Proceedings of Machine Learning Research vol 99:1–16, 2019 32nd Annual Conference on Learning Theory

Uniform Concentration and Symmetrization for Weak Interactions

Andreas Maurer AM@ANDREAS-MAURER.EU
Adalbertstr. 55, 80799 Munich, Germany.

Massimiliano Pontil M.PONTIL@UCL.AC.UK
Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, 16100 Genova, Italy
Department of Computer Science, University College London, WC1E 6BT London, UK.

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

The method to derive uniform bounds with Gaussian and Rademacher complexities is extended to the case where the sample average is replaced by a nonlinear statistic. Tight bounds are obtained for U-statistics, smoothened L-statistics and error functionals of ℓ_2-regularized algorithms.

1. Introduction

The purpose of this paper is to extend the method of Rademacher or Gaussian complexities to a more general, nonlinear setting. Suppose that $\mathbf{X} = (X_1, ..., X_n)$ is a vector of independent random variables with values in some space $\mathcal{X}$, $\mathbf{X}'$ is iid to $\mathbf{X}$, and that $\mathcal{H}$ is a finite class of functions $h : \mathcal{X} \to [0, 1]$. For $\mathbf{x} \in \mathcal{X}^n$ and $h \in \mathcal{H}$ we use $h(\mathbf{x})$ to denote the vector $h(\mathbf{x}) = (h(x_1), ..., h(x_n)) \in [0, 1]^n$ and $\mathcal{H}(\mathbf{x}) = \{h(\mathbf{x}) : h \in \mathcal{H}\} \subseteq \mathbb{R}^n$. Now let $f : [0, 1]^n \to \mathbb{R}$ be the sample average

$$f(s_1, ..., s_n) := \frac{1}{n} \sum_{i=1}^{n} s_i \text{ for } s_i \in \mathbb{R}.$$

Then it is not hard to show (see Bartlett and Mendelson (2002), Theorem 8, or Ledoux (1991), Lemma 6.3 and (4.8)) that

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{X}'}\left[f\left(h\left(\mathbf{X}'\right)\right)\right] - f\left(h\left(\mathbf{X}\right)\right)\right] \leq \frac{2}{n}\mathbb{E}\left[R\left(\mathcal{H}\left(\mathbf{X}\right)\right)\right] \leq \frac{\sqrt{2\pi}}{n}\mathbb{E}\left[G\left(\mathcal{H}\left(\mathbf{X}\right)\right)\right], \tag{1}$$

where the Rademacher and Gaussian averages of a subset $Y \subseteq \mathbb{R}^n$ are

$$R(Y) = \mathbb{E}\sup_{\mathbf{y}\in Y}\langle\boldsymbol{\epsilon}, \mathbf{y}\rangle \text{ and } G(Y) = \mathbb{E}\sup_{\mathbf{y}\in Y}\langle\boldsymbol{\gamma}, \mathbf{y}\rangle.$$

Here $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ and $\gamma = (\gamma_1, ..., \gamma_n)$ are vectors of independent Rademacher and standard normal variables respectively.

The bounded difference inequality (Theorem 11, often called McDiarmid's inequality) shows that the random variable $\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{X}'}\left[f\left(h\left(\mathbf{X}'\right)\right)\right] - f\left(h\left(\mathbf{X}\right)\right)$ is sharply concentrated about its mean, and the symmetrization inequalities (1) lead to a uniform bound on the estimation error (see Koltchinskii (2002) or Bartlett and Mendelson (2002)): for any $\delta \in (0, 1)$ with probability at least $1 - \delta$

$$\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{X}'}\left[f\left(h\left(\mathbf{X}'\right)\right)\right] - f\left(h\left(\mathbf{X}\right)\right) \leq \frac{2}{n}\mathbb{E}\left[R\left(\mathcal{H}\left(\mathbf{X}\right)\right)\right] + \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{2}$$

© 2019 A. Maurer & M. Pontil.

This fact has proven very useful in statistical learning theory, and many techniques have been developed to bound Rademacher and Gaussian averages in various contexts of classification, function learning, matrix completion, multi-task learning and unsupervised learning (see e.g. Bartlett and Mendelson (2002), Meir and Zhang (2003), Ambroladze et al. (2007), Kakade et al. (2009), Kakade et al. (2012), Biau et al. (2008)).

The sample average is particularly simple and useful, but there are many other interesting statistics, which are nonlinear, such as U-statistics, quantiles, or M-estimators to estimate other distributional properties. Concrete examples would be estimators of the median for economic applications, or the Wilcoxon two-sample statistic, which plays a role in the evaluation of ranking functions (Agarwal et al. (2005)). Nonlinear versions of (1) and (2) could be quite useful and make the abundance of techniques to bound Rademacher and Gaussian averages available in a larger context.

In this paper we show that such an extension is possible, also for vector valued function classes, if the statistic f in question has the right kind of Lipschitz property and is not too "far from linearity". To make this precise we make the following definition.

Definition 1 *Let $\mathcal{X}$ be any set. For $k \in \{1, ..., n\}$ and $y, y' \in \mathcal{X}$, define the k-th partial difference operator on the space of functions $f : \mathcal{X}^n \to \mathbb{R}$ as*

$$D^k_{yy'} f(\mathbf{x}) = f(..., x_{k-1}, y, x_{k+1}, ...) - f\left(..., x_{k-1}, y', x_{k+1}, ...\right), \text{ for } \mathbf{x} \in \mathcal{X}^n.$$

For $\mathcal{U} \subseteq \mathbb{R}^d$ we define seminorms M_{Lip} and J_{Lip} on the vector space of functions $f : \mathcal{U}^n \to \mathbb{R}$ by

$$\begin{aligned} M_{Lip}(f) &= \max_k \sup_{\mathbf{x} \in \mathcal{U}^n, y \neq y' \in \mathcal{U}} \frac{D^k_{yy'} f(\mathbf{x})}{\|y - y'\|} \text{ and} \\ J_{Lip}(f) &= n \max_{k \neq l} \sup_{\mathbf{x} \in \mathcal{U}^n, y \neq y', z, z' \in \mathcal{U}} \frac{D^l_{zz'} D^k_{yy'} f(\mathbf{x})}{\|y - y'\|}. \end{aligned}$$

So M_{Lip} is a coordinatewise Lipschitz-seminorm, while J_{Lip} quantifies nonlinearity in terms of second-order interactions in combination with a Lipschitz property.

With these definitions we can extend the Gaussian part of the symmetrization inequalities (1) to nonlinear statistics.

Theorem 2 *Let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent random variables with values in $\mathcal{X}$, $\mathbf{X}'$ iid to $\mathbf{X}$, let $\mathcal{U} \subseteq \mathbb{R}^d$, let $\mathcal{H}$ be a finite class of functions $h : \mathcal{X} \to \mathcal{U}$ and let $\mathcal{H}(\mathbf{X}) = \{h(\mathbf{X}) : h \in \mathcal{H}\} \subseteq \mathbb{R}^{dn}$. Then for $f : \mathcal{U}^n \to \mathbb{R}$*

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{X}'}\left[f\left(h\left(\mathbf{X}'\right)\right)\right] - f(h(\mathbf{X}))\right] \leq \sqrt{2\pi}\,(2M_{Lip}(f) + J_{Lip}(f))\; \mathbb{E}[G(\mathcal{H}(\mathbf{X}))]. \quad (3)$$

Remarks:

1. If $d = 1$ and f is the arithmetic mean, then it is easy to see that $M_{Lip}(f) = 1/n$ and $J_{Lip}(f) = 0$, so the Gaussian version of (1) is recovered up to a constant factor of 2.

2. Since the right hand side of (3) is invariant under a sign-change of f, the same bounds hold for $\sup_{h\in\mathcal{H}} f(h(\mathbf{X})) - \mathbb{E}[f(h(\mathbf{X}'))]$.

3. In many applications the Gaussian average $G(\mathcal{H}(\mathbf{X}))$ can be bounded in the same way as the Rademacher average. In general $G(\mathcal{H}(\mathbf{X}))$ can be bounded by $R(\mathcal{H}(\mathbf{X}))$ with an additional factor of $3\sqrt{\ln(n+1)}$ (see Ledoux (1991), (4.9)).

4. Finite cardinality of $\mathcal{H}$ is required to avoid problems of measurability, and the cardinality of $\mathcal{H}$ can be arbitrarily large. For infinite $\mathcal{H}$ one can replace expressions like $\mathbb{E}[\sup_{h\in\mathcal{H}}(\cdot)]$ by $\sup_{\mathcal{H}_0\subset\mathcal{H},|\mathcal{H}_0|<\infty}\mathbb{E}\left[\sup_{h\in\mathcal{H}_0}(\cdot)\right]$.

For a given statistic f the key to the application of Theorem 2 is the verification that $M_{Lip}(f)$ and $J_{Lip}(f)$ are of order $O(1/n)$. This is true for the sample average, but also for

- U- and V-statistics of all orders with coordinate-wise Lipschitz kernels. This includes multi-sample cases, such as smoothened versions of the Wilcoxon two-sample-statistic. A corresponding application to ranking is sketched in Section 2.1.
- Lipschitz L-statistics. These are weighted averages of order statistics with Lipschitz weighting functions and include smoothened approximations to medians, or smoothened estimators for quantiles. In Section 2.2 a potential application to robust clustering is discussed.
- a class of M-estimators with strongly convex objectives, in particular error functionals of ℓ_2-regularized classification or function estimation. In Section 2.3 we sketch an application to representation learning.

This list is not exhaustive and other examples can be generated using the fact that M_{Lip} and J_{Lip} are seminorms. Also, if $\mathcal{U}\subseteq\mathbb{R}^d$ is bounded and $M_{Lip}(f)$ and $J_{Lip}(f)$ are of order $O(1/n)$, then every twice differentiable function with bounded derivatives when composed with f has the same property (see Maurer and Pontil (2018)).

The seminorms M_{Lip} and J_{Lip} are strongly related to the seminorms M and J introduced in Maurer and Pontil (2018). For $f:\mathcal{X}^n\to\mathbb{R}$, where $\mathcal{X}$ can be any set, they are defined as

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

M and J control the nonlinear generalizations of several properties of linear statistics, such as Bernstein's inequality, sample-efficient variance estimation, empirical Bernstein bounds and Berry-Esseen type bounds of normal approximation (see Maurer (2017) and Maurer and Pontil (2018)). If $\mathcal{U}$ is bounded with diameter Δ, then clearly $M(f)\leq M_{Lip}(f)\Delta$ and $J(f)\leq J_{Lip}(f)\Delta$, and the results in Maurer and Pontil (2018) can be reformulated in terms of M_{Lip} and J_{Lip}. In particular, if $\mathcal{U}$ is bounded and $M_{Lip}(f)$ and $J_{Lip}(f)$ are of order $O(1/n)$, then f is a weakly interactive function as defined in Maurer and Pontil (2018).

Theorem 2, the definition of $M(f)$ and the bounded difference inequality (Theorem 11) applied to the random variable $\sup_{h\in\mathcal{H}}\mathbb{E}[f(h(\mathbf{X}'))] - f(h(\mathbf{X}))$ yield the nonlinear extension of (2).

Corollary 3 *Under the conditions of Theorem 2, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{h\in\mathcal{H}} \mathbb{E}\left[f\left(h\left(\mathbf{X}'\right)\right)\right] - f\left(h\left(\mathbf{X}\right)\right)$$
$$\leq \sqrt{2\pi}\left(2M_{Lip}\left(f\right)+J_{Lip}\left(f\right)\right)\,\mathbb{E}\left[G\left(\mathcal{H}\left(\mathbf{X}'\right)\right)\right] + M\left(f\right)\sqrt{n\ln\left(1/\delta\right)}.$$

The next section is devoted to applications, then we prove Theorem 2. An appendix contains some technical material.

2. Applications

In the sequel we sketch some potential applications and exhibit some generic classes of statistics, to which Theorem 2 and Corollary 3 can be applied.

2.1. Ranking, U- and V-statistics

An example for the application of Theorem 2 is given by the following variant of the Wilcoxon-two-sample statistic, which we simplify for the purpose of illustration. Let n be an even integer, $\ell : \mathbb{R} \to [0,1]$ and define $A_\ell : \mathbb{R}^n \to \mathbb{R}$ by

$$A_\ell\left(x_1,...,x_n\right) = \frac{4}{n^2}\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n}\ell\left(x_i - x_j\right).$$

Now suppose that μ_+ and μ_- are two probability measures on some space $\mathcal{X}$, and we construct a sample $\mathbf{X}=(X_1,...,X_n)$ by drawing the first half of $\mathbf{X}$ iid from μ_+ and the second half iid from μ_-, that is $\mathbf{X}\sim\mu_+^{n/2}\times\mu_-^{n/2}$. Now let $h:\mathcal{X}\to\mathbb{R}$ be some function. If $\ell = 1_{(0,\infty)}$ is the indicator of the positive reals, then $A_\ell\left(h\left(\mathbf{X}\right)\right)$ is evidently an unbiased estimator for

$$\Pr_{(x,y)\sim\mu_+\times\mu_-}\left\{h\left(x\right)>h\left(y\right)\right\}, \tag{4}$$

also called "the area under the ROC Curve" (AUC) (as explained in Agarwal et al. (2005)), and provides a criterion for the evaluation of h as a ranking functions. In this case A_ℓ is the proper Wilcoxon statistic (apart from the fact that we didn't worry about ties and consider a balanced sample for simplicity), but surrogate loss functions ℓ come into play if a good ranking function is to be chosen from a set of candidates (see Ying et al. (2016)).

Let us assume that ℓ has Lipschitz constant L. Applying the partial difference operator to the function A_ℓ, at first for $k\leq n/2$, we find for any $y,y'\in\mathbb{R}$

$$D^k_{y,y'}A_\ell\left(\mathbf{x}\right) = \frac{4}{n^2}\sum_{j=n/2+1}^{n}\ell\left(y-x_j\right)-\ell\left(y'-x_j\right)\leq\frac{2L}{n}\left|y-y'\right|.$$

Together with the analogous argument for $k>n/2$ this gives the bound

$$M_{Lip}\left(A_\ell\right) = \max_k\sup_{\mathbf{x}\in\mathbb{R}^n,y\neq y'\in\mathbb{R}}\frac{D^k_{y,y'}A_\ell\left(\mathbf{x}\right)}{\left|y-y'\right|}\leq\frac{2L}{n}. \tag{5}$$

In the same way one shows that $M(A_\ell) \le 2/n$. To bound $J_{Lip}(A_\ell)$ first let $k \le n/2$, $l \ne k$ and $y, y', z, z' \in \mathbb{R}$. Then

$$D^l_{z,z'} D^k_{y,y'} A_\ell(\mathbf{x}) = \frac{4}{n^2} \sum_{j=n/2+1}^{n} D^l_{z,z'} \left(\ell(y - x_j) - \ell(y' - x_j)\right)$$
$$= \begin{cases} 0 & \text{if} \quad l \le n/2 \\ \le \frac{4}{n^2}\left(\ell(y-z) - \ell(y'-z) - \ell(y-z') + \ell(y'-z')\right) & \text{if} \quad l > n/2 \end{cases} \le \frac{8L}{n^2}\left|y - y'\right|,$$

and analogous reasoning for $k > n/2$ gives

$$J_{Lip}(A_\ell) = n \max_{k,l:k\ne l} \sup_{\mathbf{x}\in\mathbb{R}^n, y,y',z,z'\in\mathbb{R}} \frac{D^l_{z,z'} D^k_{y,y'} A_\ell(\mathbf{x})}{|y - y'|} \le \frac{8L}{n}. \tag{6}$$

Now suppose that $\mathcal{H}$ is a set of candidate ranking functions $h : \mathcal{X} \to \mathbb{R}$, for example a ball of linear functionals in a RKHS. We wish to choose $h \in \mathcal{H}$ so as to maximize (4). If we choose $\ell \le 1_{(0,\infty)}$, then Corollary 3 states that for every $\delta \in (0, 1)$ with probability at least $1 - \delta$ in $\mathbf{X}$ we have for every potential ranking function $h \in \mathcal{H}$ that

$$\begin{aligned} \Pr_{(X,Y)\sim\mu_+\times\mu_-} \{h(X) > h(Y)\} &= \mathbb{E}\left[A_{1_{(0,\infty)}}\left(h\left(\mathbf{X}'\right)\right)\right] \ge \mathbb{E}\left[A_\ell\left(h\left(\mathbf{X}'\right)\right)\right] \\ &\ge A_\ell(h(\mathbf{X})) - \frac{12\sqrt{2\pi}L\,\mathbb{E}\left[G\left(\mathcal{H}\left(\mathbf{X}'\right)\right)\right]}{n} - 2\sqrt{\frac{\ln(1/\delta)}{n}}, \end{aligned}$$

so as to justify the strategy to optimize the AUC by the maximization of the empirical surrogate $A_\ell(h(\mathbf{X}))$. Similar bounds are obtained in Clemencon et al. (2008), even with fast rates under some additional assumptions. The point here is to illustrate the simplicity of only needing to verify the first- and second-order response properties (5) and (6).

A generalization of this example concerns the generic classes of V- and U-statistics. Let $\mathcal{U} \subseteq \mathbb{R}^d$, $m \le n$ and for each $\mathbf{j} \in \{1, ..., n\}^m$ let $\kappa_{\mathbf{j}} : \mathcal{U}^m \to \mathbb{R}$. Define V,$U : \mathcal{U}^n \to \mathbb{R}$ by

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

The next theorem shows that V and U inherit the Lipschitz properties of the worst kernel $\kappa_{\mathbf{j}}$, scaled down by a factor of m/n and m^2/n respectively.

Theorem 4 *Let f be either V or U and $\mathcal{U} \subseteq \mathbb{R}^d$. Then $M_{Lip}(f) \le \max_{\mathbf{j}} M_{Lip}(\kappa_{\mathbf{j}})\, m/n$, $J_{Lip}(f) \le 2\max_{\mathbf{j}} M_{Lip}(\kappa_{\mathbf{j}})\, m^2/n$, $M(f) \le \max_{\mathbf{j}} M(\kappa_{\mathbf{j}})\, m/n$ and $J(f) \le 2\max_{\mathbf{j}} M(\kappa_{\mathbf{j}})\, m^2/n$.*

The easy proof is given in Appendix B. Symmetrization inequalities and uniform bounds are then immediate from Theorem 2 and Corollary 3, without any symmetry assumptions on kernels or variables.

2.2. Lipschitz L-statistics and robust clustering

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

$$\mathcal{L}_F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} F(i/n)\, x_{(i)}. \tag{7}$$

The following result is shown by Maurer and Pontil (2018).

Theorem 5 *For $\alpha, \beta \in \mathbb{R}$ let $[[\alpha,\beta]]$ denote the interval $[\min\{\alpha,\beta\}, \max\{\alpha,\beta\}]$. Then*

$$\left|D^k_{y,y'}\mathcal{L}_F(\mathbf{x})\right| \leq \frac{\|F\|_\infty \, diam\left([[y,y']]\right)}{n} \tag{8}$$

$$\left|D^l_{z,z'}D^k_{y,y'}\mathcal{L}_F(\mathbf{x})\right| \leq \frac{\|F\|_{Lip} \, diam\left([[z,z']] \cap [[y,y']]\right)}{n^2} \tag{9}$$

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

It follows that $M(\mathcal{L}_F) \leq \Delta \|F\|_\infty / n$, $M_{Lip}(\mathcal{L}_F) \leq \|F\|_\infty / n$ and $J_{Lip}(\mathcal{L}_F) \leq \|F\|_{Lip}/n$. For a $\mathcal{U}$-valued function class $\mathcal{H}$ Corollary 3 implies the following uniform bound. For every $\delta \in (0,1)$ with probability at least $1-\delta$ in $\mathbf{X}$ it holds, that

$$\left|\sup_{h\in\mathcal{H}} \mathbb{E}\left[\mathcal{L}_F\left(h\left(\mathbf{X}'\right)\right)\right] - \mathcal{L}_F\left(h\left(\mathbf{X}\right)\right)\right|$$
$$\leq \frac{\sqrt{2\pi}\left(2\|F\|_\infty + \|F\|_{Lip}\right)\, \mathbb{E}\left[G\left(\mathcal{H}\left(\mathbf{X}'\right)\right)\right]}{n} + \Delta\|F\|_\infty\sqrt{\frac{\ln(2/\delta)}{n}}.$$

Lipschitz L-statistics generalize the arithmetic mean, which is obtained by choosing F identically 1. Other choices of F lead to smoothly trimmed means or smoothened sample-quantiles.

A potential use is in robust learning. It often happens that an objective can be minimized very well only if a small proportion of outliers is trimmed away previously. The problem is that minimization must already be performed to identify the outliers, which suggests a procedure to re-sort the sample according to current losses previous to each optimization step which then disregards an upper percentile of losses. Since this generally results in non-convex algorithms, it seems natural to consider problems which are already non-convex to begin with.

We illustrate this idea in the case of K-means clustering (see Garcia et al. (2007)). Here we seek a collection $\mathbf{c} = (c_1, ..., c_K)$ of vectors in some ball $\mathbb{B} \subseteq \mathbb{R}^m$ such that for a given random vector X distributed in $\mathbb{B}$ the quantity $\mathbb{E}[\ell(\mathbf{c}, X_i)]$ is small, where $\ell(\mathbf{c}, X) = \min_{k\in\{1,...,K\}} \|X - c_k\|^2$. For a sample $\mathbf{X} = (X_1, ..., X_n)$ the standard strategy tries to find $\mathbf{c} \in \mathbb{B}^K$ so as to minimize the arithmetic mean of the vector $(\ell(\mathbf{c}, X_1), ..., \ell(\mathbf{c}, X_n))$. Uniform bounds on the estimation error have been given by Biau et al. (2008).

Now we assume that a significant portion of the data (say 25%) consists of noise, which is likely to affect the positions of the centers, but we are happy to cluster only the remaining 75%, which we expect to cluster well. For $\zeta \in [0, 1/4]$ let $F_\zeta : [0,1] \to \mathbb{R}$ be the function

$$F_\zeta(t) = \begin{cases} 4/3 & \text{if} \quad t \in [0, 3/4 - \zeta] \\ -\frac{2}{3\zeta}(t - 3/4 - \zeta) & \text{if} \quad t \in (3/4 - \zeta, 3/4 + \zeta] \\ 0 & \text{if} \quad t \in (3/4 + \zeta, 0] \end{cases}.$$

Then F_0 is the step function which drops from $4/3$ to zero at $t = 3/4$ and $\mathcal{L}_{F_0}$ is a sample quantile, averaging the lower 75%. If $\zeta \in (0, 1/4]$ then F_ζ is an approximation to F_0 with Lipschitz constant $2/(3\zeta)$ and $\mathcal{L}_{F_\zeta}$ is an approximation to the sample quantile. Consider the algorithm

$$\min_{\mathbf{c}\in\mathbb{B}^K} \mathcal{L}_{F_\zeta}\left(\ell\left(\mathbf{c}, X_1\right), ..., \ell\left(\mathbf{c}, X_n\right)\right).$$

The uniform bound above then provides a statistical performance guarantee for this algorithm with respect to the transductive objective $\mathbb{E}\left[\mathcal{L}_{F_\zeta}\left(\ell\left(\mathbf{c}, X_1\right), ..., \ell\left(\mathbf{c}, X_n\right)\right)\right]$ (for a bound on the Gaussian average of $\left\{\left(\ell\left(\mathbf{c}, X_1\right), ..., \ell\left(\mathbf{c}, X_n\right)\right) : \mathbf{c} \in\mathbb{B}^K\right\}$ see Biau et al. (2008)). This method is a smoothened version of the trimmed-K-means algorithms as described in Cuesta-Albertos et al. (1997) .

The general idea of replacing the arithmetic mean of an objective function by a smoothened sample-quantile can be applied to other methods of supervised or unsupervised learning, with above statistical guarantees. The practicality of this method remains to be explored experimentally.

2.3. Differentiation, ℓ_2-regularization and representation learning

For smooth statistics the seminorms M, M_{Lip} and J_{Lip} can often be bounded by differentiation. If $\mathcal{U} \subseteq \mathbb{R}^d$ is open and $f : \mathcal{U}^n \to \mathbb{R}$ is C^2 then for $k, l \in \{1, ..., n\}$ and $i, j \in \{1, ..., d\}$ the function $(\partial f/\partial x_{ki})(\mathbf{x})$ is simply the partial derivative of f in the (k, i)-coordinate. Likewise $\left(\partial^2 f/\partial x_{ki}\partial x_{lj}\right)(\mathbf{x})$ is the partial derivative corresponding to the coordinate pair $((k, i), (l, j))$. We now introduce the notation $\partial_k f$ for the vector valued function $\partial_k f : \mathcal{U}^n \to \mathbb{R}^d$

$$\partial_k f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_{k1}}(\mathbf{x}), ..., \frac{\partial f}{\partial x_{kd}}(\mathbf{x})\right)$$

and $\partial_{kl} f$ for the matrix valued function $\partial_{kl} f : \mathcal{U}^n \to \mathbb{R}^{d\times d}$

$$\partial_{kl} f = \left(\begin{array}{ccc} \frac{\partial^2 f}{\partial x_{k1}\partial x_{l1}}(\mathbf{x}) & ... & \frac{\partial^2 f}{\partial x_{kd}\partial x_{l1}}(\mathbf{x}) \\ ... & ... & ... \\ \frac{\partial^2 f}{\partial x_{k1}\partial x_{ld}}(\mathbf{x}) & ... & \frac{\partial^2 f}{\partial x_{kd}\partial x_{ld}}(\mathbf{x}) \end{array}\right).$$

With $\|\partial_k f\| = \sup_{\mathbf{x}\in\mathcal{U}^n} \|\partial_k f(\mathbf{x})\|$ we denote the supremum of the euclidean norm $\|\partial_k f(\mathbf{x})\|$ of the vector $\partial_k f(\mathbf{x})$ in $\mathcal{U}^n$, and with $\|\partial_{kl} f\| = \sup_{\mathbf{x}\in\mathcal{U}^n} \|\partial_{kl} f(\mathbf{x})\|$ the supremum of the operator norm $\|\partial_{kl} f(\mathbf{x})\|_{op}$ of the matrix $\partial_k f(\mathbf{x})$ in $\mathcal{U}^n$.

Theorem 6 *If $\mathcal{U} \subseteq \mathbb{R}^d$ is convex and bounded with diameter Δ and $f : \mathcal{U}^n \to \mathbb{R}$ extends to a C^2-function on an open set $\mathcal{V}$ containing $\mathcal{U}^n$ then $M_{Lip}(f) \leq \max_k \|\partial_k f\|$ and $J_{Lip}(f) \leq n\Delta \max_{k\neq l} \|\partial_{kl} f\|$.*

See Appendix C for the proof. The uniform estimation properties of a smooth statistic can therefore be described in terms of bounds on the partial derivatives. Good results are obtained if first order partial derivatives are of order $O(1/n)$ and second order derivatives are of order $O\left(1/n^2\right)$.

We sketch an application to representation learning. Let $\mathbb{B}$ be the unit ball $\mathbb{R}^d$ and let $\mathcal{U} = \mathbb{B}\times[-1,1]$. Fix $\lambda\in(0,1)$. For $\mathbf{x}=((z_1,y_1),...,(z_n,y_n))\in\mathcal{U}^n$ regularized least squares returns the vector

$$w(\mathbf{x}) = \arg\min_{w\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^n(\langle w,z_i\rangle-y_i)^2+\lambda\|w\|^2.$$

The "empirical error" f on $\mathcal{U}^n$ is then

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n(\langle w(\mathbf{x}),z_i\rangle-y_i)^2.$$

Using the well known explicit formula for $w(\mathbf{x})$ and $f(\mathbf{x})$ one can show (see Maurer (2017)) by differentiation that there are absolute constants c_1 and c_2, such that for any $k,l\in\{1,...,n\}$, $k\neq l$,

$$\|\partial_k f\|\leq\frac{c_1\lambda^{-2}}{n}\text{ and }\|\partial_{kl}f\|\leq\frac{c_2\lambda^{-3}}{n^2}. \tag{10}$$

So, taking the diameter of $\mathcal{U}$ into account, we have $M(f)\leq c_3n^{-1}\lambda^{-2}$, $M_{Lip}(f)\leq c_4n^{-1}\lambda^{-2}$ and $J_{Lip}(f)\leq c_4n^{-1}\lambda^{-3}$.

Now let $\mathcal{H}$ be a class of representations of some underlying space $\mathcal{X}$ of labeled data, that is functions $h:\mathcal{X}\to\mathcal{U}$, which leave the labels invariant, and we wish to find an optimal representation. If we plan to use ridge regression in the top layer, the obvious criterion for the quality of the representation on a sample $\mathbf{X}\in\mathcal{X}^n$ is

$$\mathbb{E}[f(h(\mathbf{X}))]=\mathbb{E}\left[(\langle w(h(\mathbf{X})),Z\rangle-Y)^2\right].$$

Then Corollary 3 combined with Theorem 6 and (10) gives a high probability bound on

$$\sup_{h\in\mathcal{H}}\mathbb{E}\left[f\left(h\left(\mathbf{X}'\right)\right)\right]-f(h(\mathbf{X})),$$

so as to justify the minimization of $f(h(\mathbf{X}))$ in h if the Gaussian average $\mathbb{E}[G(\mathcal{H}(\mathbf{X}))]$ can be bounded.

3. Proof of Theorem 2

We prove the theorem for $\mathcal{U}\subseteq\mathbb{R}$, the proof for $\mathcal{U}\subseteq\mathbb{R}^d$ being the same but with additional notation. For this section and Appendix D we use the following notation. We take $f:\mathcal{U}^n\to\mathbb{R}$ as fixed and abbreviate $M=M_{Lip}(f)$ and $J=J_{Lip}(f)$, when there is no ambiguity. For any $i,j\in\mathbb{N}$ we use $[i,j]$ to denote the set of integers $[i,j]=\{i,...,j\}$ if $i\leq j$, or $[i,j]=\emptyset$ if $i>j$. Whenever two vectors in $\mathcal{X}^n$ or $\mathcal{U}^n$ are denoted $\mathbf{x}$ and $\mathbf{x}'$, and $A\subseteq[1,n]$, then we use $\mathbf{x}^A$ to denote the vector in $\mathcal{X}^n$ defined by

$$x_i^A=\begin{cases}x_i' & \text{if} \quad i\in A\\ x_i & \text{if} \quad i\notin A\end{cases}, \tag{11}$$

and we use A^c to denote the complement of A in $[1,n]$. Also $\|.\|$ denotes the euclidean norm, either on $\mathbb{R}^n$ or $\mathbb{R}^{2n}$, depending on context, and $\langle .,.\rangle$ denotes the corresponding inner product.

A random vector $\boldsymbol{\gamma}=(\gamma_1,...,\gamma_n)$ is called isonormal if all the γ_i are independent standard normal variables. If U is an orthogonal matrix and $\boldsymbol{\gamma}$ is isonormal, then $\boldsymbol{\gamma}$ and $U\boldsymbol{\gamma}$ are identically distributed. This implies the following simple lemma.

Lemma 7 *Let $v \in \mathbb{R}^d$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_d)$ be isonormal. Then (i) $\langle \boldsymbol{\gamma}, v \rangle$ is identically distributed to $\|v\| \gamma_1$, and (ii) $\|v\| = \sqrt{\pi/2}\, \mathbb{E} |\langle \boldsymbol{\gamma}, v \rangle|$.*

Proof Let U be orthogonal, mapping v to $(\|v\|, 0, ..., 0)$. Then $\langle \boldsymbol{\gamma}, v \rangle = \left\langle \boldsymbol{\gamma}, U^{-1} (\|v\|, 0, ..., 0) \right\rangle = \|v\| \langle U\boldsymbol{\gamma}, (1, 0, ..., 0) \rangle$ is identically distributed to $\|v\| \langle \boldsymbol{\gamma}, (1, 0, ..., 0) \rangle = \|v\| \gamma_1$. This proves (i), of which the expectation of the absolute value gives (ii). ∎

The next lemma is the key to the way in which the interaction-seminorm $J = J_{Lip}(f)$ enters the proof.

Lemma 8 *For any $k \in [1, n]$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{U}^n$ and $a, b \in \mathcal{U}$*

$$D^k_{a,b} f(\mathbf{x}) - D^k_{a,b} f(\mathbf{x}') \leq \frac{J}{n} \sum_{j: j \neq k} \left| x_j - x'_j \right|.$$

Proof First assume $k = 1$. Then

$$\begin{aligned} D^1_{a,b} f(\mathbf{x}) - D^1_{a,b} f(\mathbf{x}') &= \sum_{j=2}^n D^1_{a,b} f\left(\mathbf{x}^{[1,j-1]}\right) - D^1_{a,b} f\left(\mathbf{x}^{[1,j]}\right) \\ &= \sum_{j=2}^n D^1_{a,b} D^j_{x_j x'_j} f\left(\mathbf{x}^{[1,j]}\right) \leq \frac{J}{n} \sum_{j=2}^n \left| x_j - x'_j \right|. \end{aligned}$$

If $k \neq 1$ let f_π be the function $f_\pi(x) = f(\pi x)$, where π is the permutation exchanging the first and the k-th argument, observe that $J_{Lip}(f_\pi) = J_{Lip}(f) = J$, and apply the above to f_π. ∎

For $k \in \{1, ..., n\}$ define a function $F_k : \mathcal{U}^{2n} \to \mathbb{R}$ by

$$F_k(\mathbf{x}, \mathbf{x}') = \frac{1}{2^k} \sum_{A \subseteq [1, k-1]} \left(D^k_{x_k, x'_k} f\left(\mathbf{x}^A\right) + D^k_{x_k, x'_k} f\left(\mathbf{x}^{A^c}\right) \right).$$

$F_k(\mathbf{x}, \mathbf{x}')$ changes sign if we exchange x_k and x'_k. On the other hand, if we exchange x_i and x'_i with $i < k$, then $i \in [1, k-1]$, so we exchange only terms in the above sum (see (13) in Appendix D) and leave $F_k(\mathbf{x}, \mathbf{x}')$ invariant. This is the reason why we use the somewhat complicated representation of $f(\mathbf{x}) - f(\mathbf{x}')$, as given by the next lemma.

Lemma 9 *For $\mathbf{x}, \mathbf{x}' \in \mathcal{U}^n$ we have*

$$f(\mathbf{x}) - f(\mathbf{x}') = \sum_{k=1}^n F_k(\mathbf{x}, \mathbf{x}').$$

The proof is given in Appendix D.

For the remainder of this section let V be the $n \times n$-matrix

$$V_{ki} = \frac{J}{\sqrt{n-1}} + \left(2M - \frac{J}{\sqrt{n-1}} \right) I_{ki},$$

where I_{kj} is the identity. V_{ki} is $2M$ on the diagonal and $J/\sqrt{n-1}$ everywhere else. For $k \in [1,n]$ we define an operator v_k on $\mathbb{R}^n$ by $(v_k \mathbf{x})_i = V_{ki} x_i$. So v_k is multiplication by the k-th row of V. We extend v_k to $\mathbb{R}^{2n}$, setting

$$v_k\left(\mathbf{x}, \mathbf{x}'\right) = \left(v_k \mathbf{x}, v_k \mathbf{x}'\right).$$

Lemma 10 *For* $\left(\mathbf{x}, \mathbf{x}'\right), \left(\mathbf{y}, \mathbf{y}'\right) \in \mathcal{U}^{2n}$ *and* $k \in [1,n]$ *we have*

$$F_k\left(\mathbf{x}, \mathbf{x}'\right) - F_k\left(\mathbf{y}, \mathbf{y}'\right) \leq \sqrt{\pi/2}\, \mathbb{E}\left|\left\langle \boldsymbol{\gamma}_k, v_k\left(\mathbf{x}, \mathbf{x}'\right) - v_k\left(\mathbf{y}, \mathbf{y}'\right)\right\rangle\right|,$$

where $\boldsymbol{\gamma}_k = \left(\gamma_{k1}, ..., \gamma_{kn}, \gamma'_{k1}, ..., \gamma'_{kn}\right)$ *is an isonormal vector of* $2n$ *components.*

Proof Using the definition of $M = M_{Lip}(f)$ and Lemma 8 we have for any $A \subseteq \{1, ..., n\}$

$$\begin{aligned} D^k_{x_k, x'_k} f\left(\mathbf{x}^A\right) - D^k_{y_k, y'_k} f\left(\mathbf{y}^A\right) &= D^k_{x_k, y_k} f\left(\mathbf{x}^A\right) + D^k_{y'_k, x'_k} f\left(\mathbf{x}^A\right) + D^k_{y_k, y'_k}\left(f\left(\mathbf{x}^A\right) - f\left(\mathbf{y}^A\right)\right) \\ &\leq M\left(\left|x_k - y_k\right| + \left|x'_k - y'_k\right|\right) + \frac{J}{n} \sum_{i: i \neq k} \left|x_i^A - y_i^A\right|. \end{aligned}$$

This gives

$$\begin{aligned} &F_k\left(\mathbf{x}, \mathbf{x}'\right) - F_k\left(\mathbf{y}, \mathbf{y}'\right) \\ &= \frac{1}{2^k} \sum_{A \subseteq [1, k-1]} \left(D^k_{x_k, x'_k} f\left(\mathbf{x}^A\right) - D^k_{y_k, y'_k} f\left(\mathbf{y}^A\right) + D^k_{x_k, x'_k} f\left(\mathbf{x}^{A^c}\right) - D^k_{y_k, y'_k} f\left(\mathbf{y}^{A^c}\right)\right) \\ &\leq \frac{1}{2^k} \sum_{A \subseteq [1, k-1]} \left(2M\left(\left|x_k - y_k\right| + \left|x'_k - y'_k\right|\right) + \frac{J}{n} \sum_{i: i \neq k} \left|x_i^A - y_i^A\right| + \frac{J}{n} \sum_{i: i \neq k} \left|x_i^{A^c} - y_i^{A^c}\right|\right) \\ &= M\left(\left|x_k - y_k\right| + \left|x'_k - y'_k\right|\right) + \frac{J}{2n} \sum_{i: i \neq k} \left(\left|x_i - y_i\right| + \left|x'_i - y'_i\right|\right). \end{aligned}$$

Using the inequalities $|a| + |b| \leq \sqrt{2}\sqrt{a^2 + b^2}$ and $\sum_1^n |a_i| \leq \sqrt{n}\sqrt{\sum_1^n a_i^2}$ we can bound the last expression by

$$\begin{aligned} &\leq \sqrt{2} M\left(\left(x_k - y_k\right)^2 + \left(x'_k - y'_k\right)^2\right)^{1/2} + \frac{J}{\sqrt{2n}} \left(\sum_{i: i \neq k} \left(x_i - y_i\right)^2 + \left(x'_i - y'_i\right)^2\right)^{1/2} \\ &\leq \left(4M^2\left(x_k - y_k\right)^2 + \frac{J^2}{n-1} \sum_{i: i \neq k} \left(x_i - y_i\right)^2 + 4M^2\left(x'_k - y'_k\right)^2 + \frac{J^2}{n-1} \sum_{i: i \neq k} \left(x'_i - y'_i\right)^2\right)^{1/2} \\ &= \left\|v_k\left(\left(\mathbf{x}, \mathbf{x}'\right) - \left(\mathbf{y}, \mathbf{y}'\right)\right)\right\| = \sqrt{\pi/2}\, \mathbb{E}\left|\left\langle \boldsymbol{\gamma}_k, v_k\left(\left(\mathbf{x}, \mathbf{x}'\right) - \left(\mathbf{y}, \mathbf{y}'\right)\right)\right\rangle\right|, \end{aligned}$$

where the last identity follows from Lemma 7 (ii). ■

Proof (of Theorem 2) With $\mathbf{X}'$ identically distributed to $\mathbf{X}$ we have

$$\mathbb{E} \sup_h \mathbb{E}_{\mathbf{X}}\left[f\left(h\left(\mathbf{X}\right)\right)\right] - f\left(h\left(\mathbf{X}'\right)\right) \leq \mathbb{E} \sup_h f\left(h\left(\mathbf{X}\right)\right) - f\left(h\left(\mathbf{X}'\right)\right),$$

so it suffices to bound the right hand side above. We first prove that

$$\mathbb{E}\sup_h f\left(h\left(\mathbf{X}\right)\right)-f\left(h\left(\mathbf{X}'\right)\right)\leq\sqrt{\pi/2}\,\mathbb{E}\sup_h\sum_{k=1}^{n}\left\langle\boldsymbol{\gamma}_k,v_k\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle, \tag{12}$$

where the $\boldsymbol{\gamma}_1,...,\boldsymbol{\gamma}_n$ are independent copies of the isonormal vector $\boldsymbol{\gamma}_k$ in Lemma 10.

To prove (12) we show by induction on $m\in\{0,...,n\}$ that

$$\begin{aligned}&\mathbb{E}\sup_h f\left(h\left(\mathbf{X}\right)\right)-f\left(h\left(\mathbf{X}'\right)\right)\\&\qquad\leq\mathbb{E}\left[\sup_h\sqrt{\pi/2}\sum_{k=1}^{m}\left\langle\boldsymbol{\gamma}_k,v_k\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle+\sum_{k=m+1}^{n}F_k\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right].\end{aligned}$$

For $m=n$ this is (12), and for $m=0$ it is just Lemma 9. Suppose it holds for $m-1$, with some $m\leq n$, and define for each $h\in\mathcal{H}$ a real valued random variable R_h by

$$R_h=\sqrt{\pi/2}\sum_{k=1}^{m-1}\left\langle\boldsymbol{\gamma}_k,v_k\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle+\sum_{k=m+1}^{n}F_k\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right).$$

The expectation $\mathbb{E}=\mathbb{E}_{\mathbf{X}\mathbf{X}'}\mathbb{E}_{\boldsymbol{\gamma}}\left[.\right]$ is invariant under the simultaneous exchange of X_m and X_m' and, for all $k<m$, of γ_{km} and γ_{km}', which leaves R_h invariant but changes the sign of F_m. Using this fact and the induction assumption

$$\begin{aligned}&\mathbb{E}\sup_h f\left(h\left(\mathbf{X}\right)\right)-f\left(h\left(\mathbf{X}'\right)\right)\\&\leq\mathbb{E}\sup_h F_m\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)+R_h\\&=\frac{1}{2}\mathbb{E}\sup_{h,g}F_m\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)-F_m\left(g\left(\mathbf{X}\right),g\left(\mathbf{X}'\right)\right)+R_h+R_g.\end{aligned}$$

Using Lemma 10, with $(\mathbf{x},\mathbf{x}')$ replaced by $\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)$ and $(\mathbf{y},\mathbf{y}')$ replaced by $\left(g\left(\mathbf{X}\right),g\left(\mathbf{X}'\right)\right)$, we get

$$\begin{aligned}&\mathbb{E}\sup_h f\left(h\left(\mathbf{X}\right)\right)-f\left(h\left(\mathbf{X}'\right)\right)\\&\leq\frac{1}{2}\mathbb{E}\sup_{h,g}\sqrt{\pi/2}\,\mathbb{E}_{\gamma_m}\left|\left\langle\boldsymbol{\gamma}_m,v_m\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)-v_m\left(g\left(\mathbf{X}\right),g\left(\mathbf{X}'\right)\right)\right\rangle\right|+R_h+R_g\\&\leq\frac{1}{2}\mathbb{E}\sup_{h,g}\sqrt{\pi/2}\left|\left\langle\boldsymbol{\gamma}_m,v_m\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle-\left\langle\boldsymbol{\gamma}_m,v_m\left(g\left(\mathbf{X}\right),g\left(\mathbf{X}'\right)\right)\right\rangle\right|+R_h+R_g\\&=\frac{1}{2}\mathbb{E}\sup_{h,g}\sqrt{\pi/2}\left\langle\boldsymbol{\gamma}_m,v_m\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle-\sqrt{\pi/2}\left\langle\boldsymbol{\gamma}_m,v_m\left(g\left(\mathbf{X}\right),g\left(\mathbf{X}'\right)\right)\right\rangle+R_h+R_g.\end{aligned}$$

Here we could drop the absolute value because the supremum is in both h and g, and the remaining sum is invariant under the exchange of h and g. The symmetry of the standard normal distribution then gives

$$\begin{aligned}\mathbb{E}\sup_h f\left(h\left(\mathbf{X}\right)\right)-f\left(h\left(\mathbf{X}'\right)\right) &\leq \frac{1}{2}\mathbb{E}\sup_h \sqrt{\pi/2}\left\langle \boldsymbol{\gamma}_m, v_m\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle + R_h \\ &\quad +\frac{1}{2}\mathbb{E}\sup_g \sqrt{\pi/2}\left\langle -\boldsymbol{\gamma}_m, v_m\left(g\left(\mathbf{X}\right),g\left(\mathbf{X}'\right)\right)\right\rangle + R_g \\ &= \mathbb{E}\sup_h \sqrt{\pi/2}\left\langle \boldsymbol{\gamma}_m, v_m\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle + R_h.\end{aligned}$$

By definition of R_h this completes the induction and proves the claim (12).

Let $(\eta_1,...,\eta_n,\eta_1',...,\eta_n')$ be isonormal and independent of all the γ_{ki}. By Lemma 7 (i) the random variable $\sum_{k=1}^n \gamma_{ki}V_{ki}$ is identically distributed to

$$\left(\sum_{k=1}^n V_{ki}^2\right)^{1/2}\eta_i = \sqrt{4M^2+J^2}\,\eta_i.$$

Unraveling the definition of v_k it follows for every $h\in\mathcal{H}$ that

$$\sum_{k=1}^n \left\langle \boldsymbol{\gamma}_k, v_k\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle = \sum_{i=1}^n\left[\left(\sum_{k=1}^n \gamma_{ki}V_{ki}\right)h\left(X_i\right)+\left(\sum_{k=1}^n \gamma_{ki}'V_{ki}\right)h\left(X_i'\right)\right]$$

is identically distributed to

$$\sqrt{4M^2+J^2}\sum_{i=1}^n\left(\eta_i h\left(X_i\right)+\eta_i' h\left(X_i'\right)\right).$$

Combined with (12) this gives

$$\begin{aligned}\mathbb{E}\sup_h f\left(h\left(\mathbf{X}\right)\right)-f\left(h\left(\mathbf{X}'\right)\right) &\leq \sqrt{\pi/2}\mathbb{E}\sup_h\sum_{k=1}^n\left\langle \boldsymbol{\gamma}_k, v_k\left(h\left(\mathbf{X}\right),h\left(\mathbf{X}'\right)\right)\right\rangle \\ &= \sqrt{\pi/2}\sqrt{4M^2+J^2}\;\mathbb{E}\sup_h\sum_{i=1}^n \eta_i h\left(X_i\right)+\eta_i' h\left(X_i'\right) \\ &\leq \sqrt{2\pi}\left(2M+J\right)\;\mathbb{E}\,G\left(\mathcal{H}\left(\mathbf{X}\right)\right).\end{aligned}$$

■

References

S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, S., and D. Roth, D. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393-425, 2005.

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Complexity of pattern classes and the Lipschitz property. *Theoretical Computer Science*, 382(3):232–246, 2007.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3: 463–482, 2002.

G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

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. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán, C. Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576, 1997.

L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2-3):89–109, 2010.

S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In Advances in Neural Information Processing Systems, pp. 793–800, 2009.

S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Regularization Techniques for Learning with Matrices. *Journal of Machine Learning Research* 13:1865–1890, 2012.

V. Koltchinskii and D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, *The Annals of Statistics*, 30(1):1–50, 2002.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*, Springer, 1991.

A. Maurer and M. Pontil. Empirical bounds for functions with weak interactions. Proceedings of the 31st Annual Conference on Learning Theory, *PMLR*, 75:987–1010, 2018.

A. Maurer. A Second-order look at stability and generalization. Proceedings of the 30th Annual Conference on Learning Theory, *PMLR*, 65:1461-1475, 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*, pp. 195–248. Springer, Berlin, 1998.

R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In Advances in neural information processing systems, pp. 451–459, 2016.

Appendix

Appendix A. The bounded difference inequality

Theorem 11 *(McDiarmid (1998) or Boucheron et al. (2013)) Suppose $f:\mathcal{X}^n\to\mathbb{R}$ and $\mathbf{X}=(X_1,...,X_n)$ is a vector of independent random variables with values in $\mathcal{X}$, $\mathbf{X}'$ is iid to $\mathbf{X}$. Then*

$$\Pr\left\{f(\mathbf{X})-\mathbb{E}f\left(\mathbf{X}'\right)>t\right\}\leq\exp\left(\frac{-2t^2}{\sup_{\mathbf{x}\in\mathcal{X}^n}\sum_k\sup_{y,y'\in\mathcal{X}}\left(D^k_{y,y'}f(\mathbf{x})\right)^2}\right).$$

Appendix B. U- and V-statistics

Proof (of Theorem 4) We first consider V.

$$\begin{aligned}D^k_{y,y'}V(\mathbf{x}) &\leq n^{-m}\sum_{\mathbf{j}\in(\mathbb{N}_n)^m,\exists i\in\mathbb{N}_m,\,j_i=k}D^i_{y,y'}\kappa_{\mathbf{j}}\left(x_{j_1},...,x_{j_m}\right)\\ &\leq n^{-m}\sum_{\mathbf{j}\in(\mathbb{N}_n)^m,\exists i\in\mathbb{N}_m,\,j_i=k}M(\kappa_{\mathbf{j}}).\end{aligned}$$

But $n^{-m}\left|\{\mathbf{j}\in(\mathbb{N}_n)^m,\exists i\in\mathbb{N}_m,\,j_i=k\}\right|=m/n$. So $M(V)\leq m\max_{\mathbf{j}}M(\kappa_{\mathbf{j}})/n$, with exactly the same argument for $M_{Lip}(V)$. Also

$$\begin{aligned}&D^l_{z,z'}D^k_{y,y'}V(\mathbf{x})\\ \leq\;& n^{-m}\sum_{\substack{\mathbf{j}\in(\mathbb{N}_n)^m,\exists i,i'\in\mathbb{N}_m,\\ j_i=k,j_{i'}=l}}D^{i'}_{z,z'}D^i_{y,y'}\kappa_{\mathbf{j}}\left(x_{j_1},...,x_{j_m}\right)\\ \leq\;& n^{-m}\sum_{\substack{\mathbf{j}\in(\mathbb{N}_n)^m,\exists i,i'\in\mathbb{N}_m,\\ j_i=k,j_{i'}=l}}2M(\kappa_{\mathbf{j}}).\end{aligned}$$

But

$$n^{-m}\left|\left\{\mathbf{j}\in(\mathbb{N}_n)^m:\exists i,i'\in\mathbb{N}_m,\,j_i=k,j_{i'}=l\right\}\right|\leq m^2/n^2.$$

So $J(f)\leq 2m^2\max_{\mathbf{j}}M(\kappa_{\mathbf{j}})/n$, with exactly the same argument for $J_{Lip}(f)$. This completes proof for V-statistics. For the case of U-statistics we have to count the number of subsets $S\subseteq\mathbb{N}_n$ of cardinality m containing a fixed $k\in\mathbb{N}_n$ or two distict k,$l\in\mathbb{N}_n$ respectively. This is $\binom{n-1}{m-1}$ or $\binom{n-2}{m-2}$ respectively and

$$\frac{\binom{n-1}{m-1}}{\binom{n}{m}}=\frac{m!\,(n-1)!}{n!\,(m-1)!}=\frac{m}{n}\text{ or }\frac{\binom{n-2}{m-2}}{\binom{n}{m}}=\frac{m!\,(n-2)!}{n!\,(m-2)!}=\frac{m(m-1)}{n(n-1)}\leq\frac{m^2}{n^2}.$$

■

Appendix C. Differentiation

Proof (of Theorem 6) Fix $\mathbf{x} \in \mathcal{U}^n$, $y, y', z, z' \in \mathcal{X}$ and $k \neq l \in \mathbb{N}_n$. For $0 \leq s, t \leq 1$ define $\mathbf{x}(t) = S^k_{y'+t(y-y')}\mathbf{x}$ and $\mathbf{x}(s,t) = S^l_{z'+t(z-z')}S^k_{y'+t(y-y')}\mathbf{x}$. Convexity insures that $f(\mathbf{x}(t))$ and $f(\mathbf{x}(s,t))$ is defined for all values of s and t. Then

$$\begin{aligned} D^k_{yy'}f(\mathbf{x}) &= f(\mathbf{x}(1)) - f(\mathbf{x}(0)) = \int_0^1 \left\langle \partial_k f(\mathbf{x}(t)), y - y' \right\rangle dt \\ &\leq \|\partial_k f\| \left\|y - y'\right\|. \end{aligned}$$

Similarly

$$\begin{aligned} D^l_{zz'}D^k_{yy'}f(\mathbf{x}) &= (f(\mathbf{x}(1,1)) - f(\mathbf{x}(1,0))) - (f(\mathbf{x}(0,1)) - f(\mathbf{x}(0,0))) \\ &= \int_0^1\int_0^1 \left\langle \partial_{lk} f(\mathbf{x}(s,t)) \left(z - z'\right), \left(y - y'\right) \right\rangle ds dt \leq \|\partial_{lk} f\| \left\|y - y'\right\| \left\|z - z'\right\| \\ &\leq \Delta \|\partial_{lk} f\| \left\|y - y'\right\|. \end{aligned}$$

■

Appendix D. Proof of Lemma 9

We will use the following elementary fact: if Φ is a function defined on subsets of $[1, k-1]$, then for every $i \in [1, k-1]$

$$\sum_{A \subseteq [1,k-1]} \Phi(A) = \sum_{A \subseteq [1,k-1] \backslash [i]} (\Phi(A) + \Phi(A \cup [i])). \tag{13}$$

Also note that

$$\sum_{A \subseteq [1,k-1]} D^k_{x_k,x'_k} f\left(\mathbf{x}^{A^c}\right) = \sum_{A \subseteq [1,k-1]} D^k_{x_k,x'_k} f\left(\mathbf{x}^{A \cup [k,n]}\right),$$

so it suffices to prove the following

Claim: For every set $\mathcal{X}$, all $n \in \mathbb{N}$, all functions $f : \mathcal{X}^n \to \mathbb{R}$ and all vectors $\mathbf{x}$ and $\mathbf{x}' \in \mathcal{X}^n$

$$f(\mathbf{x}) - f\left(\mathbf{x}'\right) = \sum_{k=1}^n \frac{1}{2^k} \sum_{A \subseteq [1,k-1]} \left(D^k_{x_k,x'_k} f\left(\mathbf{x}^A\right) + D^k_{x_k,x'_k} f\left(\mathbf{x}^{A \cup [k,n]}\right)\right).$$

Proof By induction on n. Since the empty set is the only subset of $[1, 0]$, the case $n = 1$ reduces to

$$f(\mathbf{x}) - f\left(\mathbf{x}'\right) = \frac{1}{2}\left(f(\mathbf{x}) - f\left(\mathbf{x}'\right) + f(\mathbf{x}) - f\left(\mathbf{x}'\right)\right).$$

Assume the claim to be true for $n-1$ and fix f, $\mathbf{x}$ and $\mathbf{x}'$. Let $\mathbf{z}$ and $\mathbf{z}'$ be the $(n-1)$-dimensional vectors $(x_2, ..., x_n)$ and $(x_2', ..., x_n')$ respectively and define $g : \mathcal{X}^{n-1} \to \mathbb{R}$ by $g(\mathbf{z}) = f(x_1, \mathbf{z})$. By the induction assumption applied to g and a change of variables

$$\begin{aligned} f(\mathbf{x}) - f\left(\mathbf{x}^{[2,n]}\right) &= g(\mathbf{z}) - g(\mathbf{z}') \\ &= \sum_{k=1}^{n-1} \frac{1}{2^k} \sum_{A \subseteq [1,k-1]} D^k_{z_k,z_k'} g\left(\mathbf{z}^A\right) + D^k_{z_k,z_k'} g\left(\mathbf{z}^{A \cup [k,n-1]}\right) \\ &= \sum_{k=2}^{n} \frac{1}{2^{k-1}} \sum_{A \subseteq [2,k-1]} D^k_{x_k,x_k'} f\left(\mathbf{x}^A\right) + D^k_{x_k,x_k'} f\left(\mathbf{x}^{A \cup [k,n]}\right). \end{aligned}$$

By the same argument, replacing x_1 by x_1' in the definition of g,

$$f\left(\mathbf{x}^{[1]}\right) - f\left(\mathbf{x}^{[1,n]}\right) = \sum_{k=2}^{n} \frac{1}{2^{k-1}} \sum_{A \subseteq [2,k-1]} D^k_{x_k,x_k'} f\left(\mathbf{x}^{[1] \cup A}\right) + D^k_{x_k,x_k'} f\left(\mathbf{x}^{[1] \cup A \cup [k,n]}\right).$$

Thus, adding and subtracting $f\left(\mathbf{x}^{[1]}\right)/2$ and $f\left(\mathbf{x}^{[2,n]}\right)/2$ from $f(\mathbf{x}) - f(\mathbf{x}')$, we obtain

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}') &= \frac{1}{2}\left(f(\mathbf{x}) - f\left(\mathbf{x}^{[1]}\right) + f\left(\mathbf{x}^{[2,n]}\right) - f\left(\mathbf{x}^{[1,n]}\right)\right) + \\ &\quad + \frac{1}{2}\left(f(\mathbf{x}) - f\left(\mathbf{x}^{[2,n]}\right) + f\left(\mathbf{x}^{[1]}\right) - f\left(\mathbf{x}^{[1,n]}\right)\right) \\ &= \frac{1}{2}\left(D^1_{x_1x_1'} f(\mathbf{x}) + D^1_{x_1x_1'} f\left(\mathbf{x}^{[1,n]}\right)\right) + \\ &\quad + \frac{1}{2}\sum_{k=2}^{n} \frac{1}{2^{k-1}} \sum_{A \subseteq [2,k-1]} \left(D^k_{x_k,x_k'} f\left(\mathbf{x}^A\right) + D^k_{x_k,x_k'} f\left(\mathbf{x}^{A \cup [k,n]}\right) + \right. \\ &\qquad \left. + D^k_{x_k,x_k'} f\left(\mathbf{x}^{[1] \cup A}\right) + D^k_{x_k,x_k'} f\left(\mathbf{x}^{[1] \cup A \cup [k,n]}\right)\right) \\ &= \frac{1}{2}\left(D^1_{x_1x_1'} f(\mathbf{x}) + D^1_{x_1x_1'} f\left(\mathbf{x}^{[1,n]}\right)\right) + \\ &\quad + \sum_{k=2}^{n} \frac{1}{2^k} \sum_{A \subseteq [1,k-1]} \left(D^k_{x_k,x_k'} f\left(\mathbf{x}^A\right) + D^k_{x_k,x_k'} f\left(\mathbf{x}^{A \cup [k,n]}\right)\right). \end{aligned}$$

The last identity used (13) with $i = 1$. This completes the induction. ■