Bernoulli **25**(2), 2019, 1451–1471
https://doi.org/10.3150/18-BEJ1026

A Bernstein-type inequality for functions of bounded interaction

ANDREAS MAURER

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

We give a distribution-dependent concentration inequality for functions of independent variables. The result extends Bernstein's inequality from sums to more general functions, whose variation in any argument does not depend too much on the other arguments. Applications sharpen existing bounds for U-statistics and the generalization error of regularized least squares.

Keywords: Bernstein inequality; concentration; u-statistics

1. Introduction

If $X_1, \dots, X_n$ are independent real random variables, with $X_k - EX_k \leq 1$ almost surely, $E[X_k^2] < \infty$ and $f(X_1, \dots, X_n) = \sum_k X_k$, then Bernstein's inequality ([3]) asserts that for $t > 0$

$$\Pr\bigl\{f(X_1, \dots, X_n) - E\bigl[f(X_1, \dots, X_n)\bigr] > t\bigr\} \leq \exp\left(\frac{-t^2}{2\sum_k \sigma_k^2 + 2t/3}\right),$$

where σ_k^2 is the variance of X_k. This work gives an extension of Bernstein's inequality to more general functions f.

This extension requires two modifications. First, the variance $\sum_k \sigma_k^2$ is replaced by the Efron–Stein upper bound, or jackknife estimate, of the variance. Second, a correction term $J(f)$ is added to the coefficient $2/3$ of t in the denominator of the exponent. This correction term, which we call the interaction functional of f, vanishes for sums and represents the extent to which the variation of f in any given argument depends on other arguments.

To proceed, we introduce some notation. Let $(\Omega, \mathcal{T}) = \prod_{k=1}^n (\Omega_k, \mathcal{T})$ be some product of measurable spaces and let $\mathcal{A}(\Omega)$ be the algebra of all bounded, measurable real valued functions on Ω. For fixed $k \in \{1, \dots, n\}$ and $y, y' \in \Omega_k$ define the substitution operator S_y^k and the difference operator $D_{y,y'}^k$ on $\mathcal{A}(\Omega)$ by

$$\bigl(S_y^k f\bigr)(x_1, \dots, x_n) = f(x_1, \dots, x_{k-1}, y, x_{k+1}, \dots, x_n)$$

and $D_{y,y'}^k = S_y^k - S_{y'}^k$. With $\mathcal{A}_k(\Omega)$ we denote the subalgebra of functions in $\mathcal{A}(\Omega)$ which do not depend on the kth argument. Both $S_y^k f$ and $D_{y,y'}^k f$ are in $\mathcal{A}_k(\Omega)$.

Let a probability measure μ_k be given on each Ω_k and let μ be the product measure $\mu = \prod \mu_k$ on Ω. For $f \in \mathcal{A}(\Omega)$, the expectation Ef and variance $\sigma^2(f)$ are defined as $Ef = \int_\Omega f \, d\mu$ and $\sigma^2(f) = E[(f - Ef)^2]$. For $k \in \{1, \dots, n\}$, the conditional expectation E_k (conditional on the

σ-algebra generated by the members of $\mathcal{A}_k(\Omega)$) and the conditional variance σ_k^2 are operators on $\mathcal{A}(\Omega)$, which act on a function $f \in \mathcal{A}(\Omega)$ as

$$E_k f = E_{y\sim\mu_k}\big[S_y^k f\big] = \int_{\Omega_k} S_y^k f \sim d\mu_k(y) \quad \text{and}$$

$$\sigma_k^2(f) = E_k\big[(f - E_k f)^2\big] = \frac{1}{2} E_{(y,y')\sim\mu_k^2}\big[\big(D_{y,y'}^k f\big)^2\big],$$

where μ_k^2 is the product measure $\mu_k \times \mu_k$ on $\Omega_k \times \Omega_k$. The sum of conditional variances is the operator $\Sigma^2 : \mathcal{A}(\Omega) \to \mathcal{A}(\Omega)$ defined by

$$\Sigma^2(f) = \sum_{k=1}^n \sigma_k^2(f).$$

This operator appears in the Efron–Stein inequality ([8,18], see also Section 2.4) as

$$\sigma^2(f) \le E\big[\Sigma^2(f)\big],$$

which becomes an equality if f is a sum of real valued functions X_k on Ω_k. It also appears in the following exponential tail bound ([17], Theorem 3.8, or [15], Theorem 11).

Suppose that $f \in \mathcal{A}(\Omega)$ satisfies $f - E_k f \le b$ for all $k \in \{1, \dots, n\}$. Then

$$\Pr\{f - Ef > t\} \le \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\Omega} \Sigma^2(f)(\mathbf{x}) + 2bt/3}\right). \tag{1.1}$$

This inequality reduces to Bernstein's inequality if f is a sum, but it suffers from the worst-case choice of the configuration $\mathbf{x}$, for which $\Sigma^2(f)(\mathbf{x})$ is evaluated. One would like to replace the supremum by an expectation, just as in the Efron–Stein inequality.

This replacement is trivially possible when f is a sum, because then $\Sigma^2(f)$ is constant. It turns out that it is also possible if $\Sigma^2(f)$ has the right properties of concentration about its mean, a weak form of being constant. To insure this we control the interaction between the different arguments of f, in the sense that the variation in any argument must not depend too much on the other arguments.

Definition 1. The interaction functional $J : \mathcal{A}(\Omega) \to \mathbb{R}_0^+$ is defined by

$$J(f) = \left(\sup_{\mathbf{x}\in\Omega} \sum_{k,l:k\neq l} \sup_{z,z'\in\Omega_l} \sup_{y,y'\in\Omega_k} \big(D_{z,z'}^l D_{y,y'}^k f\big)^2(\mathbf{x})\right)^{1/2} \qquad \text{for } f \in \mathcal{A}(\Omega).$$

The distribution-dependent interaction functional J_μ is defined by

$$J_\mu(f) = 2\left(\sup_{\mathbf{x}\in\Omega} \sum_l \sup_{z\in\Omega_l} \sum_{k:k\neq l} \sigma_k^2\big(f - S_z^l f\big)(\mathbf{x})\right)^{1/2}.$$

These quantities are related and bounded using the inequalities

$$\begin{aligned} J_\mu(f) &\le J(f) \\ &\le n \sup_{\mathbf{x}\in\Omega} \max_{k,l} \sup_{z,z'\in\Omega_l} \sup_{y,y'\in\Omega_k} \left(D^l_{z,z'} D^k_{y,y'} f\right)(\mathbf{x}) \end{aligned} \tag{1.2}$$

(see the end of Section 2.3). For our applications below the last, simplest and crudest bound appears to be sufficient. The above functionals and bounds vanish for sums and are positive homogeneous of degree one. The following is our main result.

Theorem 1. *Suppose $f \in \mathcal{A}(\Omega)$ satisfies $f - E_k f \le b$ for all k. Then for all $t > 0$*

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

Remarks:

1. If this is applied to sums of independent random variables (real valued functions X_k defined on Ω_k), we recover Bernstein's inequality.

2. Consider the case that $\Omega_k = \Omega_0$, $\mu_k = \mu_0$ and a sequence of functions $f_n \in \mathcal{A}(\Omega_0^n)$, such that $J_\mu(f_n)/\sqrt{n} \to 0$ (for example if $J_\mu(f_n)$ is bounded) and such that the limit $\sigma^2 = \lim_{n\to\infty} E[\Sigma^2(f_n)]/n$ exists. Applying Theorem 1 to the sequence $f_n/\sqrt{n}$, and letting $n \to \infty$, we obtain the tail of a normal distribution with variance σ^2. In some cases, like U-statistics, this is known to be the correct limiting distribution ([10], Theorem 7.1).

3. Although the distribution dependent functional J_μ is potentially much smaller than J, in the applications considered sofar it seems sufficient to consider J or the above bounds thereof.

4. Since $E[\Sigma^2(f)] \le \sup_{\mathbf{x}} \Sigma^2(f)(\mathbf{x}) \le \sup_{\mathbf{x}} (1/4) \sum_k \sup_{y,y'} (D^k_{y,y'}(f))^2(\mathbf{x})$, the variance term above can never be larger than the variance term in (1.1), which in turn can never be larger than what we get from the bounded difference inequality ([17], Theorem 3.7, or [5], Theorem 6.5).

5. If also $f - E_k f \ge -b$, then the result can be applied to $-f$ so as to obtain a two-sided inequality.

In Theorem 2.1, [11] bounds the bias in the Efron–Stein inequality in terms of iterated jackknifes, which correspond to the expectations of higher order differences. The second of these iterates can be bounded in terms of the interaction functional and allows us to put the variance $\sigma^2(f)$ back into the inequality of Theorem 1.

Proposition 1.

$$E\left[\Sigma^2(f)\right] \le \sigma^2(f) + \frac{1}{4} H^2(f) \le \sigma^2(f) + \frac{1}{4} J^2(f),$$

where the functional $H : \mathcal{A}(\Omega) \to \mathbb{R}$ is defined by

$$H(f) = \left(E_{\mathbf{x}\sim\mu} \sum_{k,l:k\ne l} E_{(z,z')\sim\mu_l^2} E_{(y,y')\sim\mu_l^2} \left[\left(D^l_{z,z'} D^k_{y,y'} f\right)^2(\mathbf{x}) \right] \right)^{1/2}.$$

See Section 2.4 for the proof. In combination with Theorem 1 we obtain the following corollary.

Corollary 1. *Suppose $f \in \mathcal{A}$ and $f - E_k f \leq b$ for all k. Then for all $t > 0$*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{2\sigma^2(f) + H^2(f)/2 + (2b/3 + J_\mu(f))t}\right).$$

We apply Theorem 1 in two different situations. For undecoupled U-statistics of any order with bounded, symmetric kernels it is very easy to bound the interaction functional, so as to obtain an asymptotically sharp Bernstein inequality, with mild dependence on the order of the U-statistic. Section 3 gives a derivation and discussion of this inequality.

In a different context Theorem 1 sharpens a stability based generalization bound for regularized least squares. This application is discussed in Section 4.

The idea of using second differences (as in the definition of J) has been put to work in [11] to estimate the bias in the Efron–Stein inequality. The entropy method, which underlies our proof of Theorem 1, has been developed by a number of authors, notably [13] and [4]. The latter work also introduces the key-idea of combining it with the decoupling method used below. Our proof follows a thermodynamic formulation of the entropy method as laid out in [15].

The next section gives a proof of Theorem 1. Then follow the applications to U-statistics and ridge regression.

2. Proof of Theorem 1

The proof of our main result, Theorem 1, uses the entropy method ([4,5,13]), from which the next section collects a set of tools. These results are taken from [15], which provides proofs and additional motivation.

2.1. Definitions and tools

Ω and $\mathcal{A}(\Omega)$ are as in the introduction, $\mathcal{A}_k(\Omega)$ is the subalgebra of $\mathcal{A}(\Omega)$ of those bounded, measurable functions on Ω which are independent of the kth coordinate. For $f \in \mathcal{A}(\Omega)$ and $\beta \in \mathbb{R}$ define the expectation functional $E_{\beta f}$ on $\mathcal{A}(\Omega)$ by

$$E_{\beta f}[g] = Z_{\beta f}^{-1} E\left[g e^{\beta f}\right], \qquad g \in \mathcal{A}(\Omega),$$

where $Z_{\beta f} = E[e^{\beta f}]$. The entropy $S_f(\beta)$ of f at β is given by

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

where $\mathrm{KL}(\nu, \mu)$ is the Kullback–Leibler divergence.

Lemma 1 (Theorem 1 in [15]). *For any $f \in \mathcal{A}(\Omega)$ and $\beta > 0$ we have*

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

and, for $t \geq 0$,

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

Define the real function ψ by $\psi(t) := te^t - e^t + 1$.

Lemma 2 (Lemma 10 in [15]). *Let $f \in \mathcal{A}(\Omega)$ satisfy $f - E_k f \leq 1$ for all $k \in \{1, \ldots, n\}$. Then for $\beta > 0$*

$$S_f(\beta) \leq \psi(\beta) E_{\beta f}\big[\Sigma^2(f)\big].$$

Bounding $E_{\beta f}[\Sigma^2(f)] \leq \sup_{\mathbf{x}} \Sigma^2(f)(\mathbf{x})$ and using Lemma 1 quickly leads to a proof of inequality (1.1). For Theorem 1, we need more tools.

Definition 2. The operator $D : \mathcal{A}(\Omega) \to \mathcal{A}(\Omega)$ is defined by

$$Dg = \sum_k \Big(g - \inf_{y \in \Omega_k} S_y^k g\Big)^2, \qquad \text{for } g \in \mathcal{A}(\Omega).$$

To clarify: $\inf_{y\in\Omega_k} S_y^k g$ is the member of $\mathcal{A}(\Omega)$ defined by

$$\Big(\inf_{y \in \Omega_k} S_y^k g\Big)(\mathbf{x}) = \inf_{y \in \Omega_k} \big(S_y^k\big(g(\mathbf{x})\big)\big).$$

It does not depend on x_k, so $\inf_{y\in\Omega_k} S_y^k g \in \mathcal{A}_k(\Omega)$.

Lemma 3 (Lemma 15 in [15], also Proposition 5 in [14]). *We have, for $\beta > 0$, that*

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

We use this to derive the following lemma, which, together with Proposition 2 below, gives the concentration property of $\Sigma^2(f)$ alluded to in the introduction.

Lemma 4. *Suppose that*

$$Df \leq a^2 f. \tag{2.1}$$

Then for $\beta \in (0, 2/a^2)$

$$\ln E\big[e^{\beta f}\big] \leq \frac{\beta Ef}{1 - a^2\beta/2}. \tag{2.2}$$

Functions satisfying (2.1) are called weakly self-bounded in [5].

Proof of Lemma 4. Using Lemma 1 and Lemma 3 and the weak self-boundedness assumption (2.1) we have for $\beta > 0$ that

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

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

$$\ln E\big[e^{\beta f}\big] \le \frac{a^2\beta}{2}\ln E e^{\beta f} + \beta E f,$$

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

We also use the following decoupling technique: If μ and ν are two probability measures and ν is absolutely continuous w.r.t. μ, then it is easy to show that

$$E_\nu g \le \mathrm{KL}(d\nu, d\mu) + \ln E_\mu e^g.$$

Applying this inequality when ν is the measure $Z_{\beta f}^{-1} e^{\beta f}\,d\mu$, we obtain the following

Lemma 5. *We have for any $g \in \mathcal{A}(\Omega)$ that*

$$E_{\beta f}[g] \le S_f(\beta) + \ln E\big[e^g\big]. \tag{2.3}$$

2.2. Self-boundedness of the sum of conditional variances

We record some properties of the substitution operator. For $k \in \{1,\dots,n\}$ and $y \in \Omega_k$ the operator S_y^k is a homomorphism (linear and multiplicative) on $\mathcal{A}(\Omega)$ and the identity on $\mathcal{A}_k(\Omega)$. If $l \ne k$, it commutes with S_z^l and with E_l. Most importantly

$$S_y^k \sigma_l^2(f) = \frac{1}{2} S_y^k E_{(z,z')\sim\mu_l^2}\big[\big(D_{z,z'}^l f\big)^2\big] = \frac{1}{2} E_{(z,z')\sim\mu_l^2}\big[\big(D_{z,z'}^l S_y^k f\big)^2\big] = \sigma_l^2\big(S_y^k f\big).$$

Note however that for $l = k$, we get $S_y^k S_z^k = S_z^k$ and $S_y^k E_k = E_k$ and $S_y^k \sigma_k^2 = \sigma_k^2$, because S_z^k, E_k and σ_k^2 map to $\mathcal{A}_k(\Omega)$.

Proposition 2. *We have $D(\Sigma^2(f)) \le J_\mu(f)^2 \Sigma^2(f)$ for any $f \in \mathcal{A}(\Omega)$.*

Proof. Fix $\mathbf{x} \in \Omega$. Below all members of $\mathcal{A}(\Omega)$ are understood as evaluated on $\mathbf{x}$. For $l \in \{1, \dots, n\}$ let $z_l \in \Omega_l$ be a minimizer in z of $S_z^l \Sigma^2(f)$ (existence is assumed for simplicity, an approximate minimizer would also work), so that

$$\inf_{z \in \Omega_l} S_z^l \Sigma^2(f) = S_{z_l}^l \Sigma^2(f) = \sum_k S_{z_l}^l \sigma_k^2(f) = \sigma_l^2(f) + \sum_{k:k \neq l} S_{z_l}^l \sigma_k^2(f),$$

where we used the fact that $S_{z_l}^l \sigma_l^2(f) = \sigma_l^2(f)$, because $\sigma_l^2(f) \in \mathcal{A}_l(\Omega)$. Then

$$D\big(\Sigma^2(f)\big) = \sum_l \Big(\Sigma^2(f) - \inf_{z_l \in \Omega_l} S_z^l \Sigma^2(f)\Big)^2 = \sum_l \Big(\sum_{k:k \neq l} \big(\sigma_k^2(f) - S_{z_l}^l \sigma_k^2(f)\big)\Big)^2.$$

This step gave us a sum over $k \neq l$, which is important, because it allows us to use the commutativity properties mentioned above. Then, using $2\sigma_k^2(f) = E_{(y,y') \sim \mu_k^2} (D_{y,y'}^k f)^2$, we get

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

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

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

so that

$$\begin{aligned} D\big(\Sigma^2(f)\big) &\le 2\sum_l \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\big[D^k_{y,y'}\big(f - S^l_{z_l} f\big)\big]^2 \Sigma^2(f) \\ &= 4\sum_l \sum_{k:k\neq l} \sigma_k^2\big(f - S^l_{z_l} f\big)\Sigma^2(f) \\ &\le 4\sup_{\mathbf{x}\in\Omega}\sum_l \sup_{z\in\Omega_l}\sum_{k:k\neq l} \sigma_k^2\big(f - S^l_z f\big)(\mathbf{x})\Sigma^2(f) \\ &= J_\mu^2(f)\Sigma^2(f). \end{aligned}$$

□

2.3. Proof of Theorem 1

We need two more auxiliary results. Recall the definition of the function $\psi(t) := te^t - e^t + 1$.

Lemma 6. *For any $a \ge 0$ and $0 \le \gamma < 1/(1/3 + a/2)$ we have*

(i) $a\sqrt{\psi(\gamma)/2} < 1$,
(ii)

$$\frac{\psi(\gamma)}{\gamma^2(1 - a\sqrt{\psi(\gamma)/2})^2} \le \frac{1}{2(1 - (1/3 + a/2)\gamma)^2}.$$

Proof. If $0 \le \gamma < 1/(1/3+a/2)$ and $a \ge 0$, then $\gamma < 3$. In this case, we have the two convergent power series representations

$$\begin{aligned} \frac{1}{2(1-\gamma/3)^2} &= \sum_{n=0}^\infty \frac{n+1}{2} 3^{-n}\gamma^n =: \sum_{n=0}^\infty b_n\gamma^n, \\ \frac{\gamma e^\gamma - e^\gamma + 1}{\gamma^2} &= \sum_{n=0}^\infty \frac{1}{(n+2)n!}\gamma^n =: \sum_{n=0}^\infty c_n\gamma^n. \end{aligned}$$

Now $b_0 = c_0 = 1/2$ by inspection and for $n \ge 1$

$$\frac{b_n}{c_n} = \frac{(n+2)!}{2\times 3^n} = \frac{1\times 2}{2} \times \prod_{k=1}^n \left(\frac{k+2}{3}\right) \ge 1,$$

so that $b_n \ge c_n$ for all non-negative n. Term by term comparison of the two power series gives

$$\frac{\psi(\gamma)}{\gamma^2} = \frac{\gamma e^\gamma - e^\gamma + 1}{\gamma^2} \le \frac{1}{2(1-\gamma/3)^2}, \tag{2.4}$$

which is (ii) in the case that $a = 0$.

It also gives us for general $a > 0$ that

$$\sqrt{\psi(\gamma)/2} \le \frac{\gamma}{2(1-\gamma/3)} < a^{-1}, \tag{2.5}$$

since $\gamma < 1/(1/3 + a/2) \implies \gamma/(2(1-\gamma/3)) < a^{-1}$. This proves (i).

(ii) is equivalent to

$$\frac{\psi(\gamma)}{\gamma^2} \le \frac{(1 - a\sqrt{\psi(\gamma)/2})^2}{2((1-\gamma/3) - a\gamma/2)^2}.$$

To complete the proof it suffices by (2.4) to show that the right-hand side above is, for fixed γ, a non-decreasing function of $a \in [0, 2(1-\gamma/3)/\gamma)$. Let $b := \sqrt{\psi(\gamma)/2}$, $c := (1-\gamma/3)$ and $d := \gamma/2$, so the expression in question becomes $(1-ab)^2/(2(c-ad)^2)$. Calculus gives

$$\frac{d}{da}\frac{(1-ab)^2}{2(c-ad)^2} = \frac{(1-ab)(d-bc)}{(c-ad)^3}.$$

But $c - ad = 1 - (1/3 + a/2)\gamma > 0$ by assumption. Also $1 - ab > 0$ by (i) and, using (2.5),

$$\begin{aligned} d - bc &= \frac{\gamma}{2} - \sqrt{\psi(\gamma)/2}(1-\gamma/3) \\ &\ge \frac{\gamma}{2} - \frac{\gamma(1-\gamma/3)}{2(1-\gamma/3)} = 0. \end{aligned}$$

The expression $(1-ab)^2/(2(c-ad)^2)$ is therefore non-decreasing in a. □

We finally need an optimization lemma.

Lemma 7. *Let C and b denote two positive real numbers, $t > 0$. Then*

$$\inf_{\beta \in [0,1/b)} \left(-\beta t + \frac{C\beta^2}{1-b\beta} \right) \le \frac{-t^2}{2(2C+bt)}. \tag{2.6}$$

The proof of this lemma can be found in [14] (Lemma 12).

Proposition 3. *Suppose that $f \in \mathcal{A}(\Omega)$ is such that $\forall k$, $f - E_k(f) \le 1$, and that*

$$D\big(\Sigma^2(f)\big) \le a^2 \Sigma^2(f),$$

with $a \ge 0$. Then for all $t > 0$

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

Proof. By a simple limiting argument, we may assume that $a > 0$. Now let $0 < \gamma \leq \beta < 1/(1/3 + a/2)$. By Lemma 6, (i) $\theta := (1/a)\sqrt{2\psi(\gamma)} < 2/a^2$ and also $\theta > \sqrt{\psi(\gamma)/2}\sqrt{2\psi(\gamma)} = \psi(\gamma)$. By Lemma 2,

$$\begin{aligned} S_f(\gamma) &\leq \psi(\gamma) E_{\gamma f}\big[\Sigma^2(f)\big] = \theta^{-1}\psi(\gamma) E_{\gamma f}\big[\theta\Sigma^2(f)\big] \\ &\leq \theta^{-1}\psi(\gamma)\big(S_f(\gamma) + \ln E\big[e^{\theta\Sigma^2(f)}\big]\big), \end{aligned}$$

where the second inequality follows from Lemma 5. Subtracting $\theta^{-1}\psi(\gamma)S_f(\gamma)$, multiplying by θ and using Lemma 4 together with the assumed self-boundedness of $\Sigma^2(f)$ gives us

$$S_f(\gamma)\big(\theta - \psi(\gamma)\big) \leq \psi(\gamma)\ln E\big[e^{\theta\Sigma^2(f)}\big] \leq \frac{\theta\psi(\gamma)}{1 - a^2\theta/2} E\big[\Sigma^2(f)\big],$$

which holds, since $\theta < 2/a^2$. Since $\theta > \psi(\gamma)$ we can divide by $\theta - \psi(\gamma)$ to rearrange and then use the definition of θ to obtain

$$S_f(\gamma) \leq \frac{\psi(\gamma)}{(1 - a\sqrt{\psi(\gamma)/2})^2} E\big[\Sigma^2(f)\big].$$

By Lemma 6(ii) for $\beta < 1/(1/3 + a/2)$

$$\begin{aligned} \int_0^\beta \frac{S_f(\gamma)\,d\gamma}{\gamma^2} &\leq E\big[\Sigma^2(f)\big]\int_0^\beta \frac{\psi(\gamma)}{\gamma^2(1 - a\sqrt{\psi(\gamma)/2})^2}\,d\gamma \\ &\leq E\big[\Sigma^2(f)\big]\int_0^\beta \frac{d\gamma}{2(1 - (1/3 + a/2)\gamma)^2} \\ &= \frac{E[\Sigma^2(f)]}{2}\frac{\beta}{1 - (1/3 + a/2)\beta} \end{aligned}$$

and from Lemma 1

$$\begin{aligned} \Pr\{f - Ef > t\} &\leq \inf_{\beta > 0} \exp\left(\beta\int_0^\beta \frac{S_f(\gamma)}{\gamma^2}\,d\gamma - \beta t\right) \\ &\leq \inf_{\beta \in (0, 1/(1/3 + a/2))} \exp\left(\frac{E[\Sigma^2(f)]}{2}\frac{\beta^2}{1 - (1/3 + a/2)\beta} - \beta t\right) \\ &\leq \exp\left(\frac{-t^2}{2(E[\Sigma^2(f)] + (1/3 + a/2)t)}\right), \end{aligned}$$

where we used Lemma 7 in the last step. □

By Proposition 2, we can substitute $J_\mu(f)$ for a in Proposition 3, which gives Theorem 1 for the case $b = 1$. The general case follows from rescaling and the homogeneity properties of Σ^2 and J_μ.

Of the inequalities in (1.2) only the first one is not completely obvious:

$$\begin{aligned} J_\mu^2(f) &= 4 \sup_{\mathbf{x}\in\Omega} \sum_l \sup_{z\in\Omega_l} \sum_{k:k\neq l} \sigma_k^2\big(f - S_z^l f\big)(\mathbf{x}) \\ &\le 4 \sup_{\mathbf{x}\in\Omega} \sum_l \sup_{z,z'\in\Omega_l} \sum_{k:k\neq l} \sigma_k^2\big(D_{z,z'}^l f\big)(\mathbf{x}) \\ &\le \sup_{\mathbf{x}\in\Omega} \sum_l \sup_{z,z'\in\Omega_l} \sum_{k:k\neq l} \sup_{y,y'} \big(D_{y,y'}^k D_{z,z'}^l f\big)^2(\mathbf{x}) \le J^2(f). \end{aligned}$$

In the last inequality, we used the fact that the variance of a random variable is bounded by a quarter of the square of its range, so that $\sigma_k^2(f) \le (1/4)\sup_{y,y'}(D_{y,y'}^k f)^2$ for all $f \in \mathcal{A}(\Omega)$.

2.4. The bias in the Efron–Stein inequality

Since the published work of Houdré ([11]) assumes symmetric functions and identically distributed variables, we give an independent derivation of Proposition 1 (the author is grateful to Christian Houdré, who supplied a general version in private communication).

Let $X_1, \dots, X_n$ be independent variables with X_i distributed as μ_i in Ω_i, and let $X_1', \dots, X_n'$ be independent copies thereof. For $A \subseteq \{1, \dots, n\}$ denote with X^A the vector

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

and for $k, l \in \mathbb{N}$, use $[k, l]$ to denote the interval $[k, l] = \{k, \dots, l\}$ if $k \le l$ or the empty set otherwise. Also write $X = X^\varnothing$ and $X^\setminus = X^{[1,n]}$ and $X^{\setminus k} = X^{[1,n]\setminus k}$.

Let $f : \prod \Omega_i \to \mathbb{R}$ satisfy $E[f] = 0$. Then, writing $f(X) - f(X')$ as a telescopic series, we get

$$\begin{aligned} \sigma^2(f) &= E\big[f(X)\big(f(X) - f\big(X'\big)\big)\big] \\ &= \sum_{k=1}^n E\big[f(X)\big(f\big(X^{[1,k-1]}\big) - f\big(X^{[1,k]}\big)\big)\big] \\ &= -\sum_{k=1}^n E\big[f\big(X^{\{k\}}\big)\big(f\big(X^{[1,k-1]}\big) - f\big(X^{[1,k]}\big)\big)\big], \end{aligned}$$

where the last identity is obtained by exchanging X_k and X_k'. This gives the variance formula

$$\sigma^2(f) = \frac{1}{2}\sum_{k=1}^n E\big[\big(f(X) - f\big(X^{\{k\}}\big)\big)\big(f\big(X^{[1,k-1]}\big) - f\big(X^{[1,k]}\big)\big)\big] \tag{2.7}$$

(apparently due to Chatterjee). The Cauchy–Schwarz inequality then gives the Efron–Stein inequality

$$\sigma^2(f) \le E\big[\Sigma^2(f)\big] = \frac{1}{2}\sum_{k=1}^{n} E\big[\big(f(X) - f\big(X^{\{k\}}\big)\big)^2\big] = \sum_{k=1}^{n} E\big[\sigma^2\big(f(X)|X^{\setminus k}\big)\big]. \tag{2.8}$$

Now we look at the bias in this inequality.

Theorem 2. *With above conventions we have*

$$\sum_{k=1}^{n} \sigma^2\big[E\big[f(X)|X_k\big]\big] \le \sigma^2\big[f(X)\big] \le \sum_{k=1}^{n} E\big[\sigma^2\big[f(X)|X^{\setminus k}\big]\big]$$
$$\le \sum_{k=1}^{n} \sigma^2\big[E\big[f(X)|X_k\big]\big] + \frac{H^2(f)}{4}.$$

These inequalities simultaneously bound the bias of both upper and lower variance-estimators by $H^2(f)/4$.

Lemma 8.

$$\sum_{k=1}^{n} \sigma^2\big[E\big[f(X)|X_k\big]\big] \le \sigma^2\big[f(X)\big].$$

Proof. By induction on n. Recall the total variance formula

$$\sigma^2(Z) = \sigma^2\big[E[Z|X]\big] + E\big[\sigma^2[Z|X]\big].$$

With $f(X) = Z$ this gives the case $n = 1$. For $n = 2$, we get

$$\begin{aligned} 2\sigma^2\big[f(X)\big] &= \sigma^2\big[E\big[f(X)|X_1\big]\big] + \sigma^2\big[E\big[f(X)|X_2\big]\big] \\ &\quad + E\big[\sigma^2\big[f(X)|X_1\big]\big] + E\big[\sigma^2\big[f(X)|X_2\big]\big] \\ &\ge \sigma^2\big[E\big[f(X)|X_1\big]\big] + \sigma^2\big[E\big[f(X)|X_2\big]\big] + \sigma^2\big[f(X)\big], \end{aligned}$$

where we used the Efron–Stein inequality (2.8). This is where independence comes in and gives us the case $n = 2$. But if the lemma holds for $n - 1$, then

$$\begin{aligned} \sum_{k=1}^{n} \sigma^2\big[E\big[f(X)|X_k\big]\big] &= \sum_{k=1}^{n-1} \sigma^2\big[E\big[f(X)|X_k\big]\big] + \sigma^2\big[E\big[f(X)|X_n\big]\big] \\ &\le \sigma^2\big[E\big[f(X)|X_1, \ldots, X_{n-1}\big]\big] + \sigma^2\big[E\big[f(X)|X_n\big]\big] \\ &\le \sigma^2\big[E\big[f(X)\big]\big], \end{aligned}$$

where the first inequality follows from the induction hypothesis, and the second inequality follows from applying the case $n = 2$ to the two random variables $(X_1, \dots, X_{n-1})$ and X_n. □

Proof of Theorem 2. The first inequality is Lemma 8, the second is the Efron–Stein inequality (2.8). To prove the last inequality, it suffices to show that for each $k \in \{1, \dots, n\}$

$$\begin{aligned} &E\big[\sigma^2\big[f(X)|X^{\setminus k}\big]\big] - \sigma^2\big[E\big[f(X)|X_k\big]\big] \\ &\quad\le \frac{1}{4}\sum_{i:i\neq k} E\big[\big(f(X) - f\big(X^{\{i\}}\big) - f\big(X^{\{k\}}\big) + f\big(X^{\{i\}\cup\{k\}}\big)\big)^2\big], \end{aligned}$$

because summing the R.H.S. over k gives $H^2(f)/4$. We first rewrite the left-hand side as

$$\begin{aligned} &E\big[\sigma^2\big[f(X)|X^{\setminus k}\big]\big] - \sigma^2\big[E\big[f(X)|X_k\big]\big] \\ &\quad= E\big[f(X)\big(f(X) - f\big(X^{\{k\}}\big) - \big(f\big(X^{[1,n]\setminus\{k\}}\big) - f\big(X^{[1,n]}\big)\big)\big)\big] \\ &\quad= E\big[f(X)\big(f\big(X^{\varnothing}\big) - f\big(X^{[1,n]\setminus\{k\}}\big) - \big(f\big(X^{\{k\}}\big) - f\big(X^{[1,n]}\big)\big)\big)\big]. \end{aligned} \tag{2.9}$$

Let π be a permutation of $\{1, \dots, n\}$ such that $\pi(n) = k$ (to get the idea of what follows, it helps to first think of the case $k = n$, with π being the identity map). With π we can write the telescopic expansions

$$f\big(X^{\varnothing}\big) - f\big(X^{[1,n]\setminus\{k\}}\big) = \sum_{i=1}^{n-1}\big(f\big(X^{\pi[1,i-1]}\big) - f\big(X^{\pi[1,i]}\big)\big)$$

and

$$f\big(X^{\{k\}}\big) - f\big(X^{[1,n]}\big) = \sum_{i=1}^{n-1}\big(f\big(X^{\pi[1,i-1]\cup\pi\{n\}}\big) - f\big(X^{\pi[1,i]\cup\pi\{n\}}\big)\big),$$

so that, with $\pi(n) = k$

$$\begin{aligned} &f\big(X^{\varnothing}\big) - f\big(X^{[1,n]\setminus\{k\}}\big) - \big(f\big(X^{\{k\}}\big) - f\big(X^{[1,n]}\big)\big) \\ &\quad= \sum_{i=1}^{n-1}\big(f\big(X^{\pi[1,i-1]}\big) - f\big(X^{\pi[1,i]}\big) - f\big(X^{\pi[1,i-1]\cup\{k\}}\big) + f\big(X^{\pi[1,i]\cup\{k\}}\big)\big) \\ &\quad\triangleq \sum_{i=1}^{n} A_{k,i}\big(X, X'\big). \end{aligned}$$

Each summand $A_{k,i}(X, X')$ changes sign when $X_{\pi(i)}$ and $X'_{\pi(i)}$ are interchanged. Likewise it changes sign if X_k and X'_k are interchanged. But both operations leave the expression in (2.9)

unchanged. Thus,

$$
\begin{aligned}
&E\big[\sigma^2\big[f(X)|X^{\setminus k}\big]\big] - \sigma^2\big[E\big[f(X)|X_k\big]\big] \\
&\quad= \sum_{i=1}^{n-1} E\big[f(X)A_{k,i}\big(X, X'\big)\big] \\
&\quad= \frac{1}{2}\sum_{i=1}^{n-1} E\big[\big(f(X) - f\big(X^{\{i\}}\big)\big)A_{k,i}\big(X, X'\big)\big] \\
&\quad= \frac{1}{4}\sum_{i=1}^{n-1} E\big[\big(f(X) - f\big(X^{\pi\{i\}}\big) - f\big(X^{\{k\}}\big) + f\big(X^{\pi\{i\}\cup\{k\}}\big)\big)A_{k,i}\big(X, X'\big)\big] \\
&\quad\le \frac{1}{4}\sum_{i=1}^{n-1} E\big[\big(f(X) - f\big(X^{\pi\{i\}}\big) - f\big(X^{\{k\}}\big) + f\big(X^{\pi\{i\}\cup\{k\}}\big)\big)^2\big]^{1/2} \\
&\qquad\times E\big[A_{k,i}\big(X, X'\big)^2\big]^{1/2} \\
&\quad= \frac{1}{4}\sum_{i:i\neq k} E\big[\big(f(X) - f\big(X^{\{i\}}\big) - f\big(X^{\{k\}}\big) + f\big(X^{\{i\}\cup\{k\}}\big)\big)^2\big],
\end{aligned}
$$

where the inequality follows from Cauchy–Schwarz, and the last identity from replacing all the $X'_{\pi\{j\}}$ for $j < i$ by $X_{\pi\{j\}}$ in $E[A_{k,i}(X, X')^2]$. □

Since $H(f) \le J(f)$, Proposition 1 is an immediate consequence of Theorem 2.

3. Application to U-statistics

Throughout this section let $(\mathcal{X}, \mathcal{T}, \mu_0)$ be a probability space, $n \in \mathbb{N}$ and $\mu = \mu_0^n$ on $\mathcal{X}^n$. Let g be a measurable, symmetric (permutation invariant) kernel $g : \mathcal{X}^m \to [-1, 1]$ with $1 < m < n$, and let $u_n \in \mathcal{A}(\mathcal{X}^n)$ be defined by

$$
u_n(\mathbf{x}) = \binom{n}{m}^{-1} \sum_{1\le j_1<\cdots<j_m\le n} g(x_{j_1}, \ldots, x_{j_m}).
$$

First we introduce some notation. If B is a set and $m \in \mathbb{N}$, then let $\mathcal{S}_B^m$ denote the set of all those subsets of B which have cardinality m. Also, if $S \subseteq \{1, \ldots, n\}$ and $x \in \mathcal{X}^n$, we use x_S to denote the vector $(x_{j_1}, \ldots, x_{j_{|S|}}) \in \mathcal{X}^{|S|}$, where $\{j_1, \ldots, j_{|S|}\} = S$ and the j_k are increasingly ordered. For $y, z \in \mathcal{X}$ we use (y, x_S) and (y, z, x_S) to denote respectively, the vectors $(y, x_{j_1}, \ldots, x_{j_{|S|}}) \in$

$\mathcal{X}^{|S|+1}$ and $(y, z, x_{j_1}, \dots, x_{j_{|S|}}) \in \mathcal{X}^{|S|+2}$. With this notation

$$u_n(\mathbf{x}) = \binom{n}{m}^{-1} \sum_{S \in \mathcal{S}^m_{\{1,\dots,n\}}} g(x_S).$$

Lemma 9. (i) $u_n - E_k u_n \le 2m/n$, *and* (ii) $J(u_n) \le 4m^2/n$.

Proof. (i) With reference to any given $k \in \{1, \dots, n\}$, using the symmetry of g, we have

$$u_n(\mathbf{x}) = \binom{n}{m}^{-1} \sum_{S \in \mathcal{S}^{m-1}_{\{1,\dots,n\}\setminus k}} g(x_k, x_S) + \binom{n}{m}^{-1} \sum_{S \in \mathcal{S}^m_{\{1,\dots,n\}} : k \notin S} g(x_S).$$

This gives

$$u_n(\mathbf{x}) - E_k u_n(\mathbf{x}) = \binom{n}{m}^{-1} \sum_{S \in \mathcal{S}^{m-1}_{\{1,\dots,n\}\setminus k}} \big(g(x_k, x_S) - E_{y\sim\mu_k}\big[g(y, x_S)\big]\big) \le 2m/n,$$

because g takes values in an interval of diameter 2.

(ii) For $k \ne l$, $y, y' \in \Omega_k$ and $z, z' \in \Omega_l$ we get

$$D^k_{y,y'} u_n(\mathbf{x}) = \binom{n}{m}^{-1} \sum_{S \in \mathcal{S}^{m-1}_{\{1,\dots,n\}\setminus k}} \big(g(y, x_S) - g\big(y', x_S\big)\big)$$

and

$$\begin{aligned} \big|D^l_{z,z'} D^k_{y,y'} u_n(\mathbf{x})\big| &\le \binom{n}{m}^{-1} \sum_{S \in \mathcal{S}^{m-2}_{\{1,\dots,n\}\setminus\{k,l\}}} \big|\big(g(y, z, x_S) - g\big(y', z, x_S\big)\big) \\ &\qquad - \big(g\big(y, z', x_S\big) - g\big(y', z', x_S\big)\big)\big| \\ &\le 4 \frac{\binom{n-2}{m-2}}{\binom{n}{m}} = 4 \frac{m(m-1)}{n(n-1)}. \end{aligned}$$

It follows that

$$\begin{aligned} J(u_n) &\le \left(\sup_{\mathbf{x}\in\Omega} \sum_{k,l:k\ne l} \sup_{y,y',z,z'} \big(D^l_{z,z'} D^k_{y,y'} u_n(\mathbf{x})\big)^2\right)^{1/2} \\ &\le \frac{4m(m-1)}{\sqrt{n(n-1)}} \le \frac{4m^2}{n}. \end{aligned}$$

□

Substitution of this lemma in Corollary 1 gives for any $t > 0$ the Bernstein-type inequality

$$\Pr\{u_n - E[u_n] > t\} \leq \exp\left(\frac{-t^2}{2\sigma^2(u_n) + H^2(u_n)/2 + (\frac{4m}{3n} + \frac{4m^2}{n})t}\right).$$

Since $H^2(u_n) \leq J^2(u_n) = 16m^2/n \to 0$ as $n \to \infty$ this bound is asymptotically sharp by the central limit theorem for U-statistics [10]. Working directly from Theorem 1 instead of Corollary 1 one can also obtain the following inequality, which explicitly contains the variance of the limit distribution (proof omitted).

$$\Pr\{u_n - Eu_n > t\} \leq \exp\left(\frac{-nt^2}{2m^2\sigma^2_{y\sim\mu_0}(E_{\mathbf{x}\sim\mu_0^{m-1}}[g(y,\mathbf{x})]) + \frac{m^2(m-1)^2}{n-m} + 16m^2t/3}\right).$$

The elementary estimates of Lemma 9 can similarly be made when the kernel is not symmetric, or when there is a different bounded kernel $g_{\mathbf{j}}$ for every m-tuple $\mathbf{j} = (j_1, \ldots, j_m)$ of indices. Similar estimates hold for V-statistics with bounded kernels [19]. Observe also that above some improvement is possible by bounding J_μ instead of J.

To put the above result in perspective, we consider the classical work of [10], and more recent results of [2,9,12] and [1]. Hoeffding [10] and Arcones [2] consider undecoupled, nondegenerate U-statistics of arbitrary order and are therefore directly comparable to the above result. Hoeffding [10] does not have the correct variance term, while [2] gives the correct variance term but severely overestimates the scale-proxy in Bernstein's inequality to be exponential in the degree m of the U-statistics (above it is only of order m^2). This problem results from the use of the decoupling inequalities of [7] and seems to beset most works on U-statistics of higher order. Neither [10] nor [2] nor our inequality can take advantage of degeneracy. This is different for [9,12] and [1], which in this respect always improve over our result. Arcones [2], Giné et al. [9] and Adamczak [1] also consider Banach-space valued U-statistics which are inaccessible to our version of Bernstein's inequality, but in the undecoupled case of higher order they all suffer from the exponential dependence on m as introduced by the decoupling inequalities. Houdré and Reynaud-Bouret [12] contains explicit constants and is generally superior to our result (even if we improve our inequality by considering J_μ instead of J), but it is limited to the case $m = 2$. In summary: our inequality seems to be the only one, which applies to undecoupled, non-degenerate U-statistics of arbitrary degree, and both gives the correct variance term and avoids an exponential dependence on the degree. It is unsuited for application to degenerate and Banach-space-valued U-statistics. The simplicity of its derivation is perhaps its greatest merit.

4. Application to ridge regression

Let $\mathbb{B}$ be the unit ball in a separable, real Hilbert space, and let $\mathcal{Z} = \mathbb{B}\times[-1,1]$. Fix $\lambda \in (0,1)$. For $\mathbf{z} = ((x_1,y_1),\ldots,(x_n,y_n)) \in \mathcal{Z}^n$ regularized least squares returns the vector

$$w_{\mathbf{z}} = \arg\min_{w\in H} \frac{1}{n}\sum_{i=1}^{n}\big(\langle w, x_i\rangle - y_i\big)^2 + \lambda\|w\|^2.$$

Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ be a vector of independent random variables with values in $\mathcal{Z}$, where Z_i is identically distributed to $Z = (X, Y)$. We can apply Theorem 1, to obtain tail-bounds for the random variable $\Delta(\mathbf{Z}) = R(\mathbf{Z}) - \hat{R}(\mathbf{Z})$, where the "true error" R and the "empirical error" $\hat{R}$ are defined on $\mathcal{Z}^n$ by

$$R(\mathbf{z}) = E_Z\big(\langle w_{\mathbf{z}}, X\rangle - Y\big)^2 \quad \text{and} \quad \hat{R}(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n}\big(\langle w_{\mathbf{z}}, x_i\rangle - y_i\big)^2,$$

and the random variable $\Delta(\mathbf{Z})$ is the "generalization error". We can prove the following result.

Theorem 3. *There is an absolute constant c such that for every $\delta \in (0, 1/e)$ with probability at least $1 - \delta$ in $\mathbf{Z}$*

$$\Delta(\mathbf{Z}) \le E[\Delta] + \sqrt{2\sigma^2(\Delta)\ln(1/\delta)} + \frac{c\lambda^{-3}\ln(1/\delta)}{n}.$$

It can be shown ([6]) that the expectation $E[\Delta]$ is of order $1/n$, so for large sample sizes the generalization error $\Delta(\mathbf{Z})$ is dominated by the variance term. This application of Theorem 1 has now been extended to other loss functions, see [16] for more details.

A major drawback is the dependence on λ^{-3} in the last term, because in practical applications the regularization parameter λ typically decreases with n. The λ^{-3} is likely due to a very crude method of bounding $J(f)$ by differentiation. A more intelligent method might give $\lambda^{-2}n^{-1}$.

With λ fixed the inequality in Theorem 3 gives optimal rates in n up to factor $\ln n$. Suppose the distribution of Z is such, that for some constant c and $p \in [1/2, 1]$ and all $\delta > 0$ with probability at most $1 - \delta$ we have $\Delta \le cn^{-p}\ln(1/\delta)$. Then, since Δ can be shown to be bounded and letting $\delta = n^{-2p}$, it easily follows that $\sigma^2(\Delta) \le Cn^{-2p}\ln(n)$ for some other constant C, so that Theorem 3 gives a rate of $O(n^{-p}\ln(n))$.

The key to the application of Theorem 1 is the following lemma ($\mathcal{L}^+(H)$ denoting the cone of nonnegative definite operators in H).

Lemma 10. *Let $G : (0,1)^2 \to \mathcal{L}^+(H)$ and $g : (0,1)^2 \to H$ be both twice continuously differentiable, satisfying the conditions $\frac{\partial^2}{\partial s\partial t}G = 0$, $\frac{\partial^2}{\partial s\partial t}g = 0$, $\|\frac{\partial}{\partial t}G\| \le B_1$, $\|\frac{\partial}{\partial s}G\| \le B_1$, $\|\frac{\partial}{\partial t}g\| \le B_2$ and $\|\frac{\partial}{\partial s}g\| \le B_2$ for real numbers B_1 and B_2. For $\lambda > 0$ define a function $w : (0,1)^2 \to H$ by $w = (G + \lambda)^{-1}g$. Then w is twice differentiable and*

$$\left\|\frac{\partial}{\partial t}w\right\| \le \lambda^{-1}\big(B_1\|w\| + B_2\big), \tag{4.1}$$

$$\left\|\frac{\partial^2}{\partial s\,\partial t}w\right\| \le 2\lambda^{-2}\big(B_1^2\|w\| + B_1B_2\big). \tag{4.2}$$

Proof. To shorten expressions, write $T := (G + \lambda)^{-1}$. A standard argument shows that $\|T\| \le \lambda^{-1}$ (we use $\|\cdot\|$ for the operator norm and for vectors in H, depending on context) and that

$$\frac{\partial}{\partial t}T = -T\left(\frac{\partial}{\partial t}G\right)T, \qquad \text{so that } \left\|\frac{\partial}{\partial t}T\right\| \le \lambda^{-2}B_1.$$

Then

$$\frac{\partial}{\partial t}w = -T\left(\frac{\partial}{\partial t}G\right)w + T\frac{\partial}{\partial t}g.$$

This gives (4.1). Also, using the fact that the mixed partials vanish by assumption,

$$\begin{aligned}\frac{\partial^2}{\partial s\,\partial t}w &= \frac{\partial}{\partial s}\left[-T\left(\frac{\partial}{\partial t}G\right)Tg + T\frac{\partial}{\partial t}g\right]\\ &= T\left(\frac{\partial}{\partial s}G\right)T\left(\left(\frac{\partial}{\partial t}G\right)w - \frac{\partial}{\partial t}g\right) + T\left(\frac{\partial}{\partial t}G\right)T\left(\left(\frac{\partial}{\partial s}G\right)w - \frac{\partial}{\partial s}g\right),\end{aligned}$$

which gives (4.2). □

Proof of Theorem 3. It is well known and easily verified that $w_{\mathbf{z}}$ is well defined and explicitly given by the formula

$$w_{\mathbf{z}} = (G_{\mathbf{z}} + \lambda)^{-1} g_{\mathbf{z}},$$

where the positive semidefinite operator $G_{\mathbf{z}}$ and the vector $g_{\mathbf{z}} = g$ are given by

$$G_{\mathbf{z}}v = \frac{1}{n}\sum_{i=1}^{n}\langle v, x_i\rangle x_i \quad \text{and} \quad g_{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i.$$

Also we have

$$\frac{1}{n}\sum_{i=1}^{n}\big(\langle w_{\mathbf{z}}, x_i\rangle - y_i\big)^2 + \lambda\|w_{\mathbf{z}}\|^2 \le \frac{1}{n}\sum_{i=1}^{n}\big(\langle 0, x_i\rangle - y_i\big)^2 + \lambda\|0\|^2 \le 1,$$

from which we retain that $\sum(\langle w_{\mathbf{z}}, x_i\rangle - y_i)^2 \le n$ and $\|w_{\mathbf{z}}\| \le \lambda^{-1/2}$.

Now consider any sample $\mathbf{z} \in \mathcal{Z}^n$ and fix two indices $1 \le k, l \le n$ with $k \ne l$, and $z_l' = (x_l', y_l'), z_k' = (x_k', y_k'), z_l'' = (x_l'', y_l'') \in \mathcal{Z}$ and $z_k'' = (x_k'', y_k'') \in \mathcal{Z}$. For $(s,t) \in (0,1)^2$ we consider the behavior of ridge regression on the doubly modified sample $\mathbf{z}(s,t) := S^l_{z_l'+s(z_l''-z_l')}S^k_{z_k'+t(z_k''-z_k')}\mathbf{z}$ ($\mathcal{Z}$ is a convex subset of $H \times \mathbb{R}$). We write

$$G(s,t) := G_{\mathbf{z}(s,t)} \quad \text{and} \quad g(s,t) := g_{\mathbf{z}(s,t)} \quad \text{and} \quad w(s,t) := w_{\mathbf{z}(s,t)} = \big(G(s,t) + \lambda\big)^{-1} g(s,t).$$

Then

$$\begin{aligned}\left\|\left(\frac{\partial}{\partial t}G\right)v\right\| &= \frac{1}{n}\left\|\frac{\partial}{\partial t}\big\langle v, x_k' + t\big(x_k'' - x_k'\big)\big\rangle\big(x_k' + t\big(x_k'' - x_k'\big)\big)\right\|\\ &= \frac{1}{n}\left\|\big\langle v, x_k'' - x_k'\big\rangle\big(x_l' + t\big(x_k'' - x_k'\big)\big) + \big\langle v, x_k' + t\big(x_k'' - x_k'\big)\big\rangle\big(x_k'' - x_k'\big)\right\|\\ &\le \frac{2}{n}\|v\|\left\|x_k'' - x_k'\right\|\left\|x_l' + t\big(x_k'' - x_k'\big)\right\| \le \frac{4}{n}\|v\|,\end{aligned}$$

because $\|x_k'' - x_k'\| \le 2$ and $\|x_l' + t(x_k'' - x_k')\| \le 1$. Thus, $\|(\partial/\partial t)G\| \le 4/n$ and similarly $\|(\partial/\partial s)G\| \le 4/n$. Since $k \ne l$ it is clear that $(\partial^2/(\partial s \partial t))G = 0$. Also

$$\begin{aligned}\left\|\frac{\partial}{\partial t}g\right\| &= \frac{1}{n}\left\|\frac{\partial}{\partial t}\big(\big(y_k' + t\big(y_k'' - y_k'\big)\big)\big(x_k' + t\big(x_k'' - x_k'\big)\big)\big)\right\| \\ &\le \frac{1}{n}\big(\big|y_k'' - y_k'\big|\big\|x_k' + t\big(x_k'' - x_k'\big)\big\| + \big|y_k' + t\big(y_k'' - y_k'\big)\big|\big\|x_k'' - x_k'\big\|\big) \\ &\le \frac{4}{n},\end{aligned}$$

similarly $\|(\partial/\partial s)g\| \le 4/n$ and again $(\partial^2/(\partial s \partial t))g = 0$. We can then apply Lemma (10) and obtain (remembering $0 < \lambda \le 1$)

$$\left\|\frac{\partial}{\partial t}w\right\| \le \frac{4}{n}\lambda^{-1}\big(\lambda^{-1/2} + 1\big) \le \frac{8\lambda^{-3/2}}{n}$$

and

$$\left\|\frac{\partial^2}{\partial s\, l \partial t}w\right\| \le \frac{8}{n^2}\lambda^{-2}\big(\lambda^{-1/2} + 1\big) \le \frac{32\lambda^{-5/2}}{n^2},$$

where we used $\|w\| \le \lambda^{-1/2}$.

Now we define

$$\begin{aligned}R(s,t) &= E\big[\big(\big\langle w(s,t), X\big\rangle - Y\big)^2\big], \\ \hat{R}(s,t) &= \frac{1}{n}\sum_i \big(\big\langle w(s,t), x_i(s,t)\big\rangle - y_i(s,t)\big)^2.\end{aligned}$$

For the expected error, we get

$$\begin{aligned}\left|\frac{\partial}{\partial t}R(s,t)\right| &\le 2E\left|\big(\big\langle w(s,t), X\big\rangle - Y\big)\left\langle\frac{\partial}{\partial t}w(s,t), X\right\rangle\right| \\ &\le \big(\lambda^{-1/2} + 1\big)\frac{8\lambda^{-3/2}}{n} \le \frac{16\lambda^{-2}}{n}\end{aligned}$$

and

$$\begin{aligned}\left|\frac{\partial^2}{\partial s\, \partial t}R(s,t)\right| &\le 2E\left|\frac{\partial}{\partial s}\left(\big(\big\langle w(s,t), X\big\rangle - Y\big)\left\langle\frac{\partial}{\partial t}w(s,t), X\right\rangle\right)\right| \\ &\le 2E\left|\left\langle\frac{\partial}{\partial s}w(s,t), X\right\rangle\left\langle\frac{\partial}{\partial t}w(s,t), X\right\rangle\right| \\ &\quad + 2E\left|\big(\big\langle w(s,t), X\big\rangle - Y\big)\left\langle\frac{\partial^2}{\partial s\, \partial t}w(s,t), X\right\rangle\right| \\ &\le \frac{256}{n^2}\lambda^{-3}.\end{aligned}$$

By a similar (and more tedious) analysis there are absolute constants c_1 and c_2, such that

$$\left|\frac{\partial}{\partial t}\big(R(s,t)-\hat{R}(s,t)\big)\right| \le \frac{c_1\lambda^{-2}}{n}$$

and

$$\left|\frac{\partial^2}{\partial s\,\partial t}\big(R(s,t)-\hat{R}(s,t)\big)\right| \le \frac{c_2\lambda^{-3}}{n^2}.$$

Thus,

$$\begin{aligned} D^k_{z'_k,z''_k}\Delta(\mathbf{z}) &= \int_0^1 \frac{\partial}{\partial t} S^k_{z'_k+t(z''_k-z'_k)}\Delta(\mathbf{z})\,dt \\ &\le \int_0^1 \left|\frac{\partial}{\partial t}\big(R(s,t)-\hat{R}(s,t)\big)\right| dt \le \frac{c_1\lambda^{-2}}{n}. \end{aligned}$$

In particular, $\Delta - E_k[\Delta] \le c_1\lambda^{-2}/n = b$. Also

$$\begin{aligned} D^l_{z'_l,z''_l} D^k_{z'_k,z''_k}\Delta(\mathbf{z}) &= \int_0^1\int_0^1 \frac{\partial^2}{\partial s\,\partial t} S^l_{z'_l+s(z''_l-z'_l)} S^k_{z'_k+t(z''_k-z'_k)}\Delta(\mathbf{z})\,dt\,ds \\ &\le \frac{n\lambda^2}{c_1}\int_0^1\int_0^1 \left|\frac{\partial^2}{\partial s\,\partial t}\big(R(s,t)-\hat{R}(s,t)\big)\right| dt\,ds \le \frac{c_2\lambda^{-3}}{n^2}. \end{aligned}$$

Substitution in the formula gives $J(f) \le c_2\lambda^{-3}/n$. Corollary 1 and the elementary bound $H(f) \le J(f)$ give

$$\Pr\big\{\Delta - E[\Delta] > t\big\} \le \exp\left(\frac{-t^2}{2\sigma^2(\Delta) + J^2(\Delta)/2 + (2b/3 + J(\Delta))t}\right).$$

Equating the probability to $\delta \in (0, 1/e)$, solving for the deviation gives with probability at least $1-\delta$ that

$$\Delta - E[\Delta] \le \sqrt{2\sigma^2(\Delta)\ln(1/\delta)} + \big(2b/3 + 2J(\Delta)\big)\ln(1/\delta),$$

and substitution of $b = c_1\lambda^{-2}/n$ and $J(f) \le c_2\lambda^{-3}/n$ give the result. □

References

[1] Adamczak, R. (2006). Moment inequalities for U-statistics. *Ann. Probab.* **34** 2288–2314. MR2294982

[2] Arcones, M.A. (1995). A Bernstein-type inequality for U-statistics and U-processes. *Statist. Probab. Lett.* **22** 239–247. MR1323145

[3] Bernstein, S. (1924). On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math* **1** 38-49.

[4] Boucheron, S., Lugosi, G. and Massart, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.* **31** 1583–1614. MR1989444
[5] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. With a foreword by Michel Ledoux. MR3185193
[6] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* **2** 499–526. MR1929416
[7] de la Peña, V.H. (1992). Decoupling and Khintchine's inequalities for U-statistics. *Ann. Probab.* **20** 1877–1892. MR1188046
[8] Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596. MR0615434
[9] Giné, E., Latała, R. and Zinn, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability, II* (*Seattle, WA*, 1999). *Progress in Probability* **47** 13–38. Boston, MA: Birkhäuser. MR1857312
[10] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19** 293–325. MR0026294
[11] Houdré, C. (1997). The iterated jackknife estimate of variance. *Statist. Probab. Lett.* **35** 197–201. MR1483274
[12] Houdré, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic Inequalities and Applications. Progress in Probability* **56** 55–69. Basel: Birkhäuser. MR2073426
[13] Ledoux, M. (2001). *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs* **89**. Providence, RI: Amer. Math. Soc. MR1849347
[14] Maurer, A. (2006). Concentration inequalities for functions of independent variables. *Random Structures Algorithms* **29** 121–138. MR2245497
[15] Maurer, A. (2012). Thermodynamics and concentration. *Bernoulli* **18** 434–454. MR2922456
[16] Maurer, A. (2017). A second-order look at stability and generalization. In *Conference on Learning Theory* 1461–1475.
[17] McDiarmid, C. (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics. Algorithms Combin.* **16** 195–248. Berlin: Springer. MR1678578
[18] Steele, J.M. (1986). An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758. MR0840528
[19] v. Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Stat.* **18** 309–348. MR0022330

Received February 2017 and revised January 2018