $\mathbf{x}\in\Omega$. Below all members of $\mathcal{A}$ are understood as evaluated on $\mathbf{x}$. For $l\in\{1,...,n\}$ let $z_l\in\Omega_l$ be a minimizer in z of $S_z^l\Sigma^2(f)$. Then

$$D^{2}\left(\Sigma^{2}(f)\right)=\sum_{l}\left(\sum_{k:k\neq l}\left(\sigma_{k}^{2}(f)-S_{z_{l}}^{l}\sigma_{k}^{2}(f)\right)\right)^{2}.$$

The sum over $k \neq l$, since $\sigma_k^2(f) \in \mathcal{A}_k$, so $S_{z_l}^l \sigma_k^2(f) = \sigma_k^2(f)$. Then, using $2\sigma_k^2(f) = E_{(y,y')\sim\mu_k^2}\left(D_{y,y'}^k f\right)^2$, we get

$$\begin{aligned}
4D^2\left(\Sigma^2(f)\right) &= \sum_l \left(\sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left(D_{y,y'}^k f\right)^2 - S_{z_l}^l E_{(y,y')\sim\mu_k^2}\left(D_{y,y'}^k f\right)^2\right)^2 \\
&= \sum_l \left(\sum_{k\neq l} E_{(y,y')\sim\mu_k^2}\left[\left(D_{y,y'}^k f\right)^2 - \left(D_{y,y'}^k S_{z_l}^l f\right)^2\right]\right)^2 \\
&= \sum_l \left(\sum_{k\neq l} E_{(y,y')\sim\mu_k^2}\left[\left(D_{y,y'}^k f - D_{y,y'}^k S_{z_l}^l f\right)\left(D_{y,y'}^k f + D_{y,y'}^k S_{z_l}^l f\right)\right]\right)^2 \\
&\leq \sum_l \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k\left(f - S_{z_l}^l f\right)\right]^2 \times \\
&\qquad \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k f + D_{y,y'}^k S_{z_l}^l f\right]^2
\end{aligned}$$

by an application of Cauchy–Schwarz. Now, using $(a+b)^2 \leq 2a^2 + 2b^2$, we can bound the last sum independent of l by

$$\begin{aligned}
&\sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[2\left(D_{y,y'}^k f\right)^2 + 2\left(D_{y,y'}^k S_{z_l}^l f\right)^2\right] \\
&= 4\sum_{k:k\neq l}\sigma_k^2(f) + 4S_{z_l}^l \sum_{k:k\neq l}\sigma_k^2(f) \\
&\leq 4\left(\Sigma^2(f) + S_{z_l}^l \Sigma^2(f)\right) = 4\left(\Sigma^2(f) + \inf_{z\in\Omega_l} S_z^l \Sigma^2(f)\right) \leq 8\Sigma^2(f),
\end{aligned}$$

so that

$$\begin{aligned}
D^2\left(\Sigma^2(f)\right) &\leq 2\sum_l \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k\left(f - S_{z_l}^l f\right)\right]^2 \Sigma^2(f) \\
&\leq 4\sup_{\mathbf{x},\mathbf{z}\in\Omega} \sum_{k,l:k\neq l} \sigma_k^2\left(f - S_z^l f\right)(\mathbf{x})\,\Sigma^2(f) = J^2(f)\,\Sigma^2(f).
\end{aligned}$$

□

Now we can use decoupling to put these pieces together.

Proof of Theorem 40 By rescaling it suffices to prove the result for $b = 1$. We can also assume $J := J(f) > 0$. Let $0 < \gamma \leq \beta < 1/(1 + J/2)$ and set $\theta = \gamma/(J(1-\gamma))$. Then $\gamma^2/\left(2(1-\gamma)^2\right) < \theta < 2/J^2$. By the Lemma 41

$$\theta \mathrm{Ent}_f(\gamma) \le \frac{\gamma^2}{2(1-\gamma)^2} E_{\gamma f}\left[\theta \Sigma^2(f)\right] \le \frac{\gamma^2}{2(1-\gamma)^2}\left(\mathrm{Ent}_f(\gamma) + \ln E\left[e^{\theta \Sigma^2(f)}\right]\right),$$

where the second inequality follows from the decoupling inequality (24). Subtract $\gamma^2/\left(2(1-\gamma)^2\right)\mathrm{Ent}_f(\gamma)$ to get

$$\mathrm{Ent}_f(\gamma)\left(\theta - \frac{\gamma^2}{2(1-\gamma)^2}\right) \le \frac{\gamma^2}{2(1-\gamma)^2} \ln E\left[e^{\theta \Sigma^2(f)}\right].$$

Since $\gamma^2/\left(2(1-\gamma)^2\right) < \theta$ this simplifies, using the value of θ, to

$$\mathrm{Ent}_f(\gamma) \le \frac{\gamma J}{2(1-(1+J/2)\gamma)} \ln E\left[e^{\theta \Sigma^2(f)}\right]. \tag{28}$$

On the other hand $\theta < 2/J^2$, so by the self-boundedness of $\Sigma^2(f)$ (Lemma 42) and part (i) of Lemma 32 give

$$\ln E\left[e^{\theta \Sigma^2(f)}\right] \le \frac{\theta}{1 - J^2\theta/2} E\left[\Sigma^2(f)\right] = \frac{\gamma/J}{1-(1+J/2)\gamma} E\left[\Sigma^2(f)\right]. \tag{29}$$

Combining (28) and (29) to get a bound on $S_f(\gamma)$ gives

$$\mathrm{Ent}_f(\gamma) \le \frac{\gamma^2}{2(1-(1+J/2)\gamma)^2} E\left[\Sigma^2(f)\right]$$

and from Theorem 12 and Lemma 33

$$\begin{aligned}\Pr\{f - Ef > t\} &\le \inf_{\beta \in (0, 1/(1+J/2))} \exp\left(\frac{E\left[\Sigma^2(f)\right]}{2} \frac{\beta^2}{1-(1+J/2)\beta} - \beta t\right) \\ &\le \exp\left(\frac{-t^2}{2\left(E\left[\Sigma^2(f)\right] + (1+J/2)t\right)}\right).\end{aligned}$$

□

To use Theorem 40 one has to bound b and J. For the latter it is often sufficient to use the simple bound

$$J(f) \le n \max_{k \neq l} \sup_{\mathbf{x} \in \Omega} \sup_{z, z', y, y' \in \Omega_l} D^l_{z,z'} D^k_{y,y'} f(\mathbf{x}). \tag{30}$$

which can be obtained from Lemma 13.

We conclude with an application to U-statistics. Let $m < n$ be integers, $\Omega_i = \mathcal{X}$ and $\kappa : \mathcal{X}^m \to \mathbb{R}$ a symmetric kernel. For a subset of indices with cardinality m, $S = \{j_1, ..., j_m\} \subseteq \{1, ..., n\}$ define $\kappa_S : \mathcal{X}^n \to \mathbb{R}$ by $\kappa_S(\mathbf{x}) = \kappa\left(x_{j_1}, ..., x_{j_m}\right)$. The U-statistic of order m induced by κ is then the function $U : \mathcal{X}^n \to \mathbb{R}$ given by

$$U(\mathbf{x}) = \binom{n}{m}^{-1} \sum_{S \subseteq \{1,\ldots,n\}} \kappa_S(\mathbf{x}).$$

U-statistics were introduced by Hoeffding [15]. Their importance stems from the fact that for iid $\mathbf{X} = (X_1, ..., X_n)$ the random variable $U(\mathbf{X})$ is an unbiased estimator for $E[\kappa(X_1, ..., X_m)]$. Starting with the work of Hoeffding there has been a lot of work on concentration inequalities for U-statistics. To simplify the presentation we will not use the advantage of Theorem 40 over Theorem 39 and assume the kernel κ to be bounded, $\kappa : \mathcal{X}^m \to [0, 1]$ for simplicity.

Notice that,if $k \notin S$, then $\kappa_S \in \mathcal{A}_k$, so $\kappa_S(\mathbf{x}) - E_k[\kappa_S(\mathbf{x})] = 0$ and thus

$$\begin{aligned} U(\mathbf{x}) - E_k[U(\mathbf{x})] &= \binom{n}{m}^{-1} \sum_{\substack{S \subseteq \{1,\ldots,n\} \\ k \in S}} (\kappa_S(\mathbf{x}) - E_k[\kappa_S(\mathbf{x})]) \\ &\leq \binom{n}{m}^{-1} |\{S \subseteq \{1, ..., n\} : k \in S\}| \\ &= \frac{\binom{n-1}{m-1}^{-1}}{\binom{n}{m}} = \frac{m!\,(n-1)!}{n!\,(m-1)!} = \frac{m}{n}, \end{aligned}$$

so we can set the quantity b in Theorem 40 to m/n. To bound J use (30) to get

$$\begin{aligned} J(U) &\leq n \max_{k \neq l} \sup_{\mathbf{x} \in \Omega} \sup_{z,z',y,y' \in \Omega_l} D^l_{z,z'} D^k_{y,y'} U(\mathbf{x}) \\ &\leq n \binom{n}{m}^{-1} \sum_{\substack{S \subseteq \{1,\ldots,n\} \\ k,l \in S : k \neq l}} D^l_{z,z'} D^k_{y,y'} \kappa_S(\mathbf{x}) \\ &= 2n \binom{n}{m}^{-1} |\{S \subseteq \{1, ..., n\} : k, l \in S, k \neq l\}| \\ &= \frac{2n\binom{n-2}{m-2}}{\binom{n}{m}} \leq \frac{2m^2}{n}. \end{aligned}$$

Substitution in Theorem 40 gives for $t > 0$

$$\Pr\{U - EU > t\} \leq \exp\left(\frac{-t^2}{2E\left[\Sigma^2(U)\right] + 2\left(m + m^2\right)t/n}\right).$$

It can be shown (see, e.g., [21], Houdré [17]) that in general $E\left[\Sigma^2(f)\right] \leq \sigma^2(f) + J^2(f)/4$, so that for U-statistics the Efron–Stein inequality is tight in the sense that $E\left[\Sigma^2(U)\right] \leq \sigma^2(U) + m^4/n^2$. It follows that for deviations $t > 1/n$

$$\Pr\{U - EU > t\} \leq \exp\left(\frac{-t^2}{2\sigma^2(U) + 2\left(m + 2m^2\right)t/n}\right).$$

This inequality can be compared to the classical work of Hoeffding [15] and more recent results of Arcones [2], which both consider undecoupled, nondegenerate U-statistics of arbitrary order. Hoeffding [15] does not have the correct variance term, while [2] gives the correct variance term but severely overestimates the subexponential coefficient in Bernstein's inequality to be exponential in the degree m of the U-statistic (above it is only of order m^2). This exponential dependence on m results from the use of the decoupling inequalities in [24] and seems to beset most works on U-statistics of higher order (e.g., [1, 13]), which in many other ways improve over our simple inequality above.

6 Appendix I. Table of Notation

General notation

$\Omega = \prod_{k=1}^{n} \Omega_k$	underlying (product-) probability space
$\mathcal{A}$	bounded measurable functions on Ω
$\mu = \otimes_{k=1}^{n} \mu_k$	(product-) probability measure on Ω
X_k	random variable distributed as μ_k in Ω_k
$f \in \mathcal{A}$	fixed function under investigation
$g \in \mathcal{A}$	generic function
$E[g] = \int_\Omega g d\mu$	expectation of g in μ
$\sigma^2[g] = E\left[(g - E[g])^2\right]$	variance of g in μ

Notation for the entropy method

$\beta = 1/T$	inverse temperature
$E_{\beta f}[g] = E\left[ge^{\beta f}\right]/E\left[e^{\beta f}\right]$	thermal expectation of g
$Z_{\beta f} = E\left[e^{\beta f}\right]$	partition function
$d\mu_{\beta f} = Z_{\beta f}^{\cdot 1} e^{\beta f} d\mu$	thermal measure (canonical ensemble)
$\mathrm{Ent}_f(\beta) = \beta E_{\beta f}[f] - \ln Z_{\beta f}.$	(canonical) entropy
$A_f(\beta) = \frac{1}{\beta} \ln Z_{\beta f}$	free energy
$\sigma_{\beta f}^2(g) = E_{\beta f}\left[\left(g - E_{\beta f}[g]\right)^2\right]$	thermal variance of g
$\psi(t) = e^t - t - 1$	
$S_y^k F(\mathbf{x}) = F\left(x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n\right)$	substitution operator
$E_k[g](\mathbf{x}) = \int_{\Omega_k} S_y^k g \, d\mu_k(y)$	conditional expectation
$\mathcal{A}_k \subset \mathcal{A}$	functions independent of k-th variable
$Z_{k,\beta f} = E_k\left[e^{\beta f}\right]$	conditional partition function
$E_{k,\beta f}[g] = Z_{k,\beta f}^{-1} E_k\left[ge^{\beta f}\right]$	conditional thermal expectation
$\mathrm{Ent}_{k,f}(\beta) = \beta E_{k,\beta f}[g] - \ln Z_{k,\beta f}$	conditional entropy
$\sigma_{k,\beta f}^2[g] = E_{k,\beta f}\left[\left(g - E_{k,\beta f}[g]\right)^2\right]$	conditional thermal variance
$\sigma_k^2[g] = E_k\left[(g - E_k[g])^2\right]$	conditional variance

Operators on $\mathcal{A}$

$D^k_{y,y'}g = S^k_y g - S^k_{y'} g$	difference operator
$r_k\,(g) = \sup_{y,y' \in \Omega_k} D^k_{y,y'} f$	conditional range operator
$R^2\,(g) = \sum_k r^2_k\,(g)$	sum of conditional square ranges
$\Sigma^2\,(g) = \sum_k \sigma^2_k\,[g]$	sum of conditional variances
$(\inf_k g)\,(\mathbf{x}) = \inf_{y \in \Omega_k} S^k_y g\,(\mathbf{x})$	conditional infimum operator
$V^2_+ g = \sum_k E_{y \sim \mu_k}\left[\left(\left(g - S^k_y\right)_+\right)^2\right]$	Efron–Stein variance proxy
$D^2 g = \sum_k (g - \inf_k g)^2\,.$	worst case variance proxy

References

1. Adamczak, R., et al.: Moment inequalities for u-statistics. Ann. Probab. **34**(6), 2288–2314 (2006)
2. Arcones, M.A.: A Bernstein-type inequality for u-statistics and u-processes. Stat. Probab. Lett. **22**(3), 239–247 (1995)
3. Bartlett, P., Mendelson, S.: Rademacher and gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2002)
4. Bernstein, S.: Theory of Probability. Moscow (1927)
5. Boltzmann, L.: Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht. Kk Hof-und Staatsdruckerei (1877)
6. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities. Oxford University Press, Oxford (2013)
7. Boucheron, S., Lugosi, G., Massart, P., et al.: Concentration inequalities using the entropy method. Ann. Probab. **31**(3), 1583–1614 (2003)
8. Boucheron, S., Lugosi, G., Massart, P., et al.: On concentration of self-bounding functions. Electron. J. Probab. **14**, 1884–1899 (2009)
9. Bousquet, O.: A Bennett concentration inequality and its application to suprema of empirical processes. C.R. Math. **334**(6), 495–500 (2002)
10. Chatterjee, S.: Concentration inequalities with exchangeable pairs. Ph.D. thesis, Citeseer (2005)
11. Chebyshev, P.L.: Sur les valeurs limites des intégrales. Imprimerie de Gauthier-Villars (1874)
12. Gibbs, J.W.: Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics. C. Scribner's Sons, New York (1902)
13. Giné, E., Latała, R., Zinn, J.: Exponential and moment inequalities for u-statistics. High Dimensional Probability II, pp. 13–38. Springer, Berlin (2000)
14. Gross, L.: Logarithmic sobolev inequalities. Am. J. Math. **97**(4), 1061–1083 (1975)
15. Hoeffding, W.: A class of statistics with asymptotically normal distribution. Ann. Math. Stat., pp. 293–325 (1948)
16. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 301 (1963)
17. Houdré, C.: The iterated jackknife estimate of variance. Statist. Probab. Lett. **35**(2), 197–201 (1997)
18. Ledoux, M.: The Concentration of Measure Phenomenon, vol. 89. American Mathematical Society, Providence (2001)
19. Lieb, E.H.: Some convexity and subadditivity properties of entropy. Inequalities, pp. 67–79. Springer, Berlin (2002)

20. Massart, P., et al.: About the constants in Talagrand's concentration inequalities for empirical processes. Ann. Probab. **28**(2), 863–884 (2000)
21. Maurer, A., et al.: A Bernstein-type inequality for functions of bounded interaction. Bernoulli **25**(2), 1451–1471 (2019)
22. McAllester, D., Ortiz, L.: Concentration inequalities for the missing mass and for histogram rule error. J. Mach. Learn. Res. **4**(Oct), 895–911 (2003)
23. McDiarmid, C.: Concentration. Probabilistic Methods of Algorithmic Discrete Mathematics, pp. 195–248. Springer, Berlin (1998)
24. de la Peña, V.H.: Decoupling and khintchine's inequalities for u-statistics. Ann. Probab., pp. 1877–1892 (1992)
25. Popper, K.R.: Logik der forschung (1934). The Logic of Scientific Discovery. [Google Scholar] (1968)
26. Steele, J.M.: An Efron-Stein inequality for nonsymmetric statistics. Ann. Probab., pp. 753–758 (1986)
27. Talagrand, M.: Concentration of measure and isoperimetric inequalities in product spaces. Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques **81**(1), 73–205 (1995)
28. Talagrand, M.: A new look at independence. Ann. Probab., pp. 1–34 (1996)
29. Vershynin, R.: High-dimensional Probability: An Introduction with Applications in Data Science, vol. 47. Cambridge University Press, Cambridge (2018)