Entropy and Concentration

Summer school DIMA, 2017, A. Maurer

July 23, 2017

Contents

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.

The class of problems we study in this course can be described as follows: 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

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

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 [7] in the late 19th century for additive functions

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

In this case great simplifications are possible, because then variance and moment generating function are additive, which is not true in general. We will not dwell on additive functions, but develop a method to handle more general, non-additive functions. The bounds so derived will then give the most important results for additive functions, like Hoeffding's and Bennett's inequality, as easy corollaries.

The method we use, the so-called *entropy method*, is related to statistical mechanics, as developed by Boltzmann [2] and Gibbs [9]. We give an exposition of the method in Section 2 and compress it into a toolbox to derive concentration inequalities.

In Section 3 we will then use this method to prove two classical concentration inequalities, the bounded difference inequality and a generalisation 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 Section 4 we adress more difficult problems. The bounded difference inequality is used to prove the famous Gaussian concentration inequality of Tsirelson, Ibragimov and Sudakov. 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 Section 5 we describe some of the more advanced techniques, namely self-boundedness and decoupling. As examples we give sub-Gaussian lower tail bounds for convex Lipschitz functions and derive an exponential inequality for the suprema of empirical processes.

An appendix contains a summary of notation in tabular form.

I hope the course will stimulate interest in concentration inequalities. For further study I very much recommend the monographs by Ledoux [10] and Boucheron, Lugosi and Massart [5], and the excellent overview article by McDiarmid [13]. These works also have a much broader view on the subject than the narrow perspective of this course.

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 the basis of 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. The notation which we introduce along the way is also summarized in the appendix.

2 The entropy method

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 important 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

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

■

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 components the variance is just the sum of the component variances, but this doesn't work for general functions.The idea of Chebychev's inequality obviously extends to other even centered moments $E\left[(f - E[f])^{2p}\right]$.

For our purpose the most important corollary of Markov's inequality is the *exponential moment method*, an idea appearantly due to Bernstein [1].

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 then the partition function is the product of the partition functions corresponding to these components, and its logarithm (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, which are the object of this course. The approach 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 $-f(\mathbf{x})$ 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 $E[\rho] = 1$, recall that the Kullback-Leibler divergence $KL(\rho d\mu, d\mu)$ of $\rho d\mu$ to $d\mu$ is defined as

$$KL(\rho d\mu, d\mu) = E\left[\rho \ln \frac{\rho_1}{1}\right] = E[\rho \ln \rho].$$

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. Obviously we can assume $\beta = 1$. 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 $E[\rho \ln \rho] = -E[\rho \ln \phi] E = -E_\rho[\ln \phi]$ and with Jensen's inequality

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

On the other hand

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

■

The result exhibits the functions $f \in \mathcal{A} \mapsto \ln E\left[e^f\right]$ and $\rho \mapsto E[\rho \ln \rho]$ as a pair of convex conjugates.

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* in statistical mechanics, sometimes also the *canonical ensemble*. It describes 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. For any constant c we have the obvious and important identity $E_{\beta(f+c)}[g] = E_{\beta f}[g]$.

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,

$$\operatorname{Ent}_f(\beta) = E\left[\rho_{\beta f} \ln \rho_{\beta f}\right] = \beta E_{\beta f}[f] - \ln Z_{\beta f}. \tag{1}$$

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 giving the same expected value for the energy as $\rho_{\beta f}$, so that $E[\rho f] = E_{\beta f}[f]$ then

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

The thermal measure $d\mu_{\beta f} = Z_{\beta f}^{-1} e^{\beta f} d\mu$ therefore minimizes the information gain relative to the a priori measure $d\mu$, given the expected value of f. For a fixed value of the internal energy $-E_{\beta f}[f]$, the choice of the canonical ensemble is an admission of maximal ignorance.

The connection of entropy and concentration is expressed in the following result.

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

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

and, for $t \geq 0$,

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

Proof. For $\beta \neq 0$ define a function

$$A_f(\beta) = \frac{1}{\beta} \ln Z_{\beta f} = \frac{1}{\beta} \ln E\left[e^{\beta f}\right]. \tag{2}$$

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 $A_f(0) = E[f]$. Also

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

By the fundamental theorem of calculus

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

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

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

■

In statistical physics the quantity $A_f(\beta)$ as defined in (2) is called the *free energy* corresponding to the Hamiltonian (energy function) $H = -f$ and temperature $T \approx \beta^{-1}$. Dividing (1) by β and writing $U = E_{\beta f}[f]$, we recover the classical thermodynamic relation

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

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.

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 $\sigma^2_{\beta f}(g)$ and defined by

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

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

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

Lemma 6 *The following formulas hold for $f \in \mathcal{A}$*

1. $\frac{d}{d\beta}(\ln Z_{\beta f}) = E_{\beta f}[f]$.

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. $\frac{d^2}{d\beta^2}(\ln Z_{\beta f}) = \frac{d}{d\beta} E_{\beta f}[f] = \sigma^2_{\beta f}[f]$.

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 [12].

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

$$\mathrm{Ent}_f\left(\beta\right) = \int_0^\beta \int_t^\beta \sigma_{sf}^2\left[f\right] ds\ dt.$$

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

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

and

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

which we subtract to obtain

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

■

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, where all the X_k are assumed to be mutually 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 $F : \Omega \to \Xi$ the *substitution operator* S_y^k acts on F by

$$\left(S_y^k F\right)\left(x_1, ..., x_n\right) = F\left(x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n\right),$$

so the k-th argument is simply replaced by y. Note that $\left(S_y^k F\right)\left(x_1, ..., x_n\right) = F\left(S_y^k\left(x_1, ..., x_n\right)\right)$, if $\left(x_1, ..., x_n\right)$ is viewed as the identity function on Ω. In general $S_y^k\left(F \circ G\right) = F \circ S_y^k\left(G\right)$ and when restricted to functions in the algebra $\mathcal{A}$ the operator S_y^k is a homomorphism (linear and multiplicative). Since for $f \in \mathcal{A}$ the function $S_y^k f$ is independent of x_k (which had been replaced by y) we see that S_y^k is also a projection of $\mathcal{A}$ onto $\mathcal{A}_k$.

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

$$D_{y,y'}^k f = S_y^k f - S_{y'}^k f \text{ for } f \in \mathcal{A}.$$

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 *conditional range*.

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

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

The operator E_k 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 amongst each other, and for $h \in \mathcal{A}$ and $g \in \mathcal{A}_k$ we have

$$E\left[\left[E_k h \right] g \right] = E\left[E_k \left[hg \right] \right] = E\left[hg \right]. \tag{3}$$

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 $E_{k,\beta f}[g] = Z^{-1}_{k,\beta f} E_k\left[g e^{\beta f} \right]$ 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 $\sigma^2_{k,\beta f}[g] = E_{k,\beta f}\left[\left(g - E_{k,\beta f}[g] \right)^2 \right]$ for $g \in \mathcal{A}$. As $\beta \to 0$ this becomes
- The conditional variance $\sigma^2_k[g] = E_k\left[\left(g - E_k[g] \right)^2 \right]$ for $g \in \mathcal{A}$.

The previously established relations hold also for the corresponding conditional quantities, in particular the conclusion of Theorem 7

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

The following lemma 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 $E\left[E_k[h] g \right] = E\left[h E_k[g] \right]$*

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

■

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

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

Equality no longer holds in the interacting, non-linear 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.

In 1975 Elliott Lieb [11] gave a proof of this result, which was probably known some time before, at least in the classical setting relevant to our arguments.

Lemma 9 *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,

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

■

Lemma 10 *Let $\rho \in A$, $\rho > 0$. 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 $E^k[.] = E_1E_2...E_k[.]$ with E^0 being the identity map on $\mathcal{A}$ and $E^n = E$. 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 9, using $E\left[E^{k-1}\left[.\right]\right] = E\left[.\right]$,

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

■

Theorem 11

$$\mathrm{Ent}_f\left(\beta\right) \leq E_{\beta f}\left[\sum_{k=1}^{n} \mathrm{Ent}_{k,f}\left(\beta\right)\right] \tag{4}$$

Proof. Set $\rho = e^{\beta f}$ in Lemma 10 to get

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

where we used Lemma 8 in the last identity. ■

2.6 Summary of results

The results established sofar (Theorem 5, Theorem 11 and Theorem 7) already constitute a convenient toolbox to prove a number of interesting concentration inequalities. Here is a summary:

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

$$\begin{aligned} \Pr\left\{f - Ef > t\right\} &\leq \exp\left(\beta \int_0^\beta \frac{\mathrm{Ent}_f\left(\gamma\right)}{\gamma^2} d\gamma - \beta t\right) && \textit{(TB1)} \\ \ln E\left[e^{\beta\left(f - Ef\right)}\right] &= \beta \int_0^\beta \frac{\mathrm{Ent}_f\left(\gamma\right)}{\gamma^2} d\gamma && \textit{(TB2)} \\ \mathrm{Ent}_f\left(\beta\right) &\leq E_{\beta f}\left[\sum_{k=1}^{n} \mathrm{Ent}_{k,f}\left(\beta\right)\right] && \textit{(TB3)} \\ \mathrm{Ent}_f\left(\beta\right) &= \int_0^\beta \int_t^\beta \sigma_{sf}^2\left[f\right] ds\ dt && \textit{(TB4)} \\ \mathrm{Ent}_{k,f}\left(\beta\right) &= \int_0^\beta \int_t^\beta \sigma_{k,sf}^2\left[f\right] ds\ dt && \textit{(TB5)} \end{aligned}$$

3 First applications of the entropy method

We now develop some consequences of Theorem 12. First we indicate how it implies the Efron-Stein inequality, a general bound on the variance. Then we continue with the derivation of the bounded difference inequality, one of the simplest concentration inequalities, and perhaps the most useful one. Then we give a Bennett-Bernstein type inequality.

3.1 The Efron-Stein inequality

Combining the fluctation representations (TB4) and (TB5) with the subadditivity (TB3) 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

$$\sigma^2 [f] \leq E\left[\sum_k \sigma_k^2 [f]\right] = E\left[\Sigma^2 (f)\right], \tag{5}$$

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

(5) is the famous Efron-Stein-Steele inequality [16]. 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

We start with the observation that the variance of a 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*

$$\sigma^2 (f) \leq \frac{(b-a)^2}{4}$$

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 elementary 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

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

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

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

Proof. Applied to the conditional thermal variance Lemma 13 gives

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

so combining the subadditivity of entropy (TB3) and the fluctuation representation (TB4) gives

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

Bounding the thermal expectation $E_{\gamma f}$ by the supremum over $\mathbf{x} \in \Omega$ we obtain from the tail-bound (TB1) that for all $\beta > 0$

$$\begin{aligned} \Pr \left\{ f - Ef > t \right\} &\leq \exp \left(\beta \int_0^{\beta} \frac{\mathrm{Ent}_f \left(\gamma \right) d\gamma}{\gamma^2} - \beta t \right) \\ &\leq \exp \left(\frac{\beta^2}{8} \sup_{\mathbf{x} \in \Omega} R^2 \left(f \right) \left(\mathbf{x} \right) - \beta t \right) . \end{aligned}$$

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

It is important to realize, that the conditional range $r_k \left(f \right)$ is a function in $\mathcal{A}_k$ and may depend on all x_i except x_k. The sum $\sum_{k=1}^{n} r_k \left(f \right)^2$ 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 literature one often sees the following weaker result.

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 .

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 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 $\|y\|, \|y'\| \leq c_k$

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

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

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

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

$$\left\|\sum_i X_i\right\| \leq E\left\|\sum_i X_i\right\| + \sqrt{2\sum_i c_i^2 \ln(1/\delta)}. \tag{6}$$

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

$$\left\|\frac{1}{n}\sum_i X_i\right\| \leq \sqrt{\frac{E\left[\|X_1\|^2\right]}{n}} + c_1\sqrt{\frac{2\ln(1/\delta)}{n}} \tag{7}$$

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 $\mathbf{X} = (X_1, ..., X_n)$ 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 $E[f(X)]$, but, as the true distribution of X is unknown, it has to be content with minimizing the empirical surrogate

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

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

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

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 (6) 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. This is the function Rad: $\mathcal{F} \times \mathcal{X}^n$ on defined as

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

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

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

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

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

and thus obtain

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

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

$$E\left[\text{Rad}\left(\mathcal{F}, \mathbf{X}\right)\right] \leq \text{Rad}\left(\mathcal{F}, \mathbf{X}\right) + \sqrt{\frac{2 \ln\left(1/\delta\right)}{n}}. \tag{8}$$

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

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

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 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 [13, Theorem 3.8]. 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.

$$\begin{aligned}
\sigma^2_{\beta f}(f) &= \sigma^2_{\beta(f-Ef)}(f-Ef) = E_{\beta(f-Ef)}\left[(f-Ef)^2\right] - \left(E_{\beta(f-Ef)}[f-Ef]\right)^2 \\
&\leq E_{\beta(f-Ef)}\left[(f-Ef)^2\right] = \frac{E\left[(f-Ef)^2 e^{\beta(f-Ef)}\right]}{E\left[e^{\beta(f-Ef)}\right]} \\
&\leq E\left[(f-Ef)^2 e^{\beta(f-Ef)}\right] \text{ use Jensen on denominator} \\
&\leq e^\beta E\left[(f-Ef)^2\right] \text{ use hypothesis}
\end{aligned}$$

■

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. Using the first conclusion of 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 elementary 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

$$\psi(t) = e^t - t - 1, \tag{9}$$

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

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

So by Markov's inequality we have for any $\beta > 0$ that $\Pr\{f - E[f] > t\} \leq \exp(\psi(\beta)V - \beta t)$. 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 hypotheses 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.

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*

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

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

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

3.6 Vector valued concentration revisited

We look again at the situation of section 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 $f(\mathbf{x})=\|\sum_i x_i\|$ and observe that for $y,y'\in H$, $D^k_{y,y'}f(\mathbf{x})\leq\|y-y'\|$. This implies that $f-E_k[f]\leq 2c$ and also

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

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 $E\|\sum_i X_i\|$. 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 (7), 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 spectral norm of a random matrix, which is 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 [5]). 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 $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$

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

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

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

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

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 by appropriately scaled Rademacher sums $h_K(\boldsymbol{\epsilon})$ and to apply the bounded difference inequality to $f(\mathbf{h}_K(\boldsymbol{\epsilon}))$.

By an approximation argument using convolution with Gaussian kernels of decreasing width it suffices to prove the result if f is C^∞ with $\left|\left(\partial^2/x_i^2\right) f\left(\mathbf{x}\right)\right| \leq B$ for all $\mathbf{x} \in \mathbb{R}^n$ and $i \in \{1, ..., n\}$, where B is a finite, but potentially very 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

$$\begin{aligned} h_K\left(\boldsymbol{\epsilon}\right) &= \frac{1}{\sqrt{K}} \sum_{k=1}^K \epsilon_k, \text{ for } \epsilon \in \{-1,1\}^K \\ \mathbf{h}_K\left(\boldsymbol{\epsilon}\right) &= \left(h_K\left(\boldsymbol{\epsilon}_1\right), ..., h_K\left(\boldsymbol{\epsilon}_n\right)\right) \text{ for } \boldsymbol{\epsilon} = \left(\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n\right) \in \{-1,1\}^{Kn} \\ G_K &= f\left(\mathbf{h}_K\left(\boldsymbol{\epsilon}\right)\right) \text{ for } \boldsymbol{\epsilon} \in \{-1,1\}^{Kn}. \end{aligned}$$

We will use Theorem 14 on the function G_K applied to Rademacher variables $\boldsymbol{\epsilon}$.

Fix a configuration $\boldsymbol{\epsilon} \in \{-1, 1\}^{Kn}$ and let $\mathbf{x} = (x_1, ..., x_n) = \mathbf{h}_K(\boldsymbol{\epsilon})$. For each $i \in \{1, ..., n\}$ we introduce the real function $f_i(t) = S_t^i f(\mathbf{x})$. Since f is C^∞ 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 $|f_i''|$

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

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 $\left(D_{y,y'}^{(i,k)} G_K(\boldsymbol{\epsilon})\right)^2$. Now either y or y' is equal to ϵ_{ik}, so either $S_y^{(i,k)} G_K(\boldsymbol{\epsilon})$ or $S_{y'}^{(i,k)} G_K(\boldsymbol{\epsilon})$ is equal to $G_K(\boldsymbol{\epsilon})$. Without loss of generality we assume the second. Furthermore $S_y^k h_K(\boldsymbol{\epsilon}_i)$ and $h_K(\boldsymbol{\epsilon}_i)$ differ by at most $2/\sqrt{K}$, so

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

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

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

Thus

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

Now let $\boldsymbol{\epsilon}$ be Rademacher variables and $\boldsymbol{\epsilon}'$ iid to $\boldsymbol{\epsilon}$. From Theorem 14 we conclude from $f(\mathbf{h}_K(\boldsymbol{\epsilon})) = G_K(\boldsymbol{\epsilon})$ that

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

The conclusion now follows from the central limit theorem since $h_K(\boldsymbol{\epsilon}) \to \mathbf{X}$ 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

$$\frac{d}{d\beta}E_{\beta f}\left[\psi\left(f\right)\right]=E_{\beta f}\left[\psi\left(f\right)f\right]-E_{\beta f}\left[\psi\left(f\right)\right]E_{\beta f}\left[f\right]. \tag{10}$$

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

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

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

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

■

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 $\sigma_{sf}^{2}\left(f\right)\leq E_{\beta f}\left[h\right]$.

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

$$\begin{aligned}D^{2}f &= \sum_{k}\left(f-\inf_{y\in\Omega_{k}}S_{y}^{k}f\right)^{2}\\ \text{and } V_{+}^{2}f &= \sum_{k}E_{y\sim\mu_{k}}\left[\left(\left(f-S_{y}^{k}f\right)_{+}\right)^{2}\right].\end{aligned}$$

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 $h_{k}=E_{y\sim\mu_{k}}\left[\left(f-S_{y}^{k}f\right)_{+}^{2}\right]$, 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 $k\in\{1,...,n\}$

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

Using (TB3) and (TB5) we get

$$\begin{aligned}\text{Ent}_{f}\left(\beta\right) &\leq \int_{0}^{\beta}\int_{t}^{\beta}\sum_{k}E_{\beta f}\left[\sigma_{k,sf}^{2}\left(f\right)\right]dsdt\\ &\leq \int_{0}^{\beta}\int_{t}^{\beta}\sum_{k}E_{\beta f}\left[E_{k,\beta f}\left[h_{k}\right]\right]dsdt\\ &= \int_{0}^{\beta}\int_{t}^{\beta}\sum_{k}E_{\beta f}\left[h_{k}\right]dsdt\\ &= \frac{\beta^{2}}{2}E_{\beta f}\left[V^{+}\left(f\right)\right],\end{aligned}$$

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

The usual arguments involving (TB1) and an optimization in β now immediately lead to

Theorem 27 *With $t>0$*

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

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 $\beta>0$*

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

with ψ defined as in (9).

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

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

We therefore have

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

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 $E_{-\beta f}E_{k,-\beta f}=E_{-\beta f}$

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

■

Lemma 26 and Lemma 28 together with (TB2) imply the inequalities

$$\ln E\left[e^{\beta(f-E[f])}\right]\leq\frac{\beta}{2}\int_0^\beta E_{\gamma f}\left[V_+^2f\right]d\gamma. \tag{11}$$

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

$$\ln E\left[e^{\beta(E[f]-f)}\right] \leq \frac{\psi(\beta)}{\beta}\int_0^\beta E_{-\gamma f}\left[D^2 f\right] d\gamma, \tag{12}$$

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 $\Delta := \sup_{\mathbf{x}\in\mathbf{\Omega}} D^2 f(\mathbf{x})$*

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

4.3 Convex Lipschitz functions

In section 4.1 we gave a sub-gaussian bound for general Lipschitz functions of standard Gaussian processes. Now we give 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). For simplicity we assume f to be differentiable.

Theorem 30 *Let $\Omega_k = [0,1]$ 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*

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

Proof. 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. ■

4.4 The spectral norm of a random matrix

For $\mathbf{x} \in [-1,1]^{mn}$ let $M(\mathbf{x})$ be the $m \times n$ 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 $f : [-1,1]^{mn} \to \mathbb{R}$ defined by

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

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

To bound $D^2 f(\mathbf{x})$ first let $\mathbf{x} \in [-1,1]^{mn}$ 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

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

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

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

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 *For $t > 0$.*

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

and

$$\Pr \{ E[f] - f \geq t \} \leq \exp\left(\frac{-t^2}{8 + 4t/3} \right) .$$

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.

5 Beyond uniform bounds

All 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 (TB1) is then invoked to conclude that

$$\ln Ee^{\beta(f-Ef)} \leq \beta \int_0^\beta \frac{\xi(\gamma)}{\gamma^2} E_{\gamma f}[G(f)]\, d\gamma \leq \beta \sup_{\mathbf{x}} G(f)(\mathbf{x}) \int_0^\beta \frac{\xi(\gamma)\, d\gamma}{\gamma^2}. \quad (13)$$

An analogous strategy was employed for the lower tail bound Theorem 29.

The uniform estimate $E_{\gamma f}[G(f)] \leq \sup_{\mathbf{x}} G(f)(\mathbf{x})$ in (13), 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 $G(f)$ can be bounded in terms of the function f itself, a property referred to as *self-boundedness [4]*. For example, if simply $G(f) \leq f$ then $E_{\gamma f}[G(f)] \leq E_{\gamma f}[f] = (d/d\gamma) \ln Z_{\gamma f}$, and if the function ξ has some reasonable behavior, then the first integral in (13) 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 $f \in A$ there are nonnegative numbers a, b such that*

(i) $V_+^2 f \leq af + b$. Then for $0 \leq \beta < 2/a$

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

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

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

Proof. (i) We use Lemma 11 and get

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

where the last identity follows from the fact that $E_{\gamma f}[f] = (d/d\gamma)\ln Z_{\gamma f}$. 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) For $\beta < 0$ we use Lemma 12

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

Rearranging gives

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

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

To convert part (i) into a tailbound we need an optimization lemma.

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

$$\inf_{\beta\in[0,1/b)}\left(-\beta t + \frac{C\beta^2}{1-b\beta}\right) \leq \frac{-t^2}{2(2C+bt)}. \tag{14}$$

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

$$h(t) \geq \frac{t^2}{2(1+t)}. \tag{15}$$

Substituting

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

in the left side of (14) we obtain

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

where we have used (15). ■

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*

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

(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

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

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

Boucheron et al [4] have given a refined version of the lower tail bound, where the condition $a \geq 1$ is improved to $a \geq 1/3$ for the lower tail. There they also show that Theorem 34 and Theorem 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 exponent.

5.2 Convex Lipschitz functions revisited

In Section 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 [5] where the lower bound in Theorem 34 is applied to the square of the Lipschitz function f. We need the additional assumption that f^2 takes values in an interval of length at most one.

Theorem 35 *Let $\Omega_k = [0,1]$ and let $f \in \mathcal{A}$ be C^1, 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 $\mathbf{x} \in [0,1]^n$. For any k we have $f^2(\mathbf{x}) - \inf^2 f_k(\mathbf{x}) \leq 1$, and using separate convexity as in the proof of Theorem 30, $f(\mathbf{x}) - \inf f_k(\mathbf{x}) \leq |(\partial/\partial x_k) f(\mathbf{x})|$, so

$$\begin{aligned} D^2 f^2(\mathbf{x}) &= \sum_k \left(f^2(\mathbf{x}) - \inf_k f^2(\mathbf{x}) \right)^2 \\ &= \sum_k \left(f(\mathbf{x}) - \inf_k f(\mathbf{x}) \right)^2 \left(f(\mathbf{x}) + \inf_k f(\mathbf{x}) \right)^2 \\ &\leq \sum_k \left(\frac{\partial}{\partial x_k} f(\mathbf{x}) \right)^2 \left(f(\mathbf{x}) + \inf_k f(\mathbf{x}) \right)^2 \\ &\leq 4 \sum_k \left(\frac{\partial}{\partial x_k} f(\mathbf{x}) \right)^2 f(\mathbf{x})^2 \\ &\leq 4L^2 f(\mathbf{x})^2 \end{aligned}$$

By the lower tail bound of Theorem 34 we get a lower tail bound for f^2

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

Then

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

Here we used $E[f] \leq \sqrt{E[f^2]}$ and $\sqrt{E[f^2]} f \geq f^2$ 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 $\beta \in \mathbb{R}$

$$E_{\beta f}[g] \leq \mathrm{Ent}_f(\beta) + E[\ln e^g]. \tag{16}$$

In the discussion at the beginning of Section 5 where we had a general bound of the form $\mathrm{Ent}_f(\beta) \leq \xi(\beta) E_{\beta f}[G(f)]$. Using (16) we can now obtain for any $\theta > 0$

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

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

$$\mathrm{Ent}_f(\beta) \leq \frac{\xi(\beta)}{\theta - \xi(\beta)} \ln E[\exp(\theta G(f))]. \tag{17}$$

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 $\mu_{\theta G(f)}$.

5.4 The supremum of an empirical process

We will apply this trick, which has been proposed in [3], to the upwards tail of the supremum of an empirical process, sharpening the bound obtained in Section 3.4.

Theorem 36 *Let $X_1, ..., X_n$ be independent with values in $\mathcal{X}$ with X_i distributed as μ_i, and let $\mathcal{F}$ be a finite 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*

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

Then for $t > 0$

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

This result is easy and somewhat improves over Theorem 12.2 in [5], since by the triangle inequality $E[W] \leq \Sigma^2 + \sigma^2$ and the constants in the denominator of the exponent are better by a factor of two, and optimal for the variance term.

Proof. Let $0 < \gamma \leq \beta < 2$. Using Theorem 26 and (16) we get

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

Rearranging gives

$$\mathrm{Ent}_F(\gamma) \leq \frac{\gamma}{2-\gamma} \ln E e^{\gamma V^+(F)}. \tag{18}$$

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 $\left(F - S_y^k F\right)_+ \leq \left(\hat{f}(x_i) - \hat{f}(y)\right)_+$ and by the zero mean assumption

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

So $V_+(F)\leq W$. It follows from (18) that

$$\mathrm{Ent}_F(\gamma)\leq\frac{\gamma}{2-\gamma}\ln Ee^{\gamma V^+(F)}\leq\frac{\gamma}{2-\gamma}\ln E\left[e^{\gamma W}\right]. \tag{19}$$

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

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

It therefore follows from Lemma 32 above, that

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

Combining this with (19) 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 (TP1) 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

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

■

References

[1] S.Bernstein, Theory of Probability, Moscow, 1927.

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

[3] S.Boucheron,G.Lugosi,P.Massart, Concentration Inequalities using the entropy method, *Annals of Probability* 31, Nr 3, 2003

[4] S.Boucheron, G.Lugosi, P.Massart, On concentration of self-bounding functions, *Electronic Journal of Probability* Vol.14 (2009), Paper no. 64, 1884–1899, 2009

[5] S. Boucheron, G. Lugosi, P. Massart. Concentration Inequalities, Oxford University Press (2013)

[6] Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. Journal of Machine Learning Research, 2(Mar), 499-526.

[7] P.Chebychev, Sur les valeurs limites des intégrales, *Journal de Mathématiques Pures et Appliquées*, Ser. 2, 19: 157–160, 1874.

[8] Efron, B., & Stein, C. (1981). The jackknife estimate of variance. The Annals of Statistics, 586-596.

[9] J.W.Gibbs. *Elementary Principles in Statistical Mechanics with Especial Reference to the Rational Foundation of Thermodynamics.* Yale University Press, Yale, C.T. 1902.

[10] M.Ledoux, *The Concentration of Measure Phenomenon,* AMS Surveys and Monographs 89, 2001.

[11] E.H.Lieb, Some convexity and subadditivity properties of entropy, *Bull. Amer.Math. Soc.* 81, 1–14, 1975

[12] D.McAllester, L.Ortiz, Concentration inequalities for the missing mass and for histogram rule error, *NIPS*, 2002.

[13] C.McDiarmid, Concentration, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195–248. Springer, Berlin, 1998.

[14] A.Maurer. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms* 29: 121–138, 2006

[15] A.Maurer, Thermodynamics and concentration. *Bernoulli* 18.2 (2012): 434-454.

[16] J.M.Steele, An Efron-Stein inequality for nonsymmetric statistics, *Annals of Statistics* 14:753–758, 1986

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
$E\left[g\right] = \int_{\Omega} g d\mu$	expectation of g in μ
$\sigma^2\left[g\right] = E\left[\left(g - E\left[g\right]\right)^2\right]$	variance of g in μ

Notation for the entropy method

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

Operators on $\mathcal{A}$

$D_{y,y'}^k g = S_y^k g - S_{y'}^k g$	difference operator
$r_k\left(g\right) = \sup_{y,y' \in \Omega_k} D_{y,y'}^k f$	conditional range operator
$R^2\left(g\right) = \sum_k r_k^2\left(g\right)$	sum of conditional square ranges
$\Sigma^2\left(g\right) = \sum_k \sigma_k^2\left[g\right]$	sum of conditional variances
$\left(\inf_k g\right)\left(\mathbf{x}\right) = \inf_{y \in \Omega_k} S_y^k g\left(\mathbf{x}\right)$	conditional infimum operator
$V_+^2 g = \sum_k E_{y \sim \mu_k}\left[\left(\left(g - S_y^k\right)_+\right)^2\right]$	Efron-Stein variance proxy
$D^2 g = \sum_k \left(g - \inf_k g\right)^2.$	worst case variance proxy