Proceedings of Machine Learning Research vol 75:1–24, 2018 31st Annual Conference on Learning Theory

Empirical bounds for functions with weak interactions

Andreas Maurer AM@ANDREAS-MAURER.EU
Adalbertstrasse 55
D-80799 Mnchen, Germany

Massimiliano Pontil MASSIMILIANO.PONTIL@IIT.IT
Istituto Italiano di Tecnologia, 16163 Genoa, Italy
and
University College London, London WC1E 6BT, UK

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We provide sharp empirical estimates of expectation, variance and normal approximation for a class of statistics whose variation in any argument does not change too much when another argument is modified. Examples of such weak interactions are furnished by U- and V-statistics, Lipschitz L-statistics and various error functionals of ℓ_2-regularized algorithms and Gibbs algorithms.

Keywords: List of keywords

1. Introduction

A central problem of learning is to relate a finite number of observations to some underlying law. If the law is not deterministic, the appropriate model is a sequence of random variables X_i taking values in some space $\mathcal{X}$. Under the idealizing assumption of noninterfering observations of identically prepared systems, we assume these variables to be independent and identically distributed according to some probability measure μ on $\mathcal{X}$.

Any quantitative model of the law based on the observations $\mathbf{X} = (X_1, ..., X_n)$ involves the computation of functions $f : \mathcal{X}^n \to \mathbb{R}$. For example $f(\mathbf{x})$ could be a bit computed by a machine-learning program based on the training sample $\mathbf{x}$, or a statistic to estimate some parameter like a moment, quantile or correlation underlying the observed phenomenon. Here we will only consider bounded real valued functions f.

What can we say about the expectation $E[f]$ of $f(\mathbf{X})$? What about its variance, and how can we describe the distribution of $f(\mathbf{X})$?

Without any assumptions on μ, the answer depends on the class of functions under consideration. Many well known and satisfactory answers exist for the sample mean $f : [0,1]^n \to \mathbb{R}$ given by

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

The Chernov and Hoeffding inequalities (McDiarmid, 1998; Boucheron et al., 2013) give high-probability estimates of $E[f]$. Bernstein's inequality is often stronger, but contains the variance as a parameter of the distribution, which requires a separate estimate. Another highlight is the Berry-Esseen theorem (Berry, 1941) giving rates for the approximation of $f(\mathbf{X})$ by an appropriately scaled

normal variable, but again both expressions for the limiting distribution and for the approximation error contain the variance.

The variance of X_i can be estimated by the sample variance $v_n : [0,1]^n \to \mathbb{R}$

$$v_n(\mathbf{x}) = \frac{1}{2n(n-1)} \sum_{i,j \in \{1,\ldots,n\}: i \neq j} (x_i - x_j)^2, \tag{2}$$

and it is shown by Maurer and Pontil (2009) (see also Audibert et al., 2009) that, for any $\delta > 0$, with probability at least $1-\delta$ we have $\left|\sigma(X_i) - \sqrt{v_n(\mathbf{X})}\right| \leq \sqrt{\frac{2}{n-1}\ln(2/\delta)}$. This estimate can be combined with Bernstein's inequality to give a purely empirical estimate of expectation, an empirical Bernstein bound, which is superior to Hoeffding's inequality for functions of small variance (Audibert et al., 2009; Maurer and Pontil, 2009). Similarly the variance estimate can also be used in results about normal approximation.

In this paper we extend these results to general, not necessarily additive functions of independent random variables. Clearly the same quantitative results cannot be expected in great generality, but there is a class of functions whose statistical properties are in many ways very similar to those of the sample mean, even though some of these functions may look highly nonlinear at first glance.

To describe this class we introduce some notation which will be used throughout. For $k \in \{1,...,n\}$ and $y, y' \in \mathcal{X}$ we define the partial difference operator $D^k_{y,y'}$ acting on bounded functions $f : \mathcal{X}^n \to \mathbb{R}$ by

$$D^k_{y,y'} f(\mathbf{x}) = f(x_1,...,x_{k-1},y,x_{k+1},...,x_n) - f\left(x_1,...,x_{k-1},y',x_{k+1},...,x_n\right).$$

Note that $D^k_{y,y'} f(\mathbf{x})$ depends on y and y', but not on x_k.

Definition 1 *For $f : \mathcal{X}^n \to \mathbb{R}$ we define the seminorms*

$$\begin{aligned} M(f) &= \max_{k \in \{1,\ldots,n\}} \sup_{x \in \mathcal{X}^n, y, y' \in \mathcal{X}} D^k_{y,y'} f(x) \\ J(f) &= n \max_{l,k: l \neq k} \sup_{x \in \mathcal{X}^n, z, z', y, y' \in \mathcal{X}} D^l_{z,z'} D^k_{y,y'} f(x). \end{aligned}$$

For $a, b > 0$ we say that a function $f : \mathcal{X}^n \to \mathbb{R}$ has (a,b)-weak interactions, if $M(f) \leq a/n$ and $J(f) \leq b/n$.

A sequence $(f_n)_{n \geq 2}$ of functions $f_n : \mathcal{X}^n \to \mathbb{R}$ has (a,b)-weak interactions if every f_n has (a,b)-weak interactions.

The seminorm M vanishes on constant functions, the seminorm J vanishes on additive functions. They can be interpreted as distribution-independent distance measures to the linear subspaces of constant and additive functions respectively. Notice the factor n in the definition of J, so $D^l_{z,z'} D^k_{y,y'} f(x) \leq J(f)/n$.

M appears in the well known concentration inequality (McDiarmid, 1998; Boucheron et al., 2013)

$$\Pr\{f(\mathbf{X}) - E[f] > t\} \leq \exp\left(\frac{-2t^2}{nM(f)^2}\right), \tag{3}$$

often called Bounded-Difference- or McDiarmid's inequality. This inequality generalizes Hoeffding's inequality to general non-additive functions. Both seminorms M and J appear in the recent inequality (Maurer, 2017)

$$\Pr\left\{f\left(\mathbf{X}\right)-E\left[f\right]>t\right\}\leq\exp\left(\frac{-2t^{2}}{2\sigma^{2}\left(f\right)+J\left(f\right)^{2}/2+\left(2M\left(f\right)/3+J\left(f\right)\right)t}\right), \tag{4}$$

which generalizes Bernstein's inequality to non-additive functions.

In this work we give an estimator v_f for the variance $\sigma^2(f)$, also in terms of M and J (Theorem 2 below), which can be combined with inequality (4) to a purely empirical bound, so as to improve over McDiarmid's inequality for functions of small variance, just as the empirical Bernstein bound for additive functions mentioned above. We also give a result for normal approximation of general non-additive functions, also in terms of M and J (Theorem 4), which can be converted to an empirical result using our variance estimate.

If M and J cannot be appropriately controlled these results are useless. But if a sequence of functions (f_n) has weak interactions, in the sense of above definition, then $M(f_n)$ and $J(f_n)$ have linear or sublinear decay, and statistical properties resemble that of the sample mean. This is intuitively understandable, because (f_n) approaches additivity (the mixed second partial differences go to zero), n times as fast as it becomes a constant (the first partial differences go to zero). Section 2 contains our statistical results for general functions and their specialization to functions with weak interactions.

The class of functions with weak interactions contains U- and M-statistics of any order and Lipschitz L-statistics. It also contains some more exotic specimen, as error functionals for ℓ_2-regularization or the KL-divergence between the Gibbs-measures of true and empirical error for Gibbs algorithms. Section 3 describes examples of weak interactions, all of which obey the results given in Section 2. An appendix contains proofs, other technical material, and a glossary of notation in tabular form.

2. Bounds for functions with weak interactions

In this section, we give some statistical properties of the random variable $f(\mathbf{X})$ and specialize them to functions with weak interactions (a, b), so as to make them directly applicable to the examples in Section 3.

2.1. Notation, the Efron-Stein and Bernstein inequalities

In the sequel, $\mathcal{X}$ will be a measurable space and $(\mu_k)_{k\geq 1}$ a sequence of probability measures on $\mathcal{X}$. The random variables distributed as μ_k are independent and denoted X_k or $X_k \sim \mu_k$ or $(X_1, ..., X_n) \sim \prod_1^n \mu_k$. They are not necessarily identically distributed ($\mu_k = \mu$) unless explicitely mentioned. With $\mathbf{x}$ we denote a vector of the form $(x_1, ..., x_n) \in \mathcal{X}^n$ and with $\mathbf{X}$ a random vector of the form $(X_1, ..., X_n) \sim \prod_{k=1}^n \mu_k$. The algebra of bounded measurable functions $g : \mathcal{X}^n \to \mathbb{R}$ will be denoted by $\mathcal{A}_n$. If $g \in \mathcal{A}_n$ and if $\mathbf{x}$ has at least n components, then $g(\mathbf{x})$ is the function value $g(x_1, ..., x_n)$, and if $\mathbf{X}$ has at least n components then $g(\mathbf{X})$ is the random variable $g(X_1, ..., X_n)$. For $g \in \mathcal{A}_n$ expectation and variance of $g(\mathbf{X})$ will be abbreviated by $E[g]$ and $\sigma^2(g)$. A function $g \in \mathcal{A}_n$ is called additive if $f(\mathbf{x}) = \sum_{i=1}^n h_i(x_i)$ for some real valued $h_i : \mathcal{X} \to \mathbb{R}$.

For $f \in \mathcal{A}_n$ the k-th conditional variance $\sigma_k^2(f)$ and the sum of conditional variances $\Sigma^2(f)$ are the members of $\mathcal{A}_n$ defined by

$$\begin{aligned}\sigma_k^2(f)(\mathbf{x}) &= \frac{1}{2}E_{(Y,Y')\sim\mu_k\times\mu_k}\left[\left(D_{Y,Y'}^k f(\mathbf{x})\right)^2\right] \\ \Sigma^2(f)(\mathbf{x}) &= \sum_{k=1}^n \sigma_k^2(f)(\mathbf{x}).\end{aligned}$$

Note that $\sigma_k^2(f)$ does not depend on x_k, that $\sigma_k^2(f)(\mathbf{x}) \leq M(f)^2/4$ (because the variance of a bounded random variable is always bounded by a quarter of the square of its range) and that $\Sigma^2(f)(\mathbf{x}) \leq nM(f)^2/4$. For additive functions $\Sigma^2(f)(\mathbf{x})$ is independent of $\mathbf{x}$ and equals $\sigma^2(f)$. For non-additive functions this does not hold any more, instead one has the Efron-Stein inequality (Efron and Stein, 1981; Steele, 1986)

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

which gives the general bound $\sigma^2(f) \leq nM(f)^2/4$ on the variance. For functions with $M(f) \leq a/n$ (in particular for weak interactions) we get

$$\sigma^2(f) \leq \frac{a^2}{4n}. \tag{6}$$

The Efron-Stein inequality is very sharp for functions with weak interactions. We have

$$\begin{aligned}E\left[\Sigma^2(f)\right] &\leq \sigma^2(f) + \frac{1}{4}\sum_{k,l:k\neq l} E_{\mathbf{X},Z,Z',Y,Y'}\left[\left(D_{ZZ'}^l D_{YY'}^k f(\mathbf{X})\right)^2\right] \\ &\leq \sigma^2(f) + \frac{J(f)^2}{4}.\end{aligned} \tag{7}$$

The first inequality is due to Houdre (1997) (see also Maurer, 2017), the second is an elementary estimate. For weak interactions we get

$$E\left[\Sigma^2(f)\right] - \frac{b^2}{4n^2} \leq \sigma^2(f) \leq E\left[\Sigma^2(f)\right]. \tag{8}$$

In Maurer (2017) the following Bernstein-type inequality is shown to hold for every f in $\mathcal{A}_n$ and $\delta > 0$

$$\Pr\left\{f - E[f] > \sqrt{2E\left[\Sigma^2(f)\right]\ln(1/\delta)} + (2M(f)/3 + J(f))\ln(1/\delta)\right\} < \delta. \tag{9}$$

Using (8) and some elementary estimates for functions with (a,b)-weak interactions we obtain for $\delta \leq 1/e$

$$\Pr\left\{f - E[f] > \sqrt{2\sigma^2(f)\ln(1/\delta)} + (2a/3 + 3b/2)\frac{\ln(1/\delta)}{n}\right\} < \delta.$$

Since $\sigma(f)$ decays at least as quickly as $a/\sqrt{4n}$ because of (6), this achieves, for large n, at least the rate of McDiarmid's inequality (3), but it is potentially much better if $\sigma(f)$ is very small. This motivates the search for efficient estimators of $\sigma(f)$.

2.2. Variance estimation

We show that for $f \in \mathcal{A}_n$ having (a,b)-weak interactions $\sigma(f)$ can be estimated with high probability up to order $1/n$ by an estimator using only $n+1$ observations. This is one of the main results of this work.

For any $m \in \mathbb{N}$, $m > 1$, $1 \leq k \leq m$, $\mathbf{x} \in \mathcal{X}^m$ and $y \in \mathcal{X}$ define the replacement operator $S_y^k : \mathcal{X}^m \to \mathcal{X}^m$ and the deletion operator $S_-^k : \mathcal{X}^m \to \mathcal{X}^{m-1}$ by

$$\begin{aligned} S_y^k \mathbf{x} &= (x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n) \in \mathcal{X}^m \\ \text{and } S_-^k \mathbf{x} &= (x_1, ..., x_{k-1}, x_{k+1}, ..., x_n) \in \mathcal{X}^{m-1}. \end{aligned}$$

Our variance estimator is the function $v_f \in \mathcal{A}_{n+1}$ given by

$$v_f(\mathbf{x}) = \frac{1}{2(n+1)} \sum_{j=1}^{n+1} \sum_{i:i \neq j} \left(f\left(S_-^j \mathbf{x}\right) - f\left(S_-^j S_{x_j}^i \mathbf{x}\right) \right)^2. \tag{10}$$

$S_-^j \mathbf{x}$ has the j-th component deleted and $S_-^j S_{x_j}^i \mathbf{x}$ has the i-th component replaced by the j-th component and then the j-th component deleted. So both vectors differ only in one component, which is x_i in $S_-^j \mathbf{x}$ and x_j in $S_-^j S_{x_j}^i \mathbf{x}$. Also both vectors do not contain any repeated components.

It is obvious how the estimator is to be implemented in a computer program. Computation requires $(n+1)^2$ computations of f, but only a sample of size $n+1$. The latter may be a great advantage, because computing may be cheap, while collecting a sample can be very expensive (think of surveys or the results of histological examinations in medical applications). We first give the result in terms of the seminorms.

Theorem 2 *Let $\delta \in (0,1)$. If $f \in \mathcal{A}_n$ and the X_i are identically distributed, then with probability at least $1-\delta$ in $\mathbf{X} = (X_1, ..., X_{n+1})$*

$$\left| \sqrt{E\left[\Sigma^2(f)\right]} - \sqrt{v_f(\mathbf{X})} \right| \leq \sqrt{\left(2M(f)^2 + 8J(f)^2\right) \ln(2/\delta)}.$$

For one-sided bounds $2/\delta$ can be replaced by $1/\delta$.

The proof is given in appendix A. It first establishes that v_f is an unbiased estimator for $E\left[\Sigma^2(f)\right]$ and then uses a concentration inequality for self-bounded functions.

The result requires identical distribution of the X_i, in contrast to the Efron-Stein and Bernstein inequalities, but it does not require f to be symmetric. It is important to observe that our estimator requires one additional observation, as the the variance of $f(X_1, ..., X_n)$ is estimated by $v_f(X_1, ..., X_n, X_{n+1})$.

Because of (7) and (5) we have $E\left[\Sigma^2(f)\right] - J(f)^2/4 \leq \sigma^2(f) \leq E\left[\Sigma^2(f)\right]$. Thus

$$\Pr\left\{ \sqrt{v_f(\mathbf{X})} - J(f)/2 - \sqrt{\left(2M(f)^2 + 8J(f)^2\right) \ln(2/\delta)} < \sigma(f) \right\} < \delta/2,$$

which together with Theorem 2 immediately gives the following corollary (using $\delta < 1 \implies 1/2 \leq \sqrt{\ln(2/\delta)}$).

Corollary 3 *Let $\delta \in (0,1)$. If $f \in \mathcal{A}_n$ has (a,b)-weak interactions and the X_i are identically distributed, then with probability at least $1-\delta$ in $\mathbf{X} = (X_1, ..., X_{n+1})$*

$$\sqrt{v_f(\mathbf{X})} - \frac{K_-(a,b)}{n}\sqrt{\ln(2/\delta)} \leq \sigma(f) \leq \sqrt{v_f(\mathbf{X})} + \frac{K_+(a,b)}{n}\sqrt{\ln(2/\delta)},$$

where

$$\begin{aligned} K_-(a,b) &= b/2 + \sqrt{2a^2 + 8b^2} \\ K_+(a,b) &= \sqrt{2a^2 + 8b^2}. \end{aligned}$$

For one-sided bounds $2/\delta$ can be replaced by $1/\delta$.

The bounds on the variance are of order $1/n$. Since the Efron-Stein inequality implies only $\sigma(f) \leq a/\sqrt{4n}$, there is a significant estimation benefit for larger values of n.

If f is the sample mean (1), then

$$f\left(S_-^j \mathbf{x}\right) - f\left(S_-^j S_{x_j}^i \mathbf{x}\right) = \frac{1}{n}\begin{cases} x_i - x_j & \text{if} \quad i < j \\ x_{i-1} - x_{j-1} & \text{if} \quad j < i \end{cases},$$

so substitution in (10) shows that the estimator $v_f = (1/n)\, v_{n+1}$, where v_{n+1} is the sample variance (2). Since $b = 0$ for the sample mean we get the bound

$$\left|\sqrt{v_f(\mathbf{X})} - \sigma(f)\right| \leq \frac{1}{n}\sqrt{2\ln(2/\delta)},$$

so for the sample mean Corollary 3 gives the same rate as (Maurer and Pontil, 2009).

2.3. Normal approximation

Modulo a lower bound on the variance, we give a finite sample bound on normal approximation for functions with weak interactions. To formulate the result we use the following distance to normality of a real random variable W.

$$d_{\mathcal{N}}(W) = \sup\left\{\left|E\left[h\left(\frac{W - E[W]}{\sigma(W)}\right)\right] - E[h(Z)]\right| : h \text{ a real Lipschitz-1 function}\right\},$$

where $Z \sim \mathcal{N}(0,1)$. Thus $d_{\mathcal{N}}(W)$, which has also been used in Chatterjee (2008), is the Wasserstein distance between a standardized clone of W and a standard normal variable. We then have the following general result.

Theorem 4 *For $f \in \mathcal{A}_n$ let W be the random variable $W = f(\mathbf{X})$. Then*

$$d_{\mathcal{N}}(W) \leq \frac{\sqrt{n} M(f)\left(J(f) + M(f)\right)}{\sigma^2(f)} + \frac{n M(f)^3}{2\sigma^3(f)}.$$

The proof is given in appendix B. It relies on an inequality of Chatterjee (2008), which uses a variant of Stein's method (Chen et al., 2010) for normal approximation. To apply the result we need a lower bound on the variance. In the next section we use an empirical estimate, but here we simply assume a bound of the form $\sigma(f) \geq Cn^{-p}$ for some constants C and p. By (6) we must have $p \geq 1/2$. Specializing to weak interactions we obtain with some algebra

Corollary 5 *For $f \in \mathcal{A}_n$ let W be the random variable $W = f(\mathbf{X})$. If f has (a,b)-weak interactions and $\sigma(f) \geq Cn^{-p}$ then*

$$d_{\mathcal{N}}(W) \leq \frac{Ca(a+b)+a^3}{C^3 n^{2-3p}}.$$

So if a sequence f_n has (a,b)-weak interactions, $\sigma(f_n) \geq Cn^{-p}$ and $1/2 \leq p < 2/3$, then the sequence $(f_n(\mathbf{X}) - E[f_n])/\sigma(f_n)$ converges to a standard normal variable in the Wasserstein metric. For $p \geq 2/3$ the result says nothing about the asymptotic distribution. In the simplest case $p = 1/2$ (as with non-degenerate U-statistics) the rate of approach to normality is $n^{-1/2}$.

2.4. Empirical bounds for weak interactions

Now we will cast the Bernstein inequality (9) and the normal approximation inequality of the previous section into an empirical form by using the results on variance estimation of Section 2.2. In this case we will need identical distribution of the variables X_i.

To combine Bernstein's inequality (9) and the upper bound on the variance of Corollary 3 elementary estimates give

Theorem 6 (Empirical Bernstein Inequality) *If $f \in \mathcal{A}_n$ has (a,b)-weak interactions and the X_i are iid, then for $\delta > 0$ with probability at least $1-\delta$*

$$f(\mathbf{X}) \leq E[f] + \sqrt{2v_f(\mathbf{X})\ln(2/\delta)} + \frac{(8a/3+5b)\ln(2/\delta)}{n}.$$

While for Bernstein's inequality we want the variance to be small, for our normal approximation result, Theorem 4, we want it to be big. The situation is also more complicated, because the variance now appears in the denominator of the bound, so the estimate may fail. In fact it may even fail for all members of a sequence, because asymptotic normality needn't hold. We therefore precede the empirical bound by a test to verify its applicability.

Theorem 7 *Suppose that $f \in \mathcal{A}_n$ has (a,b)-weak interactions and the X_i are iid. Let W be the random variable $W = f(\mathbf{X})$. For $\delta > 0$ let $A(\delta)$ and B be the events*

$$\begin{aligned} A(\delta) &= \left\{ \frac{\sqrt{v_f(\mathbf{X})}}{2} \geq \frac{K_-(a,b)\sqrt{\ln(1/\delta)}}{n} \right\}, \\ B &= \left\{ d_{\mathcal{N}}(W) \leq \frac{4(a^2+ab)}{v_f(\mathbf{X})n^{3/2}} + \frac{4a^3}{v_f(\mathbf{X})^{3/2}n^2} \right\}. \end{aligned}$$

Then $\Pr(A(\delta) \implies B) \geq 1-\delta$.

The conclusion can also be read as $\Pr\{A(\delta) \text{ and not } B\} < \delta$ or $\Pr\{B \text{ or not } A(\delta)\} \geq 1-\delta$ or $\Pr\{B|A(\delta)\}\Pr A(\delta) \geq \Pr A(\delta) - \delta$.

Proof Let $C(\delta)$ be the event

$$C(\delta) = \left\{ \sqrt{v_f(\mathbf{X})} - \frac{K_-(a,b)\sqrt{\ln(1/\delta)}}{n} \leq \sigma(f) \right\}.$$

Then by Corollary 3 $\Pr C(\delta) \geq 1-\delta$. But under $C(\delta)$ the event A implies

$$\frac{\sqrt{v_f(\mathbf{X})}}{2} \leq \sqrt{v_f(\mathbf{X})} - \frac{K_-(a,b)\sqrt{\ln(1/\delta)}}{n} \leq \sigma(f)$$

which implies B by Theorem 4 and (a,b)-weak interactions of f. ■

On a sequence of functions f_n this result could be put to work as follows. First fix δ and n and observe $\mathbf{X}_1^{n+1}$. Then compute the variance estimator and check if $A(\delta)$ holds. If it doesn't hold then n may be to small and we may try a larger n. If we don't get it too work then the variances decay too fast and f_n may simply not be asymptotically normal, so we give up. If $A(\delta)$ holds on the other hand, we have an empirical bound on normal approximation, which can tell us a lot about the distribution of $f(\mathbf{X})$.

In the regime where Corollary 5 guarantees asymptotic normality, that is $\sigma(f_n) \geq Cn^{-p}$ and $1/2 \leq p < 2/3$, Corollary 3 guarantees that the test $A(\delta)$ succeeds with high probability for sufficiently large n.

3. Examples of functions with weak interactions

We give examples of functions having weak interactions and identify the parameters (a,b), so as to make the results of the previous section applicable. Some obvious closure relations for functions with weak interactions follow from the fact that M and J are seminorms. If f_1 and f_2 have (a_1,b_1)- and (a_2,b_2)-weak interactions respectively and $c \in \mathbb{R}$, then f_1+f_2 has (a_1+a_2, b_1+b_2)-weak interactions, f_1+c has (a_1,b_1)-weak interactions and cf_1 has $(|c|a,|c|b)$-weak interactions. The last fact allows to rescale the conveniently scaled examples we choose below.

3.1. The sample mean, V- and U-statistics

Let $\mathcal{X}=[0,1]$. The sample mean

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

has seminorm values $M(f)=1/n$ and $J(f)=0$, and therefore $(1,0)$-weak interactions. $f(\mathbf{X})$ is an unbiased estimator of the expectation of a $[0,1]$-valued random variable.

V- and U-statistics are generalizations of the sample mean. Fix $1 \leq m < n$, and for any multi-index $\mathbf{j}=(j_1,...,j_m) \in \{1,...,n\}^m$ let $\kappa_{\mathbf{j}} : \mathcal{X}^m \to [-1,1]$ and define $V,U : \mathcal{X}^m \to \mathbb{R}$,

$$\begin{aligned} V(\mathbf{x}) &= n^{-m} \sum_{\mathbf{j}\in\{1,...,n\}^m} \kappa_{\mathbf{j}}(x_{j_1},...,x_{j_m}) \\ U(\mathbf{x}) &= \binom{n}{m}^{-1} \sum_{1\leq j_1<...<j_m\leq n} \kappa_{\mathbf{j}}(x_{j_1},...,x_{j_m}). \end{aligned}$$

V-statistics have their name from Richard von Mises, who studied their asymptotic distributions (Von Mises, 1947). $V(\mathbf{x})$ receives contributions from multi-indices with multiple occurrences of individual indices. But in the expression for $D^k_{y,y'}V(\mathbf{x})$ only those multi-indices $\mathbf{j}$ survive, which contain k, with the corresponding contribution being at most $2n^{-m}$. There is a first position where

k appears in $\mathbf{j}$, for which there are m possibilities, and the remaining indices j_i can assume all values in $\{1,...,n\}$. It follows that there are at most mn^{m-1} surviving multi-indices with maximal contribution $2n^{-m}$, whence

$$M(V) = \max_k \sup_{\mathbf{x},y,y'} D^k_{y,y'} V(\mathbf{x}) \leq \frac{2mn^{m-1}}{n^m} = \frac{2m}{n}.$$

For $D^l_{z,z'} D^k_{y,y'} V(\mathbf{x})$ with $k \neq l$ each contributing index must contain both k and l. For the positions of k and l there are $m(m-1)$ possibilities. The remaining $m-2$ indices being arbitrary, there is a total of at most $m(m-1)n^{m-2}$ contributing indices, each making a contribution of at most $4n^{-m}$. Therefore $D^l_{z,z'} D^k_{y,y'} V(\mathbf{x}) \leq 4m(m-1)/n^2$ and

$$J(V) = n \max_{k \neq l} \sup_{\mathbf{x},z,z',y,y'} D^l_{z,z'} D^k_{y,y'} V(\mathbf{x}) \leq 4m(m-1)/n.$$

We conclude that V has $(2m, 4m(m-1))$-weak interactions.

U-statistics avoid multi-indices with multiple occurrences of indices. If all the $\kappa_{\mathbf{j}}$ are equal to some permutation symmetric function κ, and the X_i are iid, then $U(\mathbf{X})$ is an unbiased estimator for $E(X_1,...,X_m)$, which accounts for their name (Hoeffding, 1948). U-statistics are relevant to metric learning (Cao et al., 2016) and ranking (Clemencon et al., 2008). Similar to V-statistics it is not difficult to show that U has $\left(2m, 4m^2\right)$-weak interactions (see Maurer, 2017).

U-statistics have been extensively studied. There are normal approximation results for nondegenerate U-statistics in Chen et al. (2010), which use the Kolmogorov distance and seem to slightly improve over what we get from substituting $\left(2m, 4m^2\right)$ in Corollary 5. These results also contain variances, which would make them amenable to variance estimation as in Theorem 7.

Peel and Ralaivola (2010) use the fact that the variance of a U-statistic is itself a U-statistic and use either Hoeffding (1948) or Arcones (1995) versions of Bernstein's inequality for U-statistics to estimate the variance. These bounds are however inferior to the Bernstein inequality (9), because the first does not use the correct variance proxy and the second has a scale proxy which increases exponentially in the order m. The same problem besets the empirical Bernstein bounds given in (Peel and Ralaivola, 2010), which is inferior to the general result we get from Theorem 6 except for the first version of Peel and Ralaivola (2010) in a regime of large m/n and a kernel κ far from degeneracy.

3.2. Lipschitz L-statistics

Let $\mathcal{X} = [0,1]$ and use $\left(x_{(1)},...,x_{(n)}\right)$ to denote the order statistic of $\mathbf{x} \in \mathcal{X}^n$. Let $F : [0,1] \to \mathbb{R}$ have supremum norm $\|F\|_\infty$ and Lipschitz-constant $\|F\|_{Lip}$ and consider the function

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n F(i/n)\, x_{(i)}. \tag{11}$$

In appendix C we show that f has $\left(\|F\|_\infty, \|F\|_{Lip}\right)$-weak interactions.
Such statistics also generalize the sample mean, which is obtained by choosing F identically 1. Appropriate choices of F lead to smoothly trimmed means or smoothened quantiles. For example

with $\zeta \in (0, 1/2)$ the choice

$$F(t) = \begin{cases} 0 & \text{if} \quad t \in [0, \zeta] \\ \frac{t-\zeta}{(1/2-\zeta)^2} & \text{if} \quad t \in [\zeta, 1/2] \\ \frac{1-t-\zeta}{(1/2-\zeta)^2} & \text{if} \quad t \in [1/2, 1-\zeta] \\ 0 & \text{if} \quad t \in [1-\zeta] \end{cases}$$

effects a smoothened median where F has the Lipschitz constant $\|F\|_{Lip} = (1/2 - \zeta)^{-2}$. The case $\zeta = 0$ has the best guaranteed estimation properties, but its expectation is the coarsest substitute of the median. As $\zeta \to 1/2$ estimation deteriorates, but the expectation becomes closer to a median.

Normal approximation results for these statistics in terms of the Kolmogorov distance are also given in Chen et al. (2010), similar to what we obtain by substituting $\left(\|F\|_\infty, \|F\|_{Lip}\right)$ in Corollary 5. We are not aware of any results giving Bernstein-type inequalities or tight variance estimation in this case.

3.3. ℓ_2-regularization

While the previous examples had a certain kinship to the sample mean, the following looks quite different. Let $(H, \langle \cdot, \cdot \rangle, \|\cdot\|)$ be a real Hilbert space with unit ball $\mathbb{B}_1 = \mathcal{X}$ and define $g : \mathcal{X}^n \to H$ by

$$g(\mathbf{x}) = \arg\min_{w \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle x_i, w \rangle) + \lambda \|w\|^2, \tag{12}$$

where the non-negative real loss function ℓ is assumed to be convex, three times differentiable and satisfies $\ell(0) = 1$, and the regularization parameter λ satisfies $0 < \lambda < 1$. Then g is a well-known regularized algorithm which upon thresholding can be used for linear classification.

Define the empirical and the true losses $\hat{L}$ and $L : \mathcal{X}^n \to \mathbb{R}$ by

$$\hat{L}(\mathbf{x}) = \frac{1}{n} \sum_i \ell(\langle x_i, g(\mathbf{x}) \rangle) \text{ and } L(\mathbf{x}) = E_{x \sim \mu}[\ell(\langle x, g(\mathbf{x}) \rangle)],$$

where μ is some probability measure on $\mathbb{B}_1$. Let

$$\Delta(\mathbf{x}) = L(\mathbf{x}) - \hat{L}(\mathbf{x}),$$

which measures how much the true and empirical loss of the algorithm differ. It has been shown in (Proposition 5 Maurer, 2017a), that Δ has $\left(c_1 \lambda^{-3/2}, c_2 \lambda^{-4}\right)$-weak interactions, where the constants c_i depend on the derivatives of the loss-function ℓ. In (Maurer, 2017a) this is used to apply the Bernstein inequality (9) to the random variable $\Delta(\mathbf{X})$. Here we complement this result by simply substituting the weak interaction parameters in Corollary 3 and Corollary 5 so as to obtain bounds to estimate the variance of $\Delta(\mathbf{X})$ and to give bounds on normal approximation.

3.4. A chain rule

We interrupt the presentation of examples, to show how new interesting examples of functions with weak interactions can be generated from given ones, in addition to the obvious closure relations

which follow from M and J being seminorms. First we extend the definitions of M and J to Banach-space valued functions $f : \mathcal{X}^n \to B$ in an obvious way by setting

$$M(f) = \max_k \sup_{\mathbf{x},y,y'} \left\| D^k_{yy'} f(x) \right\| \text{ and } J(f) = n \max_{k \neq l} \sup_{\mathbf{x},y,y',z,z'} \left\| D^l_{zz'} D^k_{yy'} f(x) \right\|,$$

and we say that f has (a,b)-weak interactions if $M(f) \leq a/n$ and $J(f) \leq b/n$. Then we have the following chain rule, whose proof will be given in appendix D.

Lemma 8 *Let B be a Banach space, $U \subseteq B$ convex, $f : \mathcal{X}^n \to U$, and assume that the function $F : U \to \mathbb{R}$ is twice Fréchet-differentiable. Then*

$$\begin{aligned} M(F \circ f) &\leq \sup_{v \in U} \left\| F'(v) \right\| M(f) \text{ and} \\ J(F \circ f) &\leq n \sup_{v \in U} \left\| F''(v) \right\| M(f)^2 + \sup_{v \in U} \left\| F'(v) \right\| J(f), \end{aligned}$$

where $\|F'(v)\|$ and $\|F''(v)\|$ are the norms of the linear respectively bilinear functionals $F'(v)$ and $F''(v)$.

$$\left\| F'(v) \right\| = \sup_{w \in B, \|w\| \leq 1} \left\| F'(v)(w) \right\| \text{ and } \left\| F''(v) \right\| = \sup_{w_1, w_2 \in B, \|w_i\| \leq 1} \left\| F''(v)(w_1, w_2) \right\|.$$

The lemma shows that if f has (a,b)-weak interactions and $\|F''(v)\|$ and $\|F'(v)\|$ are bounded on U, then $F \circ f$ has (a',b')-weak interactions, where

$$a' = a \sup_{v \in U} \left\| F'(v) \right\| \text{ and } b' = a^2 \sup_{v \in U} \left\| F''(v) \right\| + b \sup_{v \in U} \left\| F'(v) \right\|.$$

It also shows our definition of weak interactions with its $1/n$-scaling is the only definition of a class of functions such that M and J are of the same order in n, and the class is invariant under compositions with C^2 functions with bounded derivatives.

3.5. The Gibbs algorithm

We use the chain rule, Lemma 8, to show that several quantities related to the Gibbs algorithm have weak interactions and thus satisfy the conditions for the results in Section 2.

Let Ω be some space of "models" endowed with some positive a-priori measure ρ and suppose that $\ell : (\omega, x) \in \Omega \times \mathcal{X} \mapsto \ell(\omega, x) \in [0,1]$ is the loss of the model ω on the datum $x \in \mathcal{X}$. The function $H : \Omega \times \mathcal{X}^n \to [0,1]$ defined by $H(\omega, \mathbf{x}) = (1/n) \sum_{i=1}^n \ell(\omega, x_i)$ is then just the sample average, or empirical error of ω on $\mathbf{x}$. Let β be some positive constant, or "inverse temperature". The Gibbs algorithm returns the distribution

$$d\pi_{\mathbf{x}}(\omega) = Z^{-1}(\mathbf{x}) e^{-\beta H(\omega, \mathbf{x})} d\rho(\omega) \text{ where } Z(\mathbf{x}) = \int_\Omega e^{-\beta H(\omega, \mathbf{x})} d\rho(\omega).$$

Typically this distribution is the stationary distribution of some sample-controlled stochastic process characterizing the algorithm. The Gibbs algorithm plays a role in the simulation of the equilibrium state in statistical mechanics (Binder, 1997) or in non-convex optimization such as simulated annealing (Kirkpatrick, 1983). There is also some recent attention because $d\pi_{\mathbf{x}}$ can be the limiting distribution of randomized algorithms in the training of deep neural networks (Rakhlin et al., 2017).

To analyze the Gibbs algorithm we define the function

$$f : \mathbf{x} \in \mathcal{X}^n \mapsto H(\cdot, \mathbf{x}) \in L_\infty(\Omega). \tag{13}$$

It is easy to verify that this function, which is just a Banach space-valued sample average, has $(1, 0)$-weak interactions. Its range is contained in the unit ball of $L_\infty(\Omega)$.

Related to the Gibbs algorithm is the free energy

$$\Lambda(\mathbf{x}) = \ln Z(\mathbf{x}) = \ln \int_\Omega e^{-\beta H(\omega, \mathbf{x})} d\rho(\omega),$$

which is interesting, because it generates the sample error averaged under the Gibbs distribution

$$\frac{d}{d\beta}\Lambda(\mathbf{x}) = -\int_\Omega H(\omega, \mathbf{x})\, d\pi_{\mathbf{x}}(\omega).$$

Then $\Lambda(\mathbf{x}) = \Xi \circ f(\mathbf{x})$ where Ξ is defined as

$$\Xi : G(.) \in L_\infty(\Omega) \mapsto \ln \int_\Omega e^{-\beta G(\omega)} d\rho(\omega).$$

It is easy to show that $\|\Xi'(G)\| \leq \beta$ and $\|\Xi''(G)\| \leq 2\beta^2$ (see appendix E). The chain rule Lemma 8 then shows that Λ has $\left(\beta, 2\beta^2\right)$-weak interactions, with corresponding consequences for a Bernstein inequality, normal approximation and estimation of variance for the random free energy $\Lambda(\mathbf{X})$.

Let X be a random variable with values in $\mathcal{X}$. Then the "true" error is given by the function $H_0 : \omega \mapsto E_{\mathbf{X}}[H(\omega, \mathbf{X})]$ and the corresponding Gibbs measure is

$$d\pi(\omega) = Z^{-1} e^{-\beta H_0(\omega)} d\rho(\omega).$$

A question of generalization is how much the measures $d\pi_{\mathbf{x}}$ and $d\pi$ differ. We might measure this difference by the Kullback-Leibler divergence $KL(d\pi_{\mathbf{x}}, d\pi)$ of the two measures. A mechanical computation using the chain rule (see appendix E) shows that the function $\mathbf{x} \mapsto KL(d\pi_{\mathbf{x}}, d\pi)$ has $\left(4\beta^2 + 2\beta, 12\beta^3 + 6\beta^2\right)$-weak interactions, which again gives useful information about the random variable $KL(d\pi_{\mathbf{X}}, d\pi)$.

There is an intuitive parallel to the case of ℓ_2-regularization of Section 3.3. In both cases the weak interaction parameters increase, with a corresponding deterioration of estimation, as we tune more closely to the sample, which for ℓ_2-regularization means decreasing λ and for the Gibbs algorithm increasing β, or lowering the "temperature". This fits with the general paradigm of regularization.

4. Summary and some open questions

We have shown that functions with weak interactions have tractable statistical properties, and that the class of such functions is quite rich, containing a number of well known statistics and other functions relevant to machine learning and statistics.

Our preliminary survey provides a small probabilistic toolbox which could be used in statistical learning theory. Apart from the application to ℓ_2-regularized classification, and the analysis of Gibbs algorithms, are there other applications to supervised learning? What is the benefit of finite-sample bounds for normal approximation? Can the empirical Bernstein bound for non-additive functions be used in the analysis of reinforcement learning algorithms, just as its additive counterpart? On the theoretical side, is there a general large deviation principle for weak interactions, in the spirit of Cramer's theorem?

References

M. A. Arcones. A Bernstein-type inequality for U-statistics and U-processes. *Statistics and Probability Letters*, 22(3):239-247, 1995.

J. Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876-1902, 2009.

A. C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122-136, 1941.

K. Binder. Applications of Monte Carlo methods to statistical physics. *Reports on Progress in Physics*, 60(5):487, 1997.

S. Boucheron, G. Lugosi, P. Massart. *Concentration Inequalities*, Oxford University Press, 2013.

Q. Cao, Z. C. Guo, and Y. Ying. Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115-132, 2016.

S. Chatterjee. A new method of normal approximation. *The Annals of Probability*, 36.4:1584-1610, 2008.

L. H. Chen, L. Goldstein, and Q. M. Shao. *Normal approximation by Stein's method*. Springer Science and Business Media, 2010.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 844-874, 2008.

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 586-596, 1981.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 293-325, 1948.

C. Houdré. The iterated jackknife estimate of variance. *Statistics and Probability Letters*, 35(2):197-201, 1997.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671-680, 1983.

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

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In COLT. 2009.

A. Maurer. A Second-order Look at Stability and Generalization. Conference on Learning Theory. 2017.

A. Maurer. A Bernstein-type inequality for functions of bounded interaction. *Bernoulli* (Forthcoming), (see also arXiv preprint arXiv:1701.06191).

C. McDiarmid. Concentration. In *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195–248. Springer, Berlin, 1998.

T. Peel, S. Anthoine, and L. Ralaivola. Empirical Bernstein inequalities for u-statistics. In Advances in Neural Information Processing Systems, pp. 1903-1911, 2010.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. arXiv preprint arXiv:1702.03849.

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

R. V. Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309-348, 1947.

The appendix contains technical material and a table of notations.

Appendix A. Proof of the variance estimation theorem

Define an operator D^2 on $\mathcal{A}_n$ by

$$D^2 f(\mathbf{x}) = \sum_k \left(f(\mathbf{x}) - \inf_{y \in \mathcal{X}} S_y^k f(\mathbf{x}) \right)^2.$$

The proof of Theorem 2 uses the following concentration inequality which can be found in (Maurer, 2006, Theorem 13) or Boucheron et al. (2013).

Theorem 9 *Suppose $f : \mathcal{X}^n \to \mathbb{R}$ satisfies for some $a > 0$*

$$D^2 f(\mathbf{x}) \leq a f(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}^n, \tag{14}$$

and let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent variables. Then for all $t > 0$

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

If in addition $f(\mathbf{x}) - \inf_{y \in \mathcal{X}} S_y^k f(\mathbf{x}) \leq 1$ for all $k \in \{1, ..., n\}$ and all $\mathbf{x} \in \mathcal{X}^n$ then

$$\Pr\{E[f] - f(\mathbf{X}) > t\} \leq \exp\left(\frac{-t^2}{2\max\{a, 1\} E[f(X)]} \right).$$

Corollary 10 *If $f \in \mathcal{A}_n$ satisfies (14) and for some $b > 0$ $f(\mathbf{x}) - \inf_{y \in \mathcal{X}} S_y^k f(\mathbf{x}) \leq b$ for all $k \in \{1, ..., n\}$ and all $\mathbf{x} \in \mathcal{X}^n$ then for all $\delta > 0$ with probability at least $1 - \delta$*

$$\sqrt{f(\mathbf{X})} - \sqrt{2a \ln(2/\delta)} \leq \sqrt{E[f]} \leq \sqrt{f(\mathbf{X})} + \sqrt{2\max\{a, b\} \ln(2/\delta)}.$$

For a one-sided bound $2/\delta$ can be replaced by $1/\delta$.

Proof If $f(\mathbf{x}) - \inf_{y\in\mathcal{X}} S_y^k f(\mathbf{x}) \leq b$ then $(f(\mathbf{x})/b) - \inf_{y\in\mathcal{X}} S_y^k (f(\mathbf{x})/b) \leq 1$ and (14) implies $D^2(f(\mathbf{x})/b) \leq (a/b)(f(\mathbf{x})/b)$, so by the second conclusion of Theorem 9

$$\begin{aligned}\Pr\{E[f] - f(\mathbf{X}) > t\} &= \Pr\{E[f/b] - f(\mathbf{X})/b > t/b\} \\ &\leq \exp\left(\frac{-(t/b)^2}{2\max\{a/b,1\}E[f/b]}\right) = \exp\left(\frac{-t^2}{2\max\{a,b\}E[f]}\right)\end{aligned}$$

(this is really an alternative formulation of the second conclusion of Theorem 9). Equating the R.H.S. to δ solving for t and elementary algebra then give with probability at least $1-\delta$ that

$$\sqrt{E[f]} \leq \sqrt{f(\mathbf{X})} + \sqrt{2\max\{a,b\}\ln(1/\delta)}.$$

In a similar way the first conclusion of Theorem 9 gives with probability at least $1-\delta$ that

$$\sqrt{f(\mathbf{X})} - \sqrt{2a\ln(1/\delta)} \leq \sqrt{E[f]}.$$

A union bound concludes the proof. ■

Proof of Theorem 2 First we show that v_f is an unbiased estimator for the Efron-Stein upper bound $E\left[\Sigma^2(f)\right]$. Observe that for $1 \leq i < j \leq n+1$

$$E\left[\left(f\left(S_-^j\mathbf{X}\right) - f\left(S_-^j S_{X_j}^i \mathbf{X}\right)\right)^2\right] = 2E\left[\sigma_i^2(f)\right],$$

while for $1 \leq j < i \leq n+1$

$$E\left[\left(f\left(S_-^j\mathbf{X}\right) - f\left(S_-^j S_{X_j}^i \mathbf{X}\right)\right)^2\right] = 2E\left[\sigma_{i-1}^2(f)\right].$$

Thus

$$\begin{aligned}E[v_f] &= \frac{1}{2(n+1)}\sum_{i=1}^{n+1}\sum_{j:j\neq i} E\left[\left(f\left(S_-^j\mathbf{X}\right) - f\left(S_-^j S_{x_j}^i \mathbf{X}\right)\right)^2\right] \\ &= \frac{1}{n+1}\left(\sum_{i=2}^{n+1}\sum_{j=1}^{i-1} E\left[\sigma_{i-1}^2(f)\right] + \sum_{i=1}^{n}\sum_{j=i+1}^{n+1} E\left[\sigma_i^2(f)\right]\right) \\ &= \frac{1}{n+1}\sum_{i=1}^{n}(n+1)E\left[\sigma_i^2(f)\right] \\ &= E\left[\Sigma^2(f)\right].\end{aligned}$$

We then apply Corollary 10 to the function v_f. Fix $\mathbf{x} \in \mathcal{X}^{n+1}$, and for each $k \in \{1,...,n+1\}$ let $y_k := \arg\min_{y\in\mathcal{X}} S_y^k v_f(\mathbf{x})$. For $i,j,k \in \{1,...n+1\}$ let

$$a_{ij} := f\left(S_-^j\mathbf{x}\right) - f\left(S_-^j S_{x_j}^i \mathbf{x}\right) \text{ and } a_{ijk} := f\left(S_-^j S_{y_k}^k \mathbf{x}\right) - f\left(S_-^j S_{x_j}^i S_{y_k}^k \mathbf{x}\right).$$

Then

$$v_f(\mathbf{x}) = \frac{1}{2(n+1)}\sum_i\sum_{j:j\neq i} a_{ij}^2. \tag{15}$$

Observe that $|a_{ij}|, |a_{ijk}| \leq M(f)$ and that $J(f \circ S_-^j) = J(f)$ so for $j \neq i \neq k \neq j$ $|a_{ij} - a_{ijk}| \leq J(f)/n$. Also the replacement of a component, which is then deleted, has no effect, so

$$a_{ikk} = f\left(S_-^k S_{y_k}^k \mathbf{x}\right) - f\left(S_-^k S_{x_k}^i S_{y_k}^k \mathbf{x}\right) = f\left(S_-^k \mathbf{x}\right) - f\left(S_-^k S_{x_k}^i \mathbf{x}\right) = a_{ik}.$$

With reference to a fixed index $k \in \{1, ..., n+1\}$ we can write

$$v_f(\mathbf{x}) = \frac{1}{2(n+1)} \left(\sum_{j:j\neq k} a_{kj}^2 + \sum_{i:i\neq k} a_{ik}^2 + \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} a_{ij}^2 \right).$$

In the expression for $v_f(\mathbf{x}) - S_{y_k}^k v_f(\mathbf{x})$ the second sum in the last expression cancels, so

$$\begin{aligned} 0 &\leq v_f(\mathbf{x}) - S_{y_k}^k v_f(\mathbf{x}) \\ &\leq \frac{1}{2(n+1)} \left(\sum_{j:j\neq k} a_{kj}^2 + \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} \left(a_{ij}^2 - a_{ijk}^2\right) \right) \\ &= \frac{1}{2(n+1)} \left(\sum_{j:j\neq k} a_{kj}^2 + \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} (a_{ij} - a_{ijk})(a_{ij} + a_{ijk}) \right) \\ &\leq M(f)^2/2 + M(f) J(f). \end{aligned} \tag{16}$$

We square and sum over k, and use $(s+t)^2 \leq 2s^2 + 2t^2$ for real s, t, and Cauchy-Schwarz to obtain

$$\begin{aligned} D^2 v_f(\mathbf{x}) &= \sum_k \left(v_f(\mathbf{x}) - S_{y_k}^k v_f(\mathbf{x}) \right)^2 \\ &\leq \frac{1}{2(n+1)^2} \sum_k \left(\sum_{j:j\neq k} a_{kj}^2 \right)^2 + \\ &\quad + \frac{1}{2(n+1)^2} \sum_k \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} (a_{ij} - a_{ijk})^2 \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} (a_{ij} + a_{ijk})^2 \\ &= : A + B. \end{aligned}$$

We treat the two terms in turn. For A we get

$$\begin{aligned} A &= \frac{1}{2(n+1)^2} \sum_k \left(\sum_{j:j\neq k} a_{kj}^2 \right)^2 \\ &\leq \frac{M(f)^2}{2(n+1)} \sum_k \sum_{j:j\neq k} a_{kj}^2 = M(f)^2 v_f(\mathbf{x}). \end{aligned}$$

For B we again use $a_{ij} - a_{ijk} \leq J(f)/n$ and $(s+t)^2 \leq 2s^2 + 2t^2$ to get

$$\begin{aligned} B &= \frac{1}{2(n+1)^2} \sum_k \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} (a_{ij} - a_{ijk})^2 \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} (a_{ij} + a_{ijk})^2 \\ &\leq \frac{2J(f)^2}{n+1} \sum_k \left(\frac{1}{2(n+1)} \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} a_{ij}^2 + \frac{1}{2(n+1)} \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} a_{ijk}^2 \right) \end{aligned}$$

But by (15) for every $k \in \{1, ..., n+1\}$

$$\frac{1}{2(n+1)} \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} a_{ij}^2 \leq v_f(\mathbf{x})$$

and also, by the definition of y_k,

$$\frac{1}{2(n+1)} \sum_{i,j:i\neq j\wedge k\notin\{i,j\}} a_{ijk}^2 \leq S_{y_k}^k v_f(\mathbf{x}) \leq v_f(\mathbf{x}).$$

It follows that $B \leq 4J(f)^2 v_f(\mathbf{x})$ and

$$D^2 v_f(\mathbf{x}) \leq \left(M(f)^2 + 4J(f)^2\right) v_f(x).$$

Together with (16) this can be used in Corollary 10. Since

$$M(f)^2/2 + M(f) J(f) \leq \frac{1}{2}(M(f) + J(f))^2 \leq M(f)^2 + J(f)^2 \leq M(f)^2 + 4J(f)^2,$$

the corollary gives us for any $\delta > 0$ with probability at least $1-\delta$

$$\left|\sqrt{E[\Sigma^2(f)]} - \sqrt{v_f(\mathbf{X})}\right| \leq \sqrt{\left(2M(f)^2 + 8J(f)^2\right)\ln(2/\delta)}.$$

■

Appendix B. Proof of the normal approximation theorem

To prove Theorem 4 we use a result of Chatterjee (Chatterjee (2008), Theorem 2.2), for which we need extra notation. Let $\mathbf{X}' = (X_1', ..., X_n')$ be an independent copy of $\mathbf{X} = (X_1, ..., X_n)$. For a proper subset $A \subsetneqq \{1, ..., n\}$ define the vector $\mathbf{X}^A = \mathbf{X}^A(\mathbf{X}, \mathbf{X}')$ to be

$$X_i^A = \begin{cases} X_i' & \text{if} \quad i \in A \\ X_i & \text{if} \quad i \notin A \end{cases}.$$

For $A \subsetneqq \{1, ..., n\}$ define the random variables

$$\begin{aligned} T_A &= T_A(\mathbf{X}, \mathbf{X}') = \sum_{j\notin A} \left(D^j_{X_j, X_j'} f(\mathbf{X})\right)\left(D^j_{X_j, X_j'} f(\mathbf{X}^A)\right) \\ \text{and } T &= T(\mathbf{X}, \mathbf{X}') = \frac{1}{2} \sum_{A\subsetneqq\{1,...,n\}} \frac{T_A}{\binom{n}{|A|}(n-|A|)}. \end{aligned}$$

Theorem 11 *(Chatterjee) Let $f : \mathcal{X}^n \to \mathbb{R}$ and suppose $E[f] = 0$ and $\sigma^2(f) < \infty$. Then*

$$d_{\mathcal{N}}(f(\mathbf{X})) \leq \frac{\sqrt{\sigma^2(E[T|\mathbf{X}])}}{\sigma^2(f)} + \frac{1}{2\sigma^3(f)} \sum_{j=1}^n E\left[\left|D^j_{X_j, X_j'} f(\mathbf{X})\right|^3\right]. \tag{17}$$

Proof of Theorem 4 Both sides of the inequality we wish to prove do not change when a constant is added to f. We can therefore assume $E[f] = 0$ and use Chatterjee's theorem. We can bound the second term in (17) immediately by $nM(f)^3 / \left(2\sigma^3(f)\right)$, so the main work is in bounding $\sqrt{\sigma^2\left(E\left[T|\mathbf{X}\right]\right)}$. By the L_2-triangle inequality (Minkovsky-inequality) we have

$$\begin{aligned}\sqrt{\sigma^2\left(E\left[T|\mathbf{X}\right]\right)} &\leq \frac{1}{2}\sum_{A\subset\{1,...,n\}}\frac{\sqrt{\sigma^2\left(E\left[T_A|\mathbf{X}\right]\right)}}{\binom{n}{|A|}\left(n-|A|\right)} \\ &\leq \frac{1}{2}\sum_{A\subset\{1,...,n\}}\frac{\sqrt{E\left[\sigma^2\left(T_A|\mathbf{X}'\right)\right]}}{\binom{n}{|A|}\left(n-|A|\right)},\end{aligned}$$

where we used Lemma 4.4 in Chatterjee (2008) for the second inequality. So we first need to bound $E\left[\sigma^2\left(T_A|\mathbf{X}'\right)\right]$ for fixed $A \subsetneqq \{1,...,n\}$. This is done with the Efron Stein inequality Efron and Stein (1981), which gives

$$E\left[\sigma^2\left(T_A|\mathbf{X}'\right)\right] \leq \frac{1}{2}E\left[\sum_{i=1}^{n}\left(T_A\left(\mathbf{X},\mathbf{X}'\right) - S_i''T_A\left(\mathbf{X},\mathbf{X}'\right)\right)^2|\mathbf{X}'\right],$$

where $\mathbf{X}''$ is yet another independent copy of $\mathbf{X}$, and the operator S_i'' acts on functions of $2n$ variables and substitutes every occurence of X_i by X_i''

$$\left(S_i''F\right)\left(\mathbf{X},\mathbf{X}'\right) = F\left(X_1,...,X_{i-1},X_i'',X_{i+1},...,X_n,\mathbf{X}'\right).$$

Now let $V_j := D^j_{X_j,X_j'}f\left(\mathbf{X}\right)$, $W_j := D^j_{X_j,X_j'}f\left(\mathbf{X}^A\right)$, $V_{ij} := S_i''V_j = S_i^{''}D^j_{X_j,X_j'}f\left(\mathbf{X}\right)$ and $W_{ij} := S_i''W_j = S_i''D^j_{X_j,X_j'}f\left(\mathbf{X}^A\right)$. Observe that all of V_j, W_j, V_{ij} and W_{ij} have absolute value bounded by $M(f)$, and that for $i \neq j$

$$|V_j - V_{ij}| \leq J(f)/n \text{ and } |W_j - W_{ij}| \leq J(f)/n.$$

Then

$$\begin{aligned}&\sum_{i=1}^{n}\left(T_A\left(\mathbf{X},\mathbf{X}'\right) - S_i''T_A\left(\mathbf{X},\mathbf{X}'\right)\right)^2 \\ &= \sum_{i=1}^{n}\left(\sum_{j\notin A}V_jW_j - V_{ij}W_{ij}\right)^2 \\ &= \sum_{i=1}^{n}\left(\sum_{j\notin A,j\neq i}\left(V_jW_j - V_{ij}W_{ij}\right) + 1_{\{i\notin A\}}\left(V_iW_i - V_{ii}W_{ii}\right)\right)^2 \\ &\leq 2\sum_{i=1}^{n}\left(\sum_{j\notin A,j\neq i}V_jW_j - V_{ij}W_{ij}\right)^2 + 2\sum_{i\notin A}\left(V_iW_i - V_{ii}W_{ii}\right)^2 \\ &\leq 2\sum_{i=1}^{n}\left(\sum_{j\notin A,j\neq i}V_jW_j - V_{ij}W_{ij}\right)^2 + 8nM(f)^4.\end{aligned}$$

Now, using Cauchy Schwarz,

$$
\begin{aligned}
&2\sum_{i=1}^{n}\left(\sum_{j\notin A,j\neq i} V_jW_j - V_{ij}W_{ij}\right)^2 \\
&= 2\sum_{i=1}^{n}\left(\sum_{j\notin A,j\neq i} (V_j - V_{ij})\,W_j + V_{ij}\,(W_j - W_{ij})\right)^2 \\
&\leq 4\sum_{i=1}^{n}\left(\sum_{j\notin A,j\neq i} (V_j - V_{ij})\,W_j\right)^2 + 4\sum_{i=1}^{n}\left(\sum_{j\notin A,j\neq i} V_{ij}\,(W_j - W_{ij})\right)^2 \\
&\leq 4\sum_{i=1}^{n}\sum_{j\notin A,j\neq i}(V_j - V_{ij})^2 \sum_{j\notin A,j\neq i} W_j^2 + 4\sum_{i=1}^{n}\sum_{j\notin A,j\neq i} V_{ij}^2 \sum_{j\notin A,j\neq i}(W_j - W_{ij})^2 \\
&\leq 8\sum_{i=1}^{n}\sum_{j\notin A,j\neq i}\frac{J(f)^2}{n^2}\sum_{j\notin A,j\neq i} M(f)^2 \\
&\leq 8nM(f)^2\,J(f)^2\,.
\end{aligned}
$$

Putting the chains of inequalities together and using $\sqrt{s+t} \leq \sqrt{s} + \sqrt{t}$ we conclude that

$$\sqrt{E\left[\sigma^2\left(T_A|\mathbf{X}'\right)\right]} \leq 2\sqrt{n}M(f)\left(M(f) + J(f)\right).$$

Thus

$$
\begin{aligned}
\sqrt{\sigma^2\left(E\left[T|X\right]\right)} &\leq \frac{1}{2}\sum_{A\subsetneq\{1,\ldots,n\}}\frac{\sqrt{E\left[\sigma^2\left(T_A|X'\right)\right]}}{\binom{n}{|A|}\left(n-|A|\right)} \\
&\leq \sqrt{n}M(f)\left(J(f) + M(f)\right)\sum_{k=1}^{n-1}\sum_{A:|A|=k}\frac{\binom{n-1}{k}}{\binom{n}{k}\left(n-k\right)} \\
&= \sqrt{n}M(f)\left(J(f) + M(f)\right)
\end{aligned}
$$

By Theorem 11 and the bound on the last term of (17)

$$d_{\mathcal{N}}\left(f\left(\mathbf{X}\right)\right) \leq \frac{\sqrt{n}M(f)\left(J(f) + M(f)\right)}{\sigma^2(f)} + \frac{nM(f)^3}{2\sigma^3(f)}.$$

■

Appendix C. Lipschitz L-statistics revisited

We show that the Lipschitz L-statistics of Section 3.2 have $\left(\|F\|_\infty, \|F\|_{Lip}\right)$-weak interactions. For $\alpha, \beta \in \mathbb{R}$ let $[[\alpha, \beta]]$ be the interval $[\min\{\alpha,\beta\}, \max\{\alpha,\beta\}]$. That f as defined in equation (11) has $\left(\|F\|_\infty, \|F\|_{Lip}\right)$-weak interactions is clearly implied by

Theorem 12 *With f as (11) we have*

$$\left|D^k_{y,y'}f(\mathbf{x})\right| \leq \frac{\|F\|_\infty \, diam\left([[y,y']]\right)}{n} \tag{18}$$

$$\left|D^l_{z,z'}D^k_{y,y'}f(\mathbf{x})\right| \leq \frac{\|F\|_{Lip} \, diam\left([[z,z']]\cap[[y,y']]\right)}{n^2} \tag{19}$$

for any $\mathbf{x}\in[0,1]^n$, all $k\neq l$ and all $y,y',z,z'\in[0,1]$.

Proof Suppose we can prove the inequalities (18) and (19) for all $\mathbf{x}\in[0,1]^n$ and all $k\neq l$ and in the three cases

a	$z'\leq z<y'\leq y$	$[[z,z']]\cap[[y,y']]=\emptyset$, non-intersection
b	$z'\leq y'\leq y\leq z$	$[[y,y']]\subseteq[[z,z']]$, inclusion
c	$z'\leq y'\leq z\leq y$	partial intersection.

The right collumn above enumerates all possible relationships between $[[z,z']]$ and $[[y,y']]$. Then, as (18) and (19) are invariant under the exchanges of $k\leftrightarrow l$, $z\leftrightarrow z'$ and $y\leftrightarrow y'$, we have proven these inequalities for all possible orderings of z,z',y and y'. It therefore suffices to prove the above inequality in the three cases a, b and c.

To further simplify the problem we introduce the vector $\hat{\mathbf{x}}\in[0,1]^n$ defined by

$$\hat{x}_i=\left(S^l_{z'}S^k_{y'}\mathbf{x}\right)_{(i)}.$$

Then $\hat{\mathbf{x}}$ is already ordered, and there are $\hat{l}$ and $\hat{k}$ in $\{1,...,n\}$ such that $\hat{l}\neq\hat{k}$ and $\hat{x}_{\hat{l}}=z'$ and $\hat{x}_{\hat{k}}=y'$. Write $F_i=F(i/n)$, so that $|F_i|\leq\|F\|_\infty$ and $|F_i-F_{i-1}|\leq\|F\|_{Lip}/n$. Transcribing to the new variables and omitting the "^"-symbols, it becomes appearant that we have to prove the inequalities

$$A := \left|\sum_{i=1}^n F_i\left(\mathbf{x}_{(i)}-\left(S^k_y\mathbf{x}\right)_{(i)}\right)\right|\leq\|F\|_\infty \operatorname{diam}\left(\left[\left[y,y'\right]\right]\right) \text{ and}$$

$$B := \left|\sum_{i=1}^n F_i\left(\mathbf{x}_{(i)}-\left(S^k_y\mathbf{x}\right)_{(i)}-\left(\left(S^l_z\mathbf{x}\right)_{(i)}-\left(S^l_zS^k_y\mathbf{x}\right)_{(i)}\right)\right)\right|$$

$$\leq \frac{\|F\|_{Lip}\operatorname{diam}\left([[z,z']]\cap[[y,y']]\right)}{n}$$

for all $\mathbf{x}\in[0,1]^n$, which are already ordered with $x_i\leq x_{i+1}$, and all $k\neq l$ and in the three cases

a	$x_l\leq z<x_k\leq y$
b	$x_l\leq x_k\leq y\leq z$.
c	$x_l\leq x_k\leq z\leq y$

We let $p,q\in\{1,...,n\}$ be such that

$$\left(S^k_y\mathbf{x}\right)_{(p)}=y \text{ and } \left(S^l_z\mathbf{x}\right)_{(q)}=z.$$

The effect which modifying an argument has on the order statistic is a shift and the replacement of a boundary term. For $x_k\leq y$ we have

$$\left(S^k_y\mathbf{x}\right)_{(i)}=\begin{cases} x_i & \text{if} \quad i\notin\{k,...,p\} \\ x_{i+1} & \text{if} \quad i\in\{k,...,p-1\} \\ y & \text{if} \quad i=p \end{cases}.$$

It follows that in all cases a, b and c

$$\begin{aligned} A &= \left|\sum_{i=k}^{p-2} F_i\left(x_i - x_{i+1}\right) + F_{p-1}\left(x_{p-1} - y\right)\right| \\ &\leq \|F\|_\infty \left(\sum_{i=k}^{p-2} |x_i - x_{i+1}| + |x_{p-1} - y|\right) \leq \|F\|_\infty (y - x_k), \end{aligned}$$

which gives the bound on A and therefore (18).

For the second inequality it is easy to see that $B = 0$ whenever $[[x_k, y]]$ and $[[x_l, z]]$ don't intersect, as in **Case a**, so we consider only the cases b and c.

Case b (inclusion, $x_l \leq x_k \leq y \leq z$). By partial summation we get

$$\begin{aligned} B &= \left|\sum_{i=k}^{p-1} \left(F_i - F_{i-1}\right)\left(x_i - x_{i+1}\right) + \left(F_p - F_{p-1}\right)\left(x_p - y\right)\right| \\ &\leq \frac{\|F\|_{Lip}}{n} \sum_{i=k}^{p-1} |x_i - x_{i+1}| + \frac{\|F\|_{Lip}}{n} |x_p - y| \\ &= \frac{\|F\|_{Lip}}{n} (y - x_k). \end{aligned}$$

The general principle here is partial summation and the fact that the sum of absolute differences always collapses to the diameter of an interval because of the ordering.

Case c, (partial intersection, $x_l \leq x_k \leq z \leq y$).

$$\begin{aligned} B &= \left|\sum_{i=k}^{q-1} \left(F_i - F_{i-1}\right)\left(x_i - x_{i+1}\right) + \left(F_q - F_{q-1}\right)\left(x_q - z\right)\right| \\ &\leq \frac{\|F\|_{Lip}}{n} \sum_{i=k}^{q-1} |x_i - x_{i+1}| + \frac{\|F\|_{Lip}}{n} |x_q - z| \\ &= \frac{\|F\|_{Lip}}{n} (z - x_k). \end{aligned}$$

■

Appendix D. Proof of the chain rule

Proof of Lemma 8 Take arbitrary $\mathbf{x} \in \mathcal{X}^n$, $y, y', z, z' \in \mathcal{X}$ and any $k, l, k \neq l$. Define a linear function $h : [0,1] \to U$ by

$$h(t) = tf\left(S_y^k \mathbf{x}\right) + (1-t) f\left(S_{y'}^k \mathbf{x}\right).$$

Then $h'(t) = D_{y,y'}^k f(\mathbf{x})$ and

$$\begin{aligned} D_{y,y'}^k F \circ f(\mathbf{x}) &= F(h(1)) - F(h(0)) = \int_0^1 F'(h(t)) h'(t)\, dt \\ &\leq \int_0^1 \left\|F'(h(t))\right\| \left\|D_{y,y'}^k f(\mathbf{x})\right\| dt \leq \sup_{v \in U} \left\|F'(v)\right\| M(f). \end{aligned}$$

This proves the first inequality. For the bound on J define a bilinear function $g:[0,1]\times[0,1]\to U$ by

$$g(s,t)=stf\left(S_z^lS_y^k\mathbf{x}\right)+s(1-t)f\left(S_z^lS_{y'}^k\mathbf{x}\right)+t(1-s)f\left(S_{z'}^lS_y^k\mathbf{x}\right)+(1-s)(1-t)f\left(S_{z'}^lS_{y'}^k\mathbf{x}\right).$$

Then $\left\|\frac{\partial}{\partial t}g(s,t)\right\|=\left\|sD_{y,y'}^kf\left(S_z^l\mathbf{x}\right)+(1-s)D_{y,y'}^kf\left(S_{z'}^l\mathbf{x}\right)\right\|\leq M(f)$ and similarly $\left\|\frac{\partial}{\partial t}g(s,t)\right\|\leq M(f)$ and also $\left\|\frac{\partial^2}{\partial s\partial t}g(s,t)\right\|=\left\|D_{zz'}^lD_{yy'}^k\right\|\leq J(f)/n$. Thus

$$\begin{aligned}\left|\frac{\partial^2}{\partial s\partial t}F(g(s,t))\right| &= \left|F''(g(s,t))\frac{\partial}{\partial t}g(s,t)\frac{\partial}{\partial s}g(s,t)+F'(g(s,t))\frac{\partial^2}{\partial s\partial t}g(s,t)\right|\\ &\leq \left\|F''(g(s,t))\right\|\left\|\frac{\partial}{\partial t}g(s,t)\right\|\left\|\frac{\partial}{\partial s}g(s,t)\right\|+\left\|F'(g(s,t))\right\|\left\|\frac{\partial^2}{\partial s\partial t}g(s,t)\right\|\\ &\leq \left\|F''(g(s,t))\right\|M(f)^2+\left\|F'(g(s,t))\right\|J(f)/n\end{aligned}$$

So that

$$\begin{aligned}D_{zz'}^lD_{yy'}^kF\circ f(\mathbf{x}) &= F(g(1,1))-F(g(1,0))-(F(g(0,1))-F(g(0,0)))\\ &= \int_0^1\int_0^1\frac{\partial^2}{\partial s\partial t}F(g(s,t))\,ds\,dt\\ &\leq \left\|F''(g(s,t))\right\|M(f)^2+\left\|F'(g(s,t))\right\|J(f)/n.\end{aligned}$$

The second inequality follows. ■

Appendix E. The Gibbs algorithm

We substantiate the claims in Section 3.5. For $G\in L_\infty(\Omega)$ define

$$Z(G)=\int_\Omega e^{-\beta G(\omega)}d\rho(\omega)$$

and an expectation functional

$$E_G[h]:=Z(G)^{-1}\int_\Omega h(\omega)e^{-\beta G(\omega)}d\rho(\omega)\ \text{ for } h\in L_\infty(\Omega).$$

Then

$$\begin{aligned}KL(d\pi_{\mathbf{x}},d\pi) &= E_{H(\cdot,\mathbf{x})}\left[\ln\left(\frac{Z^{-1}(\mathbf{x})e^{-\beta H(\omega,\mathbf{x})}}{Z_0^{-1}e^{-\beta H_0(\omega)}}\right)\right]\\ &= F\circ f(\mathbf{x}),\end{aligned}$$

where f is defined in equation (13) and F the real function defined on the unit ball $\mathbb{B}_1$ of $L_\infty(\Omega)$ by

$$F(G):=E_G\left[\ln\left(\frac{Z(G)^{-1}e^{-\beta G}}{Z_0^{-1}e^{-\beta H_0}}\right)\right]=\beta E_G[H_0-G]-\ln Z(G)+\ln Z_0.$$

To apply the chain rule we have to bound the derivatives of $F = \beta(\Psi - \Phi) - \Xi + \ln Z_0$, where $\Psi, \Phi, \Xi : \mathbb{B}_1 \to \mathbb{R}$ are the functions

$$\Xi(G) = \ln Z(G), \Phi(G) = E_G[G] \text{ and} \Psi(G) = E_G[H_0].$$

Differentiating we find

$$\begin{aligned} \Xi'(G)[u] &= -\beta E_G[u] \\ \Xi''(G)[u][v] &= \beta^2 (E_G[uv] - E_G[u] E_G[v]) \end{aligned}$$

so that $\|\Xi'\| \leq \beta$ and $\|\Xi''\| \leq 2\beta^2$. We also have

$$\begin{aligned} \Phi'(G)[u] &= \beta E_G[G] E_G[u] - \beta E_G[Gu] + E_G[u] \\ \Psi'(G)[u] &= \beta E_G[H_0] E_G[u] - \beta E_G[H_0 u]. \end{aligned}$$

Since $\|H_0\|, \|G\| \in \mathbb{B}_1$ we have $\|\Phi'\| \leq 2\beta + 1$ and $\|\Psi'\| \leq 2\beta$. By a somewhat tedious computation

$$\begin{aligned} \Phi''[u][v] &= 2\beta^2 E_G[G] E_G[v] E_G[u] - \beta^2 E_G[G] E_G[vu] + \beta^2 E_G[Guv] \\ &\quad -\beta^2 E_G[Gv] E_G[u] - \beta^2 E_G[Gu] E_G[v] - 2\beta E_G[uv] + 2\beta E_G[u] E_G[v], \end{aligned}$$

which gives $\|\Phi''\| \leq 6\beta^2 + 4\beta$. Similarly, and a bit simpler, one obtains $\|\Psi''\| \leq 6\beta^2$. Adding these estimates we get

$$\begin{aligned} F &= \beta(\Psi - \Phi) - \Xi + \ln Z_0 \\ \left\|F'\right\| &\leq 4\beta^2 + 2\beta \\ \left\|F''\right\| &\leq 12\beta^3 + 6\beta^2. \end{aligned}$$

The chain rule then gives

$$\begin{aligned} M(F \circ f) &\leq \left(4\beta^2 + 2\beta\right) M(f) \leq \frac{4\beta^2 + 2\beta}{n} \\ J(F \circ f) &\leq n\left(12\beta^3 + 6\beta^2\right) M(f)^2 + 0 \leq \frac{12\beta^3 + 6\beta^2}{n}. \end{aligned}$$

Appendix F. Table of notation

Symbol	Quick description	Section
$\mathcal{X}$	space of observarions	1
X_i	independent random variables in $\mathcal{X}$	1
μ_i	distribution of X_i	1
$\mathbf{X}$	random vector composed of the X_i	2.1
$\mathcal{A}_n$	bounded measurable functions $f : \mathcal{X}^n \to \mathbb{R}$	2.1
$\mathbf{x}$	vector in $\mathcal{X}^n$	2.1
$E[f]$	$E[f] = E[f(\mathbf{X})] = E[f(X_1, ..., X_n)]$ for $f \in \mathcal{A}_n$	2.1
$\sigma^2(f)$	Variance of $f(X_1, ..., X_n)$ for $f \in \mathcal{A}_n$	2.1
$D^k_{y,y'}$	partial difference operator	1
S^k_y	substitution operator	2.2
S^k_-	deletion operator	2.2
$M(f)$	distance to constant functions	1
$J(f)$	distance to additive functions	1
$\sigma^2_k(f)$	k-th conditional variance	2.1
$\Sigma^2(f)$	sum of conditional variances	2.1
v_n	sample variance	1
v_f	variance estimator for $f \in \mathcal{A}_n$. Note $v_f \in \mathcal{A}_{n+1}$	2.2
K_-, K_+	estimation error coefficients for v_f	2.2
$d_{\mathcal{N}}$	distance to normality	2.3