Bernoulli **18**(2), 2012, 434–454
DOI: 10.3150/10-BEJ341

Thermodynamics and concentration

ANDREAS MAURER

Adalbertstr. 55, D-80799 München, Germany. E-mail: am@andreas-maurer.eu

We show that the thermal subadditivity of entropy provides a common basis to derive a strong form of the bounded difference inequality and related results as well as more recent inequalities applicable to convex Lipschitz functions, random symmetric matrices, shortest travelling salesmen paths and weakly self-bounding functions. We also give two new concentration inequalities.

Keywords: concentration; entropy method; tail bounds

1. Introduction

Concentration inequalities bound the probabilities that random quantities deviate from their average, median or otherwise typical values. They are at the heart of empirical science and play an important role in the study of natural and artificial learning systems.

An early concentration inequality for sums was given by Chebychev and Bienaymé in the 19th century [6] and allowed a rigorous proof of the weak law of large numbers. The subject has since been developed by Bernstein, Chernoff, Bennett, Hoeffding and many others [1,9], and results were extended from sums to more general and complicated nonlinear functions. During the past few decades, research activity has been stimulated by the contributions of Michel Talagrand [22,23] 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 [19]. Nevertheless, a unified and elementary basis for the derivation of the many available results is still missing.

One of the more recent methods used to derive concentration inequalities, the so-called *entropy method*, is rooted in the early investigations of Boltzmann [2] and Gibbs [7] into the foundations of statistical mechanics. A general problem of statistical mechanics is to demonstrate the "equivalence of ensembles", which can be interpreted as an exponential concentration property of the Hamiltonian, or energy function. While the modern entropy method evolved along a complicated historical path via quantum field theory and the logarithmic Sobolev inequality of Leonard Gross [8], its hidden simplicity was understood and emphasized by Michel Ledoux, who also recognized the key role that the subadditivity of entropy can play in the derivation of concentration inequalities [10,11]. Recently, Boucheron *et al.* [4] showed that the entropy method is sufficiently strong to derive a form of Talagrand's convex distance inequality.

The purpose of this paper is to advertise the subadditivity of entropy as a unified basis for the derivation of concentration inequalities for functions on product spaces and to demonstrate the benefits of formulating the concentration problem in the language of statistical thermodynamics, an approach proposed by David McAllester [18].

Our method consists of three steps. The first step (Theorem 1) expresses the log-Laplace transform (or, more directly, the deviation probability) in terms of an integral of the thermal entropy over a range of inverse temperatures. This step encapsulates the so-called Herbst argument.

The second step (Theorem 6) is the tensorization inequality, or, more properly, a thermal subadditivity property of entropy. It asserts that the entropy of a system is no greater than the thermal average of the sum of entropies of the constituent subsystems.

The third step (Theorem 3) expresses the entropy of the subsystem in terms of thermal energy fluctuations.

All three steps are elementary and their combination leads to a general concentration result (Theorem 7) that can be used whenever we succeed in controlling the latter fluctuations.

We then use the method to first derive a strong form of the bounded difference inequality and an inequality given by McDiarmid and related to Bennett's inequality [19]. These results are normally not associated with the entropy method. Then monotonicity properties of thermal energy fluctuations, or bounds thereof, are exploited to derive two apparantly novel sub-Gaussian tail-bounds and to give a new proof of an upper tail-bound in [16] that improves on some results obtained from Talagrand's convex distance inequality. Finally, we show how our method can be extended in a generic way using self-boundedness and/or decoupling, and illustrate this extension by deriving a concentration inequality that underlies the recent new proof of the convex distance inequality [4].

Clearly statement and proof of all the results presented in this paper would be possible on a purely formal basis without any reference to physics. The author believes, however, that positioning the subject in a broader scientific context highlights its historical connections and gives access to a valuable source of intuition.

In the next section, we describe the connection between entropy and concentration and introduce several thermodynamic functions. We then transfer these concepts to product spaces and present the tensorization inequality. The remaining sections are dedicated to applications, and we conclude with a tabular summary of the notation used in the paper.

2. Entropy and concentration

Let (Ω, Σ, μ) be a probability space and $f \in L_\infty[\mu]$ be a fixed function whose concentration properties are to be studied.

We interpret the points $x \in \Omega$ as possible states of a physical system and f as the negative energy (or Hamiltonian) function, so that $-f(\mathbf{x})$ is the system's energy in the state x. The measure μ models an a priori probability distribution of states in the absence of any constraining information.

We will ignore questions of measurability. If it seems necessary to the reader, Ω may be taken as a potentially very large finite set, the cardinality of which will play no role in our results. The boundedness assumption is a simplification that is justified by the fact that most of our results are vacuous for $f \notin L_\infty[\mu]$. In the remaining cases, we will mention optimal conditions on f.

For any $g \in L_\infty[\mu]$ we write $E[g] = \int_\Omega g \, \mathrm{d}\mu$ and $\sigma^2[g] = E[(g - E[g])^2]$.

2.1. Thermal equilibrium and thermodynamic functions

Our function f defines a one-parameter family $\{E_{\beta f}\colon\ \beta \in \mathbb{R}\}$ of expectation functionals by

$$E_{\beta f}[g] = \frac{E[g\mathrm{e}^{\beta f}]}{E[\mathrm{e}^{\beta f}]}, \qquad g \in L_\infty[\mu].$$

In statistical thermodynamics, $E_{\beta f}[g]$ is the *thermal expectation* of the observable g at temperature $T = 1/\beta$. The normalizing expectation is called the *partition function*,

$$Z_{\beta f} = E[\mathrm{e}^{\beta f}].$$

The corresponding probability measure on Ω,

$$\mathrm{d}\mu_{\beta f} = Z_{\beta f}^{-1}\mathrm{e}^{\beta f}\,\mathrm{d}\mu,$$

is called the canonical ensemble. It describes a system in thermal equilibrium with a heat reservoir at temperature $T = 1/\beta$. The canonical ensemble has the density $\rho = Z_{\beta f}^{-1}\mathrm{e}^{\beta f}$, which maximizes the Kullback–Leibler divergence or relative entropy $KL(\rho\,\mathrm{d}\mu, \mathrm{d}\mu) := E[\rho \ln \rho]$, given the expected internal energy $-E[\rho f]$. The parameter β is the Lagrange multiplier corresponding to this constraint. For a constant c we have the obvious and important identity $E_{\beta(f+c)}[g] = E_{\beta f}[g]$.

The corresponding maximal value of the Kullback–Leibler divergence is the *canonical entropy*

$$S_f(\beta) = KL(Z_{\beta f}^{-1}\mathrm{e}^{\beta f}\,\mathrm{d}\mu, \mathrm{d}\mu) = \beta E_{\beta f}[f] - \ln Z_{\beta f}. \tag{2.1}$$

Note that $S_{-f}(\beta) = S_f(-\beta)$, a simple but very useful fact to pass from upper to lower tails.

For $\beta \neq 0$ the Helmholtz free energy is defined by

$$A_f(\beta) = \frac{1}{\beta}\ln Z_{\beta f}.$$

Dividing (2.1) by β and writing $U = E_{\beta f}[f]$, we obtain the classical thermodynamic relation

$$A = U - TS,$$

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

By L'Hôpital's rule, we have $\lim_{\beta\to 0} A_f(\beta) = E[f]$, so the free energy A_f extends continuously to $\mathbb{R}$ by setting $A_f(0) = E[f]$. We find

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

Integrating this identity from zero to β and multiplying with β, we obtain:

Theorem 1. *For any $\beta > 0$ we have*

$$\ln E\big[\mathrm{e}^{\beta(f-Ef)}\big] = \beta \int_0^\beta \frac{S_f(\gamma)}{\gamma^2}\,\mathrm{d}\gamma$$

and, for $t \geq 0$,

$$\Pr\{f - Ef > t\} \leq \exp\left(\beta \int_0^\beta \frac{S_f(\gamma)}{\gamma^2}\,\mathrm{d}\gamma - \beta t\right).$$

Proof.

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

Combining this with Markov's inequality gives the second assertion. □

The theorem shows how bounds on the canonical entropy can lead to concentration results. In the following we present ways to arrive at such bounds.

2.2. Entropy and energy fluctuations

The *thermal variance* of a function $g \in L_\infty[\mu]$ is denoted $\sigma^2_{\beta f}(g)$ and defined by

$$\sigma^2_{\beta f}(g) = E_{\beta f}\big[(g - E_{\beta f}[g])^2\big] = E_{\beta f}[g^2] - (E_{\beta f}[g])^2.$$

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

We first give some simple results pertaining to the derivatives of the partition function and the thermal expectations.

Lemma 2. *The following formulas hold*:

1. $\frac{\mathrm{d}}{\mathrm{d}\beta}(\ln Z_{\beta f}) = E_{\beta f}[f]$.
2. *If $h : \beta \mapsto h(\beta) \in L_\infty[\mu]$ is differentiable and $(\mathrm{d}/\mathrm{d}\beta)h(\beta) \in L_\infty[\mu]$, then*

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

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

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

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 [18].

Theorem 3. *We have for* $\beta > 0$

$$S_f(\beta) = \int_0^\beta \int_t^\beta \sigma_{sf}^2[f]\,\mathrm{d}s\,\mathrm{d}t.$$

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

$$\beta E_{\beta f}[f] = \int_0^\beta E_{\beta f}[f]\,\mathrm{d}t = \int_0^\beta \left(\int_0^\beta \sigma_{sf}^2[f]\,\mathrm{d}s + E[f] \right) \mathrm{d}t$$

and

$$\ln Z_{\beta f} = \int_0^\beta E_{tf}[f]\,\mathrm{d}t = \int_0^\beta \left(\int_0^t \sigma_{sf}^2[f]\,\mathrm{d}s + E[f] \right) \mathrm{d}t,$$

which we subtract to obtain

$$\begin{aligned} S_f(\beta) = \beta E_{\beta f}[f] - \ln Z_{\beta f} &= \int_0^\beta \left(\int_0^\beta \sigma_{sf}^2[f]\,\mathrm{d}s - \int_0^t \sigma_{sf}^2[f]\,\mathrm{d}s \right) \mathrm{d}t \\ &= \int_0^\beta \left(\int_t^\beta \sigma_{sf}^2[f]\,\mathrm{d}s \right) \mathrm{d}t. \end{aligned}$$

□

Since bounding $\sigma_{\beta f}^2[f]$ 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 2 we have

$$\frac{\mathrm{d}}{\mathrm{d}T}(-E_{\beta f}[f]) = \frac{1}{T^2}\sigma_{\beta f}^2[f],$$

which exhibits the proportionality of heat capacity and energy fluctuations.

2.3. A variational entropy bound

While Theorem 3 is just an elementary way of rewriting the canonical entropy, the following lemma is typically a strict inequality that leads to the modified logarithmic Sobolev inequality proposed by Massart in [13]. To state it, we define the real function

$$\psi(t) = \mathrm{e}^t - t - 1, \tag{2.2}$$

which arises from deleting the first two terms in the power series expansion of the exponential function.

Lemma 4. *If $c \in \mathbb{R}$, then*

$$S_f(\beta) \le E_{\beta f}\big[\psi\big(-\beta(f-c)\big)\big].$$

Proof. Using $\ln t \le t - 1$, we get

$$\beta f - \ln Z_{\beta f} = \beta(f-c) + \ln \frac{\mathrm{e}^{\beta c}}{Z_{\beta f}} \le \beta(f-c) + \left(\frac{\mathrm{e}^{\beta c}}{Z_{\beta f}} - 1\right).$$

Taking the thermal expectation then gives

$$\begin{aligned} S_f(\beta) &\le E_{\beta f}[\beta(f-c)] + \frac{\mathrm{e}^{\beta c}}{Z_{\beta f}} - 1 \\ &= E_{\beta f}[\beta(f-c)] + E\left[\frac{\mathrm{e}^{-\beta(f-c)}\mathrm{e}^{\beta f}}{Z_{\beta f}}\right] - 1 \\ &= E_{\beta f}\big[\mathrm{e}^{-\beta(f-c)} + \beta(f-c) - 1\big]. \end{aligned}$$

□

3. Product spaces

We now assume that $\Omega = \prod_{k=1}^n \Omega_k$ and $\mu = \bigotimes_{k=1}^n \mu_k$, where each μ_k is the probability measure representing the distribution of some variable X_k in the space Ω_k, where all the X_k are assumed to be mutually independent. The X_k are irrelevant for the derivation of our inequalities, but they are convenient in the discussion of applications.

If $\mathbf{x} = (x_1, \dots, x_n) \in \Omega$ describes a state of a physical system, we can think of $x_k \in \Omega_k$ as the state of the kth subsystem, which may be a particle or a more abstract object, such as a spin assigned to the vertex of a graph. The a priori measure μ assigns independent probabilities μ_k to the states of the subsystems. If the total energy is a sum of energies of the subsystems, $f = \sum f_k$, with $f_k \in L_\infty[\mu_k]$, then this is also true for the canonical ensemble $Z_{\beta f}\mathrm{e}^{\beta f}\,\mathrm{d}\mu$ corresponding to non-interaction of the subsystems.

3.1. Conditional expectations

For $\mathbf{x} \in \Omega$, $1 \le k \le n$ and $y \in \Omega_k$ we use $\mathbf{x}_{y,k}$ to denote the vector in Ω, which is obtained by replacing x_k with y. We also write, for $g \in L_\infty[\mu]$,

$$E_k[g](\mathbf{x}) = \int_{\Omega_k} g(\mathbf{x}_{y,k})\,\mathrm{d}\mu_k(y) = \int_{\Omega_k} g(x_1, \dots, x_{k-1}, y, x_{k+1}, \dots, x_n)\,\mathrm{d}\mu_k(y).$$

The operator E_k corresponds to an expectation conditional to all variables with indices different to k. We denote with $\mathcal{A}_k$ the sub-algebra of $L_\infty[\mu]$ consisting of those functions that are independent of the kth variable. E_k is evidently a linear projection onto $\mathcal{A}_k$. Also, the E_k commute amongst each other and, for $h \in L_\infty[\mu]$ and $g \in \mathcal{A}_k$, we have

$$E[[E_k h]g] = E[E_k[hg]] = E[hg]. \tag{3.1}$$

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 $Z_{k,\beta f} = E_k[\mathrm{e}^{\beta f}]$,
- the conditional thermal expectation $E_{k,\beta f}[g] = Z_{k,\beta f}^{-1} E_k[g\mathrm{e}^{\beta f}]$ for $g \in L_\infty[\mu]$,
- the conditional entropy $S_{k,f}(\beta) = \beta E_{k,\beta f}[f] - \ln Z_{k,\beta f}$,
- the conditional free energy $A_{k,f}(\beta) = \beta^{-1} \ln Z_{k,\beta f}$,
- the conditional thermal variance $\sigma^2_{k,\beta f}[g] = E_{k,\beta f}[(g - E_{k,\beta f}[g])^2]$ for $g \in L_\infty[\mu]$. As $\beta \to 0$, this becomes
- the conditional variance $\sigma^2_k[g] = E_k[(g - E_k[g])^2]$ for $g \in L_\infty[\mu]$.

If we fix all variables except x_k, then E_k just becomes an ordinary expectation, and it becomes evident that all the previously established relations also hold for the corresponding conditional quantities; in particular, the conclusions of Theorem 3,

$$S_{k,f}(\beta) = \int_0^\beta \int_t^\beta \sigma^2_{k,sf}[f]\,\mathrm{d}s\,\mathrm{d}t,$$

and of Lemma 4,

$$S_{k,f}(\beta) \le E_{k,\beta f}\big[\psi\big(-\beta(f - f_k)\big)\big] \qquad \text{if } f_k \in \mathcal{A}_k.$$

Other members of $\mathcal{A}_k$ that will play a role in the sequel are:

- the conditional supremum $(\sup_k g)(\mathbf{x}) = \sup_{y \in \Omega_k} g(\mathbf{x}_{y,k})$ for $g \in L_\infty[\mu]$,
- the conditional infimum $(\inf_k g)(\mathbf{x}) = \inf_{y \in \Omega_k} g(\mathbf{x}_{y,k})$ for $g \in L_\infty[\mu]$ and
- the conditional range $\mathrm{ran}_k(g) = \sup_k g - \inf_k g$ for $g \in L_\infty[\mu]$.

3.2. Tensorization of entropy

In the non-interacting case, when the energy function f is a sum, $f = \sum f_k$, with $f_k \in L_\infty[\mu_k]$, it is easily verified that $S_{k,f}(\beta)(\mathbf{x}) = S_{k,f}(\beta)$ is independent of $\mathbf{x}$ and that

$$S_f(\beta) = \sum_{k=1}^n S_{k,f}(\beta). \tag{3.2}$$

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

The tensorization inequality states that the total entropy is no greater than the thermal average of the sum of the conditional entropies. In 1975, Elliott Lieb [12] gave a proof of this result, which was probably known some time before, at least in the classical setting relevant to our arguments.

Lemma 5. *Let $h, g > 0$ be bounded measurable functions on Ω. Then, for any expectation E,*

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

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 $\Phi'' = 1/t > 0$. Thus, by Jensen's inequality,

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

Theorem 6.

$$S_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^{n} S_{k,f}(\beta)\right]. \tag{3.3}$$

Proof. We denote the canonical density with ρ, so $\rho = \mathrm{e}^{\beta f}/Z_{\beta f}$. Writing $\rho = \rho/E[\rho]$ as a telescopic product and using the previous lemma, we get

$$\begin{aligned} E\left[\rho \ln \frac{\rho}{E[\rho]}\right] &= E\left[\rho \ln \prod_{k=1}^{n} \frac{E_1 \cdots E_{k-1}[\rho]}{E_1 \cdots E_{k-1} E_k[\rho]}\right] \\ &= \sum E\left[E_1 \cdots E_{k-1}[\rho] \ln \frac{E_1 \cdots E_{k-1}[\rho]}{E_1 \cdots E_{k-1}[E_k[\rho]]}\right] \\ &\leq \sum E\left[\rho \ln \frac{\rho}{E_k[\rho]}\right] = E\left[\sum E_k\left[\rho \ln \frac{\rho}{E_k[\rho]}\right]\right]. \end{aligned}$$

From the definition of ρ, we then obtain

$$\begin{aligned} S_f(\beta) &= \beta E_{\beta f}[f] - \ln Z_{\beta f} = E\left[\rho \ln \frac{\rho}{E[\rho]}\right] \leq E\left[\sum E_k\left[\rho \ln \frac{\rho}{E_k[\rho]}\right]\right] \\ &= E\left[\sum_{k=1}^{n}\left(E_k\left[\frac{\mathrm{e}^{\beta f}}{Z_{\beta f}} \ln \frac{\mathrm{e}^{\beta f}}{Z_{\beta f}}\right] - E_k\left[\frac{\mathrm{e}^{\beta f}}{Z_{\beta f}}\right] \ln E_k\left[\frac{\mathrm{e}^{\beta f}}{Z_{\beta f}}\right]\right)\right] \\ &= Z_{\beta f}^{-1} \sum_{k=1}^{n} E[E_k[\mathrm{e}^{\beta f}] S_{k,f}(\beta)] = Z_{\beta f}^{-1} \sum_{k=1}^{n} E[\mathrm{e}^{\beta f} S_{k,f}(\beta)] \qquad \text{by (3.1)} \\ &= E_{\beta f}\left[\sum_{k=1}^{n} S_{k,f}(\beta)\right]. \end{aligned}$$

$\square$

3.3. The Efron–Stein–Steele inequality

Combining (3.3) with Theorem 3 and dividing by β^2, we obtain

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

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

$$\sigma^2[f] \leq E\left[\sum_k \sigma_k^2[f]\right],$$

which is the well-known Efron–Stein–Steele inequality [21]. Observe that we may drop the assumption $f \in L_\infty[\mu]$, but we still require the existence of exponential moments in an interval containing zero, so the inequality so derived is slightly weaker than the inequality in [21].

3.4. A modified logarithmic Sobolev inequality

Suppose we have a sequence of functions $f_k \in \mathcal{A}_k$, so that f_k is independent of the kth coordinate. Combining (3.3) with Lemma 4 and using the identity $E_{\beta f} E_{k,\beta f} = E_{\beta f}$, we obtain

$$S_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^n \psi\big(-\beta(f - f_k)\big)\right], \tag{3.4}$$

which is the modified logarithmic Sobolev inequality proposed by Massart [13,14]. Many consequences of this powerful inequality have been explored (e.g., [3–5,13,16]). Here we will concentrate on the consequences of combining the tensorization inequality with the fluctuation representation of entropy in Theorem 3. Since the fluctuation representation is an identity, this combination is stronger than (3.4) and leads to some results that apparently cannot be recovered from (3.4). We will also re-derive some results that can be derived from (3.4) in cases where we believe that the proposed method gives some additional insight.

3.5. Conditional thermal variance and exponential concentration

Theorems 1, 6 and 3 (applied to the conditional entropy) form the backbone of the proposed method. Combining them, we obtain the following generic concentration result:

Theorem 7. *For any $\beta > 0$ we have the entropy bound*

$$S_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^n \int_0^\beta \int_t^\beta \sigma^2_{k,sf}[f]\,\mathrm{d}s\,\mathrm{d}t\right],$$

the bound on the log-Laplace transform

$$\ln E\big[\mathrm{e}^{\beta(f-Ef)}\big] = \beta \int_0^\beta \frac{S_f(\gamma)}{\gamma^2}\,\mathrm{d}\gamma$$

and the concentration inequality

$$\Pr\{f - Ef > t\} \le \exp\left(\beta \int_0^\beta \frac{S_f(\gamma)}{\gamma^2}\,\mathrm{d}\gamma - \beta t\right).$$

The obvious strategy is to start by bounding the conditional thermal variance $\sigma^2_{k,sf}[f]$. Typically, this leads to considerable simplifications and we will follow this method in the sequel.

4. Two classical concentration inequalities

We begin with the derivation of two classical results: the bounded difference inequality and a similar result, which reduces to the familiar Bennett inequality when f is the sum of its arguments. These inequalities are not new, but they are very useful. We obtain them in their strongest forms and they provide a good illustration of our proposed method.

For $a, b \in \mathbb{R}$, $a < b$ define $\zeta_{a,b} : \mathbb{R} \to \mathbb{R}$ by

$$\zeta_{a,b}(t) = (b - t)(t - a).$$

We state some elementary facts without proof.

Lemma 8. (i) *If X is a random variable with values in $[a, b]$, then*

$$\sigma^2[X] \le (b - EX)(EX - a) = \zeta_{a,b}(EX) \le \frac{(b-a)^2}{4}.$$

(ii) *The function $\zeta_{a,b}$ is non-increasing in $[(a + b)/2, \infty)$.*

4.1. The bounded difference inequality

By Lemma 8(i), we get for all $s \in \mathbb{R}$ that $\sigma^2_{k,sf}[f] \le \mathrm{ran}^2_k(f)/4$, so by the first conclusion of Theorem 7,

$$S_f(\gamma) \le \frac{1}{4}\int_0^\gamma \int_t^\gamma E_{\gamma f}\left[\sum_{k=1}^n \mathrm{ran}^2_k(f)\right] \mathrm{d}s\,\mathrm{d}t \le \frac{\gamma^2}{8} E_{\gamma f}[R^2(f)], \tag{4.1}$$

where we introduced the abbreviation $R^2(f) := \sum_{k=1}^n \mathrm{ran}^2_k(f)$. Bounding the thermal expectation by the uniform norm, we obtain from the third conclusion of Theorem 7 that for all $\beta > 0$

$$\Pr\{f - Ef > t\} \le \exp\left(\beta \int_0^\beta \frac{S_f(\gamma)\,\mathrm{d}\gamma}{\gamma^2} - \beta t\right) \le \exp\left(\frac{\beta^2}{8}\|R^2(f)\|_\infty - \beta t\right).$$

Substitution of the minimizing value $\beta = 4t/\|R^2(f)\|_\infty$ gives

$$\Pr\{f - Ef > t\} \le \exp\left(\frac{-2t^2}{\|R^2(f)\|_\infty}\right),$$

which is the well-known bounded difference inequality (with correct exponent) in the strong version given by McDiarmid [19], Theorem 3.7, where the supremum is outside of the sum of squared conditional ranges. Note that the result is vacuous for $f \notin L_\infty[\mu]$.

4.2. A Bennett–Bernstein concentration inequality

The proof of the bounded difference inequality relied on bounding the thermal variance $\sigma_{k,\beta f}(f)$ uniformly in β, using constraints on the conditional range 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 [19], Theorem 3.8. To state it, we abbreviate the sum of conditional variances of f as

$$\Sigma^2(f) = \sum \sigma_k^2(f).$$

Again, we start with a bound on the thermal variance.

Lemma 9. *Assume $f - Ef \leq 1$. Then, for $\beta > 0$,*

$$\sigma_{\beta f}^2(f) \leq \mathrm{e}^{\beta}\sigma^2(f).$$

Proof.

$$\begin{aligned}
\sigma_{\beta f}^2(f) = \sigma_{\beta(f-Ef)}^2(f - Ef) &= E_{\beta(f-Ef)}[(f-Ef)^2] - \big(E_{\beta(f-Ef)}[f - Ef]\big)^2 \\
&\leq E_{\beta(f-Ef)}[(f-Ef)^2] = \frac{E[(f-Ef)^2\mathrm{e}^{\beta(f-Ef)}]}{E[\mathrm{e}^{\beta(f-Ef)}]} \\
&\leq E\big[(f-Ef)^2\mathrm{e}^{\beta(f-Ef)}\big] \qquad \text{use Jensen on denominator} \\
&\leq \mathrm{e}^{\beta}E[(f-Ef)^2] \qquad \text{use hypothesis.}
\end{aligned}$$

□

Next we bound the total entropy $S_f(\beta)$.

Lemma 10. *Assume that $f - E_k f \leq 1$ for all $k \in \{1, \ldots, n\}$. Then, for $\beta > 0$,*

$$S_f(\beta) \leq (\beta\mathrm{e}^{\beta} - \mathrm{e}^{\beta} + 1)E_{\beta f}[\Sigma^2(f)].$$

Proof. Using the first conclusion of Theorem 7 and the previous lemma, we get

$$S_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^{n}\int_0^\beta\int_t^\beta \sigma_{k,sf}^2[f]\,\mathrm{d}s\,\mathrm{d}t\right] \leq \int_0^\beta\int_t^\beta \mathrm{e}^s\,\mathrm{d}s\,\mathrm{d}t\,E_{\beta f}[\Sigma^2(f)].$$

The conclusion follows from the elementary formula

$$\int_0^\beta \int_t^\beta \mathrm{e}^s \,\mathrm{d}s\,\mathrm{d}t = \int_0^\beta (\mathrm{e}^\beta - \mathrm{e}^t)\,\mathrm{d}t = \beta\mathrm{e}^\beta - \mathrm{e}^\beta + 1. \qquad \square$$

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

Theorem 11. *Assume* $f - E_k f \le 1, \forall k$. *Let* $t > 0$ *and denote* $V = \|\Sigma^2(f)\|_\infty$. *Then*

$$\begin{aligned}\Pr\{f - E[f] > t\} &\le \exp\bigl(-V\bigl((1+tV^{-1})\ln(1+tV^{-1}) - tV^{-1}\bigr)\bigr)\\ &\le \exp\biggl(\frac{-t^2}{2V + 2t/3}\biggr).\end{aligned}$$

Proof. Fix $\beta > 0$. Recall the definition of the function ψ in (2.2) and observe that

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

because $(\mathrm{d}/\mathrm{d}\gamma)(\gamma^{-1}(\mathrm{e}^\gamma - 1)) = \gamma^{-2}(\gamma\mathrm{e}^\gamma - \mathrm{e}^\gamma + 1)$ and $\lim_{\gamma\to 0}\gamma^{-1}(\mathrm{e}^\gamma - 1) = 1$. Theorem 7 and Lemma 10 combined with a uniform bound then give

$$\begin{aligned}\ln E\mathrm{e}^{\beta(f-Ef)} &= \beta\int_0^\beta \frac{S_f(\gamma)\,\mathrm{d}\gamma}{\gamma^2}\\ &\le \beta\biggl(\int_0^\beta \frac{\gamma\mathrm{e}^\gamma - \mathrm{e}^\gamma + 1}{\gamma^2}\,\mathrm{d}\gamma\biggr)\|\Sigma^2(f)\|_\infty = \psi(\beta)V.\end{aligned}$$

So, by Markov's inequality, we have for any $\beta > 0$ that $\Pr\{f - E[f] > t\} \le \exp(\psi(\beta)V - \beta t)$. Substitution of $\beta = \ln(1 + tV^{-1})$ gives the first inequality; the second is Lemma 2.4 in [19]. $\square$

Observe that f is assumed bounded above by the hypotheses of the theorem. The existence of exponential moments $E[\mathrm{e}^{\beta f}]$ is needed only for $\beta \ge 0$, so the assumption $f \in L_\infty[\mu]$ can be dropped in this case.

5. Exploiting monotonicity

Sometimes an appropriately chosen bound on the conditional thermal variance $\sigma^2_{k,sf}[f]$ can be shown to have a monotonicity property in the variable s, which can be used to find a bound uniform in the the region of integration. The remaining part of the fluctuation integral then just becomes $\beta^2/2$, which leads to sub-Gaussian tail estimates, just as for the bounded difference inequality. In this section, we give three examples.

5.1. Functions with large conditional expectations

The following is our first novel result, the proof of which is hardly more difficult than that of the bounded difference inequality. It depends on the assumption that the conditional expectations are consistently in the upper halves of the conditional ranges for all k and all configurations $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$ of the conditioning data. If this condition is met, the result is much stronger than the bounded difference inequality, and, for large deviations t, also much stronger than Bennett's inequality.

Theorem 12. *Suppose that*

$$E_k[f] \geq \frac{\sup_k f + \inf_k f}{2} \qquad \forall k \in \{1, \dots, n\} \tag{5.1}$$

and let

$$A = \left\| \sum_k \Big(\sup_k f - E_k[f] \Big) \Big(E_k[f] - \inf_k f \Big) \right\|_\infty .$$

Then

$$\Pr\{ f - Ef > t \} \leq \mathrm{e}^{-t^2/(2A)}.$$

Proof. By Lemma 2, the function $\beta \mapsto E_{k,\beta f}[f]$ is non-decreasing, so for $\beta \geq 0$ we have

$$E_k[f] \in \left[\frac{\sup_k f + \inf_k f}{2}, E_{k,\beta f}[f] \right].$$

Since the function $\zeta_{\inf_k f, \sup_k f}$ (of Lemma 8) is non-increasing in this interval, we get

$$\sigma^2_{k,\beta f}(f) \leq \zeta_{\inf_k f, \sup_k f}(E_{k,\beta f}[f]) \leq \zeta_{\inf_k f, \sup_k f}(E_k[f]),$$

and, from the first conclusion of Theorem 7,

$$\begin{aligned} S_f(\beta) &\leq E_{\beta f}\left[\sum_{k=1}^n \int_0^\beta \int_t^\beta \sigma^2_{k,sf}[f] \, \mathrm{d}s \, \mathrm{d}t \right] \leq \frac{\beta^2}{2} E_{\beta f}\left[\sum_{k=1}^n \zeta_{\inf_k f, \sup_k f}(E_k[f]) \right] \\ &\leq \frac{\beta^2}{2} \left\| \sum_{k=1}^n \zeta_{\inf_k f, \sup_k f}(E_k[f]) \right\|_\infty = \frac{\beta^2 A}{2}. \end{aligned}$$

The result now follows as in the proof of the bounded difference inequality. □

5.2. Monotonicity of variational bounds on the thermal variance

A related strategy first finds a simple variational bound on the conditional thermal variance. We have

$$\sigma^2[g] = \min_{t \in \mathbb{R}} E[(g-t)^2] \leq E[(g-c)^2] \qquad \forall c \in \mathbb{R}.$$

Applied to the conditional thermal variance, this translates to

$$\sigma^2_{k,\beta f}[f] \le E_{k,\beta f}[(f - f_k)^2] \qquad \forall f_k \in \mathcal{A}_k. \tag{5.2}$$

We will use $\inf_k f$ for f_k and combine this observation with the following.

Proposition 13. *The function $\beta \mapsto E_{k,\beta f}[(f - \inf_k f)^2]$ is non-decreasing.*

Proof. Write $h = f - \inf_k f$ and define a real function ξ by $\xi(t) = (\max\{t, 0\})^2$. Since $h \ge 0$, we have

$$E_{k,\beta f}\Big[\Big(f - \inf_k f\Big)^2\Big] = E_{k,\beta(f-\inf_k f)}\Big[\Big(f - \inf_k f\Big)^2\Big] = E_{k,\beta h}[\xi(h)].$$

By Lemma 2, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\beta} E_{\beta h}[\xi(h)] = E_{\beta h}[\xi(h)h] - E_{\beta h}[\xi(h)]E_{\beta h}[h] \ge 0,$$

where the last inequality uses the well-known fact that for any expectation $E[\xi(h)h] \ge E[\xi(h)]E[h]$ whenever ξ is a non-decreasing function. □

A first consequence is a lower tail bound somewhat similar to Bernstein's inequality, Theorem 11.

Theorem 14. *Let $t > 0$ and denote*

$$W = \Big\| \sum_k E_k\Big(f - \inf_k f\Big)^2 \Big\|_\infty .$$

Then

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

Proof. We use inequality (5.2) and Proposition 13 to get for $s \ge 0$

$$\sigma^2_{k,-sf}[f] \le E_{k,-sf}\Big[\Big(f - \inf_k f\Big)^2\Big] \le E_k\Big[\Big(f - \inf_k f\Big)^2\Big].$$

We therefore obtain from Theorem 7

$$\begin{aligned} S_{-f}(\beta) &\le E_{-\beta f}\left[\sum_{k=1}^n \int_0^\beta \int_t^\beta \sigma^2_{k,-sf}[f]\,\mathrm{d}s\,\mathrm{d}t\right] \le \frac{\beta^2}{2} E_{-\beta f}\left[\sum_{k=1}^n E_k\Big(f - \inf_k f\Big)^2\right] \\ &\le \frac{\beta^2 W}{2} \end{aligned}$$

and then proceed as in the proof of the bounded difference inequality. □

If we take the function f to be an average of real random variables, then Theorem 14 reduces to an inequality given in [20] and [15]. In [15] it is argued that for very heterogeneous variables this inequality is superior to Bernstein's inequality. Similar arguments apply to the present, more general case.

When we apply the same method to obtain upper tail bounds we arrive at a surprisingly powerful result. To state it, we introduce worst-case variance proxies, which will play an important role in the sequel.

Definition 1. *Let $g \in L_\infty[\mu]$. The worst-case variance proxy of g is the function $Dg \in L_\infty[\mu]$ defined by*

$$Dg = \sum_k \Big(g - \inf_k g\Big)^2.$$

The function Dg is a local measure of the sensitivity of g to modifications of its individual arguments.

Lemma 15. *We have, for $\beta > 0$,*

$$S_f(\beta) \leq \frac{\beta^2}{2} E_{\beta f}[Df].$$

Proof. We use inequality (5.2) and Proposition 13 to get for $0 \leq s \leq \beta$

$$\sigma^2_{k,sf}[f] \leq E_{k,sf}\Big[\Big(f - \inf_k f\Big)^2\Big] \leq E_{k,\beta f}\Big[\Big(f - \inf_k f\Big)^2\Big].$$

So, using Theorem 7 again,

$$\begin{aligned} S_f(\beta) &\leq E_{\beta f}\left[\sum_{k=1}^n \int_0^\beta \int_t^\beta \sigma^2_{k,sf}[f]\,\mathrm{d}s\,\mathrm{d}t\right] \leq \frac{\beta^2}{2} E_{\beta f}\left[\sum_{k=1}^n E_{k,\beta f}\Big(f - \inf_k f\Big)^2\right] \\ &= \frac{\beta^2}{2} E_{\beta f}\left[\sum_{k=1}^n \Big(f - \inf_k f\Big)^2\right], \end{aligned}$$

where we used the identity $E_{\beta f} E_{k,\beta f} = E_{\beta f}$ in the last equation. □

The usual arguments now immediately lead to the following.

Theorem 16. *With $t > 0$,*

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\|Df\|_\infty}\right).$$

In [16] this result was derived from inequality (3.4), and it is shown that it improves the exponent on upper tail bounds derived from Talagrand's convex distance inequality in many cases, as for shortest travelling salesmen paths, Steiner trees and the eigenvalues of random symmetric matrices. Here we only give one example of how $\|Df\|_\infty$ may be bounded and consider a convex Lipschitz function f defined on the cube $[0,1]^n$. For simplicity, we assume f to be differentiable.

Let $x \in [0,1]^n$ and suppose that for some fixed k there is $y \in [0,1]$ such that $f(\mathbf{x}_{y,k}) \le f(\mathbf{x})$. Then by convexity (using really only the fact that f is separately convex in each coordinate),

$$f(\mathbf{x}) - f(\mathbf{x}_{y,k}) \le \langle \mathbf{x} - \mathbf{x}_{y,k}, \partial f(\mathbf{x}) \rangle_{\mathbb{R}^n} = (x_k - y)\partial_k f(\mathbf{x}) \le |\partial_k f(\mathbf{x})|.$$

We therefore have $f(\mathbf{x}) - \inf_y f(\mathbf{x}_{y,k}) \le |\partial_k f(\mathbf{x})|$ and

$$Df(\mathbf{x}) = \sum_{k=1}^n \Big(f(\mathbf{x}) - \inf_y f(\mathbf{x}_{y,k}) \Big)^2 \le \|\partial f(\mathbf{x})\|_{\mathbb{R}^n}^2 \le \|f\|_{\mathrm{Lip}}^2.$$

In combination with Theorem 16 we obtain upper tail bounds for f with an exponent twice as good as obtained from the convex distance inequality [11], Corollary 4.10, or an earlier application of the entropy method [11], Theorem 5.9.

For a corresponding lower tail bound, we have to use an estimate similar to what was used in the proof of Bennett's inequality.

Lemma 17. *If $f - \inf_k f \le 1, \forall k$, then for $\beta > 0$,*

$$S_{-f}(\beta) \le \psi(\beta) E_{-\beta f}[Df],$$

with ψ defined as in (2.2).

Proof. Let $k \in \{1, \dots, n\}$. We write $h_k := f - \inf_k f$. Then $h_k \in [0,1]$ and for $s \le \beta$

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

We therefore have

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

where we used the formula

$$\int_0^\beta \int_t^\beta \mathrm{e}^{-s} \, \mathrm{d}s \, \mathrm{d}t = 1 - \mathrm{e}^{-\beta} - \beta \mathrm{e}^{-\beta}.$$

Thus, using Theorem 7 and the identity $E_{-\beta f}E_{k,-\beta f} = E_{-\beta f}$,

$$S_{-f}(\beta) \le E_{-\beta f}\left[\sum_k \int_0^\beta \int_t^\beta \sigma^2_{k,-sf}[f]\,\mathrm{d}s\,\mathrm{d}t\right] \le E_{-\beta f}\left[\sum_k \int_0^\beta \int_t^\beta E_{k,-sf}[h_k^2]\,\mathrm{d}s\,\mathrm{d}t\right]$$
$$\le \psi(\beta)E_{-\beta f}\left[\sum_k E_{k,-\beta f}[h_k^2]\right] = \psi(\beta)E_{-\beta f}[Df].$$

□

Lemmas 15 and 17 together with Theorem 1 imply the inequalities

$$\ln E\left[\mathrm{e}^{\beta(f-E[f])}\right] \le \frac{\beta}{2}\int_0^\beta E_{\gamma f}[Df]\,\mathrm{d}\gamma \tag{5.3}$$

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

$$\ln E\left[\mathrm{e}^{\beta(E[f]-f)}\right] \le \frac{\psi(\beta)}{\beta}\int_0^\beta E_{-\gamma f}[Df]\,\mathrm{d}\gamma, \tag{5.4}$$

where in the last inequality we also used the fact that $\gamma \mapsto \psi(\gamma)/\gamma^2$ is non-decreasing. Bounding the thermal expectation with the uniform norm and substitution of $\beta = \ln(1 + t\|Df\|_\infty^{-1})$ gives the following lower tail bound that can also be found in [16].

Theorem 18. *If $f - \inf_k f \le 1$ for all k, then for $t > 0$,*

$$\Pr\{Ef - f > t\} \le \exp\left(-\|Df\|_\infty\left(\left(1 + \frac{t}{\|Df\|_\infty}\right)\ln\left(1 + \frac{t}{\|Df\|_\infty}\right) - \frac{t}{\|Df\|_\infty}\right)\right)$$
$$\le \exp\left(\frac{-t^2}{2\|Df\|_\infty + 2t/3}\right).$$

The two inequalities (5.3) and (5.4) are the keys to obtaining concentration inequalities in terms of the worst-case variance proxy Df. Both results can also be deduced from Massart's inequality (3.4) as shown in [16]. We do not claim that the derivations given above are per se superior. We presented them because they follow the same principles as the proofs of the other results given above (the bounded difference inequality and Theorems 11, 12 and 14), which do not follow from inequality (3.4).

6. Self-boundedness and canonical decoupling

We conclude by presenting two general principles to extend the utility of the proposed method. All the above applications of Theorem 7 involved a chain of inequalities of the form 5

$$S_{\varepsilon f}(\gamma) \le E_{\varepsilon\gamma f}\left[\sum_{k=1}^n \int_0^\gamma \int_t^\gamma \sigma^2_{k,\varepsilon sf}[f]\,\mathrm{d}s\,\mathrm{d}t\right] \le \xi(\gamma)E_{\varepsilon\gamma f}[G(f)],$$

where $\varepsilon = 1$ for upper tail results and $\varepsilon = -1$ for lower tail results, ξ is some non-negative real function and $G(f)$ is some function on Ω derived from f. For the bounded difference inequality, for example, $\xi(\gamma) = \gamma^2/8$ and $G = R^2(f)$; for the Bennett inequality $\xi(\gamma) = \gamma e^{\gamma} - e^{\gamma} + 1$ and $G(f) = \Sigma^2(f)$; for Theorem 16 we had $\xi(\gamma) = \gamma^2/2$ and $G(f) = Df$; while for the corresponding lower tail bound, Theorem 18, we had $\xi(\gamma) = \psi(\gamma)$ and also $G(f) = Df$, etc. Theorem 7 is then invoked to conclude that

$$\ln E e^{\varepsilon\beta(f-Ef)} \le \beta \int_0^\beta \frac{\xi(\gamma)}{\gamma^2} E_{\varepsilon\gamma f}[G(f)]\,d\gamma \le \beta \|G(f)\|_\infty \int_0^\beta \frac{\xi(\gamma)\,d\gamma}{\gamma^2}. \tag{6.1}$$

Here the uniform estimate $E_{\varepsilon\beta f}[G(f)] \le \|G(f)\|_\infty$, while being very simple, is somewhat loose. We now sketch how it can sometimes be avoided by exploiting special properties of the thermal expectation.

6.1. Self-boundedness

The first possibility we consider is that the function $G(f)$ can be bounded in terms of the function f itself, a property referred to as *self-boundedness* [4]. For example, if $G(f) \le f$, then $E_{\gamma f}[G(f)] \le E_{\gamma f}[f] = (d/d\gamma) \ln E[\exp(\gamma f)]$, and if the function ξ has some reasonable behavior, then the first integral in (6.1) above can be bounded by partial integration or even more easily. As an example, we apply this idea in the setting of Theorems 16 and 18.

Theorem 19. *Suppose that there are non-negative numbers a, b such that $Df \le af + b$. Then, for $t > 0$, we have*

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

If, in addition, $a \ge 1$ and $f - \inf_k f \le 1, \forall k \in \{1, \dots, n\}$, then

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

Proof. We only prove the lower tail bound; for the upper tail we refer to [16]. As for the lower tail, it follows from (5.4) and Lemma 2 that

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

Rearranging gives

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

where one verifies that for $\beta > 0$ and $a \geq 1$ we have $\psi(\beta)(1 + a\beta^{-1}\psi(\beta))^{-1} \leq \beta^2/2$. The usual analysis with Markov's inequality and optimization in β conclude the proof. □

Recently Boucheron *et al.* [4] have given a refined version of this result, where the condition $a \geq 1$ is improved to $a \geq 1/3$ for the lower tail. There they also show that Theorems 19 and 16 together suffice to derive a version of the convex distance inequality that differs from Talagrand's original result only in that it has an inferior exponent.

It must be stressed that the same method of proof can be used to yield self-bounded versions of all concentration inequalities derived from Theorem 7, such as the bounded-difference and Bennett inequalities.

6.2. Decoupling

A second method to avoid the uniform bound on the thermal expectation uses decoupling. Recall that for any two probability measures ν and μ and a measurable function g we have

$$E_{x\sim\nu}[g(x)] \leq KL(\nu, \mu) + \ln E_{x\sim\mu} e^{g(x)},$$

which can be regarded as an instance of convex duality and easily verified directly from the definition of the Kullback–Leibler divergence. Applying this inequality when ν is the canonical ensemble and μ is the a priori measure, we obtain for any $\theta > 0$

$$S_{\varepsilon f}(\beta) \leq \xi(\beta)\theta^{-1} E_{\varepsilon\beta f}[\theta G(f)] \leq \xi(\beta)\theta^{-1}\big(S_{\varepsilon f}(\beta) + \ln E[\exp(\theta G(f))]\big).$$

For values of β and θ where $\theta > \xi(\beta)$ we obtain

$$S_{\varepsilon f}(\beta) \leq \frac{\xi(\beta)}{\theta - \xi(\beta)} \ln E[\exp(\theta G(f))].$$

Hence, if we can control the upwards deviations of $G(f)$ (or some suitable bound thereof), we obtain concentration inequalities for f in terms of the expectation of $G(f)$ (or the bound thereof). Again, this method, which was proposed in [3], can be applied to all the versions of $G(f)$ we introduced above and combined with all methods to control the upwards deviation of $G(f)$, which leads to a proliferation of concentration inequalities. Perhaps not all of these deserve to be documented. We just quote a corresponding result in [17] that uses $G(f) = Df$ and combines with self-boundedness.

Theorem 20. *Suppose that there is $g \in L_\infty[\mu]$ and $a \geq 1$ such that $0 \leq f \leq g$, $Df \leq ag$ and $Dg \leq ag$. Then, for $t > 0$,*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{4aE[g] + 3at/2}\right).$$

If, in addition, $f - \inf_k f \leq 1$ *for all* k, *then*

$$\Pr\{Ef - f > t\} \leq \exp\left(\frac{-t^2}{4aE[g] + at}\right).$$

In [17] the theorem is used to show that the concentration of eigenvalues (f) of the Gram matrix of a sample of independent, bounded random vectors in a Hilbert space is controlled by the size of the largest eigenvalue (g).

7. A glossary of notation

We conclude with a tabular summary of notation.

$\Omega = \prod_{k=1}^n \Omega_k$	underlying (product-) probability space.
$\mu = \bigotimes_{k=1}^n \mu_k$	(product-) probability measure on Ω.
X_k	random variable distributed as μ_k in Ω_k.
$f \in L_\infty[\mu]$	fixed function (negative energy) under investigation.
$g \in L_\infty[\mu]$	generic function.
$E[g] = \int_\Omega g \, \mathrm{d}\mu$	expectation of g in μ.
$\sigma^2[g] = E[(g - E[g])^2]$	variance of g in μ.
$\beta = 1/T$	inverse temperature.
$E_{\beta f}[g] = E[g\mathrm{e}^{\beta f}]/E[\mathrm{e}^{\beta f}]$	thermal expectation of g.
$Z_{\beta f} = E[\mathrm{e}^{\beta f}]$	partition function.
$S_f(\beta) = \beta E_{\beta f}[f] - \ln Z_{\beta f}$.	canonical entropy.
$A_f(\beta) = \frac{1}{\beta} \ln Z_{\beta f}$	Helmholtz free energy.
$\sigma^2_{\beta f}(g) = E_{\beta f}[(g - E_{\beta f}[g])^2]$	thermal variance of g.
$\psi(t) = \mathrm{e}^t - t - 1$	
$\mathbf{x}_{y,k}$	vector $\mathbf{x} \in \Omega$ with x_k replaced by $y \in \Omega_k$.
$E_k[g](\mathbf{x}) = \int_{\Omega_k} g(\mathbf{x}_{y,k}) \, \mathrm{d}\mu_k(y)$	conditional expectation.
$\mathcal{A}_k \subset L_\infty[\mu]$	functions independent of kth variable.
$Z_{k,\beta f} = E_k[\mathrm{e}^{\beta f}]$	conditional partition function.
$E_{k,\beta f}[g] = Z^{-1}_{k,\beta f} E_k[g\mathrm{e}^{\beta f}]$	conditional thermal expectation.
$S_{k,f}(\beta) = \beta E_{k,\beta f}[g] - \ln Z_{k,\beta f}$	conditional entropy.
$\sigma^2_{k,\beta f}[g] = E_{k,\beta f}[(g - E_{k,\beta f}[g])^2]$	conditional thermal variance.
$\sigma^2_k[g] = E_k[(g - E_k[g])^2]$	conditional variance.
$(\sup_k g)(\mathbf{x}) = \sup_{y \in \Omega_k} g(\mathbf{x}_{y,k})$	conditional supremum.
$(\inf_k g)(\mathbf{x}) = \inf_{y \in \Omega_k} g(\mathbf{x}_{y,k})$	conditional infimum.
$\mathrm{ran}_k(g) = \sup_k g - \inf_k g$	conditional range.
$R^2(g) = \sum_k \mathrm{ran}^2_k(g)$	sum of conditional square ranges.
$\Sigma^2(g) = \sum_k \sigma^2_k[g]$	sum of conditional variances.
$Dg = \sum_k (g - \inf_k g)^2$	worst case variance proxy.

References

[1] Bernstein, S. (1946). *Theory of Probability*. Gastehizdal Publishing House, Moscow.

[2] Boltzmann, L. (1877). Ueber die Beziehung zwischen dem zweiten Hauptsatz der mechanischen Waermetheorie und der Wahrscheinlichkeitsrechnung respektive den Saetzen ueber das Waermegleichgewicht. *Wiener Berichte* **76** 373–435.

[3] Boucheron, S., Lugosi, G. and Massart, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.* **31** 1583–1614. MR1989444

[4] Boucheron, S., Lugosi, G. and Massart, P. (2009). On concentration of self-bounding functions. *Electron. J. Probab.* **14** 1884–1899. MR2540852

[5] Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris* **334** 495–500. MR1890640

[6] Chebychev, P. (1874). Sur les valeurs limites des intégrales. *J. Math. Pures Appl.* **19** 157–160.

[7] Gibbs, J.W. (1902). *Elementary Principles in Statistical Mechanics with Especial Reference to the Rational Foundation of Thermodynamics*. Yale, CT: Yale Univ. Press.

[8] Gross, L. (1975). Logarithmic Sobolev inequalities. *Amer. J. Math.* **97** 1061–1083. MR0420249

[9] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[10] Ledoux, M. (1996). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.* **1** 63–87 (electronic). MR1399224

[11] Ledoux, M. (2001). *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs* **89**. Providence, RI: Amer. Math. Soc. MR1849347

[12] Lieb, E.H. (1975). Some convexity and subadditivity properties of entropy. *Bull. Amer. Math. Soc.* **81** 1–13. MR0356797

[13] Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863–884. MR1782276

[14] Massart, P. (2000). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.* (6) **9** 245–303. MR1813803

[15] Maurer, A. (2003). A bound on the deviation probability for sums of non-negative random variables. *JIPAM. J. Inequal. Pure Appl. Math.* **4** 1–6 (electronic). MR1965995

[16] Maurer, A. (2006). Concentration inequalities for functions of independent variables. *Random Structures Algorithms* **29** 121–138. MR2245497

[17] Maurer, A. (2010). Dominated concentration. *Statist. Probab. Lett.* **80** 683–689. MR2595147

[18] McAllester, D. and Ortiz, L. (2002). Concentration inequalities for the missing mass and for histogram rule error. *J. Mach. Learn. Res.* **4** 895–911.

[19] McDiarmid, C. (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics. Algorithms Combin.* **16** 195–248. Berlin: Springer. MR1678578

[20] Pinelis, I.S. and Utev, S.A. (1989). Sharp exponential estimates for sums of independent random variables. *Theory Probab. Appl.* **34** 340–346. MR1005745

[21] Steele, J.M. (1986). An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758. MR0840528

[22] Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. Inst. Hautes Études Sci.* **81** 73–205. MR1361756

[23] Talagrand, M. (1996). A new look at independence. *Ann. Probab.* **24** 1–34. MR1387624

Received August 2009 and revised May 2010