Concentration in unbounded metric spaces and algorithmic stability

Aryeh Kontorovich

May 2, 2014

Abstract

We prove an extension of McDiarmid's inequality for metric spaces with unbounded diameter. To this end, we introduce the notion of the *subgaussian diameter*, which is a distribution-dependent refinement of the metric diameter. Our technique provides an alternative approach to that of Kutin and Niyogi's method of weakly difference-bounded functions, and yields nontrivial, dimension-free results in some interesting cases where the former does not. As an application, we give apparently the first generalization bound in the algorithmic stability setting that holds for unbounded loss functions. This yields a novel risk bound for some regularized metric regression algorithms. We give two extensions of the basic concentration result. The first enables one to replace the independence assumption by appropriate strong mixing. The second generalizes the subgaussian technique to other Orlicz norms.

1 Introduction

Concentration of measure inequalities are at the heart of statistical learning theory. Roughly speaking, concentration allows one to conclude that the performance of a (sufficiently "stable") algorithm on a (sufficiently "close to iid") sample is indicative of the algorithm's performance on future data. Quantifying what it means for an algorithm to be *stable* and for the sampling process to be *close to iid* is by no means straightforward and much recent work has been motivated by these questions. It turns out that the various notions of stability are naturally expressed in terms of the Lipschitz continuity of the algorithm in question (Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002; Rakhlin et al., 2005; Shalev-Shwartz et al., 2010), while appropriate relaxations of the iid assumption are achieved using various kinds of strong mixing (Karandikar and Vidyasagar, 2002; Gamarnik, 2003; Rostamizadeh and Mohri, 2007; Mohri and Rostamizadeh, 2008; Steinwart and Christmann, 2009; Steinwart et al., 2009; Zou et al.; Mohri and Rostamizadeh, 2010; London et al., 2012, 2013; Shalizi and Kontorovich, 2013).

Many of the aforementioned results are based on McDiarmid's inequality (McDiarmid, 1989):

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n w_i^2}\right), \tag{1}$$

where φ is a real-valued function of the sequence of independent random variables $X = (X_1, \ldots, X_n)$, such that

$$|\varphi(x) - \varphi(x')| \leq w_i \tag{2}$$

whenever x and x' differ only in the i^{th} coordinate. Aside from being instrumental in proving PAC bounds (Boucheron et al., 2005), McDiarmid's inequality has also found use in algorithmic stability results (Bousquet and Elisseeff, 2002). Non-iid extensions of (1) have also been considered (Marton, 1996; Rio, 2000; Chazottes et al., 2007; Kontorovich and Ramanan, 2008).

The distribution-free nature of McDiarmid's inequality makes it an attractive tool in learning theory, but also imposes inherent limitations on its applicability. Chief among these limitations is the inability of (1) to provide risk bounds for unbounded loss functions. Even in the bounded case, if the Lipschitz condition (2) holds not everywhere but only with high probability — say, with a much larger constant on a small set of exceptions — the bound in (1) still charges the full cost of the worst-case constant. To counter this difficulty, Kutin (2002); Kutin and Niyogi (2002) introduced an extension of McDiarmid's inequality to *weakly difference-bounded* functions and used it to analyze the risk of "almost-everywhere" stable algorithms. This influential result has been invoked in a number of recent papers (El-Yaniv and Pechyony, 2006; Mukherjee et al., 2006; Hush et al., 2007; Agarwal and Niyogi, 2009; Shalev-Shwartz et al., 2010; Rubinstein and Simma, 2012).

However, the approach of Kutin and Niyogi entails some difficulties as well. These come in two flavors: analytical (complex statement and proof) and practical (conditions are still too restrictive in some cases); we will elaborate upon this in Section 3. In this paper, we propose an alternative approach to the concentration of "almost-everywhere" or "average-case" Lipschitz functions. To this end, we introduce the notion of the *subgaussian diameter* of a metric probability space. The latter may be finite even when the metric diameter is infinite, and we show that this notion generalizes the more restrictive property of bounded differences.

Main results. This paper's principal contributions include defining the subgaussian diameter of a metric probability space and identifying its role in relaxing the bounded differences condition. In Theorem 1, we show that the subgaussian diameter can essentially replace the far more restrictive metric diameter in concentration bounds. This result has direct ramifications for algorithmic stability (Theorem 2), with applications to regularized regression. We furthermore extend our concentration inequality to non-independent processes (Theorem 3) and to other Orlicz norms (Theorem 4).

Motivation. The concentration properties of unbounded functions become important in settings related to regression, such as sample bias correction, domain adaptation, and boosting (Cortes and Mohri, 2014; Dasgupta and Long, 2003; Ben-David et al., 2006; Dudík et al., 2005; Mansour et al., 2009). Subgaussian distributions occur in many practical applications, such as the histogram features in computer vision (Torralba et al., 2008; Deng et al., 2009). This class of distributions subsumes the Gaussian random variables, as well as all the bounded ones (such as Bernoulli, uniform, and multinomial).

Outline of paper. In Section 2 we define the subgaussian diameter and relate it to (weakly) bounded differences in Section 3. We state and prove the concentration inequality based on this notion in Section 4 and give an application to algorithmic stability in Section 5. We then give an extension to non-independent data in Section 6 and discuss other Orlicz norms in Section 7. Conclusions and some open problems are presented in Section 8.

2 Preliminaries

A *metric probability space* $(\mathcal{X}, \rho, \mu)$ is a measurable space $\mathcal{X}$ whose Borel σ-algebra is induced by the metric ρ, endowed with the probability measure μ. Our results are most cleanly presented when $\mathcal{X}$ is a discrete set but they continue to hold verbatim for general metric probability spaces. In particular, it will be convenient to write $\mathbb{E}\varphi = \sum_{x\in\mathcal{X}} \mathbb{P}(x)\varphi(x)$ even when the latter is an integral. Random variables are capitalized (X), specified sequences are written in lowercase, the notation $X_i^j = (X_i, \ldots, X_j)$ is used for all sequences, and sequence concatenation is denoted multiplicatively: $x_i^j x_{j+1}^k = x_i^k$. We will frequently use the shorthand $\mathbb{P}(x_i^j) = \prod_{k=i}^j \mathbb{P}(X_k = x_k)$. Standard order of magnitude notation such as $O(\cdot)$ and $\Omega(\cdot)$ will be used.

A function $\varphi : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if

$$|\varphi(x) - \varphi(x')| \le L\rho(x, x'), \qquad x, x' \in \mathcal{X}.$$

Let $(\mathcal{X}_i, \rho_i, \mu_i)$, $i = 1, \ldots, n$ be a sequence of metric probability spaces. We define the product probability space

$$\mathcal{X}^n = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n$$

with the product measure

$$\mu^n = \mu_1 \times \mu_2 \times \ldots \times \mu_n$$

and ℓ_1 product metric

$$\rho^n(x, x') = \sum_{i=1}^n \rho_i(x_i, x_i'), \qquad x, x' \in \mathcal{X}^n. \tag{3}$$

We will denote partial products by

$$\mathcal{X}_i^j = \mathcal{X}_i \times \mathcal{X}_{i+1} \times \ldots \times \mathcal{X}_j.$$

We write $X_i \sim \mu_i$ to mean that X_i is an $\mathcal{X}_i$-valued random variable with law μ_i — i.e., $\mathbb{P}(X_i \in A) = \mu_i(A)$ for all Borel $A \subset \mathcal{X}_i$. This notation extends naturally to sequences: $X_1^n \sim \mu^n$. We will associate to each $(\mathcal{X}_i, \rho_i, \mu_i)$ the *symmetrized distance* random variable $\Xi(\mathcal{X}_i)$ defined by

$$\Xi(\mathcal{X}_i) = \epsilon_i \rho_i(X_i, X_i'), \tag{4}$$

where $X_i, X_i' \sim \mu_i$ are independent and $\epsilon_i = \pm 1$ with probability $1/2$, independent of X_i, X_i'. We note right away that $\Xi(\mathcal{X}_i)$ is a *centered* random variable:

$$\mathbb{E}[\Xi(\mathcal{X}_i)] = 0. \tag{5}$$

A real-valued random variable X is said to be *subgaussian* if it admits a $\sigma > 0$ such that

$$\mathbb{E}e^{\lambda X} \leq e^{\sigma^2 \lambda^2 / 2}, \qquad \lambda \in \mathbb{R}. \tag{6}$$

The smallest σ for which (6) holds will be denoted by $\sigma^*(X)$. We define the *subgaussian diameter* $\Delta_{\text{SG}}(\mathcal{X}_i)$ of the metric probability space $(\mathcal{X}_i, \rho_i, \mu_i)$ in terms of its symmetrized distance $\Xi(\mathcal{X}_i)$:

$$\Delta_{\text{SG}}(\mathcal{X}_i) = \sigma^*(\Xi(\mathcal{X}_i)). \tag{7}$$

In words, $\Xi(\mathcal{X}_i)$ is the signed distance between two points independently drawn from $\mathcal{X}_i$ and σ^* is the subgaussian moment of that random variable. If a metric probability space $(\mathcal{X}, \rho, \mu)$ has finite diameter,

$$\text{diam}(\mathcal{X}) := \sup_{x, x' \in \mathcal{X}} \rho(x, x') < \infty,$$

then its subgaussian diameter is also finite:

$$\Delta_{\text{SG}}(\mathcal{X}) \leq \text{diam}(\mathcal{X}). \tag{8}$$

The bound in (8) is nearly tight in the sense that for every $\varepsilon > 0$ there is a metric probability space $(\mathcal{X}, \rho, \mu)$ for which

$$\text{diam}(\mathcal{X}) < \Delta_{\text{SG}}(\mathcal{X}) + \varepsilon \tag{9}$$

(see the Appendix for proofs of (8) and (9), and related discussion).

On the other hand, there exist unbounded metric probability spaces with finite subgaussian diameter. A simple example is $(\mathcal{X}, \rho, \mu)$ with $\mathcal{X} = \mathbb{R}$, $\rho(x, x') = |x - x'|$ and μ the standard Gaussian probability measure $d\mu = (2\pi)^{-1/2} e^{-x^2/2} dx$. Obviously, $\text{diam}(\mathcal{X}) = \infty$. Now the symmetrized distance $\Xi = \Xi(\mathcal{X})$ is distributed as the difference (=sum) of two standard Gaussians: $\Xi \sim N(0, 2)$. Since $\mathbb{E}e^{\lambda \Xi} = e^{\lambda^2}$, we have

$$\Delta_{\text{SG}}(\mathcal{X}) = \sqrt{2}. \tag{10}$$

More generally, the subgaussian distributions on $\mathbb{R}$ are precisely those for which $\Delta_{\text{SG}}(\mathbb{R}) < \infty$.

3 Related work

McDiarmid's inequality (1) suffers from the limitations mentioned above: it completely ignores the distribution and is vacuous if even one of the w_i is infinite.[1] In order to address some of these issues, Kutin (2002); Kutin and Niyogi (2002) proposed an extension of McDiarmid's inequality to "almost everywhere" Lipschitz functions $\varphi : \mathcal{X}^n \to \mathbb{R}$. To formalize this, fix an $i \in [n]$ and let $X_1^n \sim \mu^n$ and $X_i' \sim \mu_i$ be independent. Define $\tilde{X}_1^n = \tilde{X}_1^n(i)$ by

$$\tilde{X}_j(i) = \begin{cases} X_j, & j \neq i \\ X_i', & j = i. \end{cases} \tag{11}$$

Kutin and Niyogi define φ to be *weakly difference-bounded* by (b, c, δ) if

$$\mathbb{P}\Big(|\varphi(X) - \varphi(\tilde{X}(i))| > b\Big) = 0 \tag{12}$$

and

$$\mathbb{P}\Big(|\varphi(X) - \varphi(\tilde{X}(i))| > c\Big) < \delta \tag{13}$$

for all $1 \leq i \leq n$. The precise result of Kutin (2002, Theorem 1.10) is somewhat unwieldy to state — indeed, the present work was motivated in part by a desire for simpler tools. Assuming that φ is weakly difference-bounded by (b, c, δ) with

$$\delta = \exp(-\Omega(n)) \tag{14}$$

and $c = O(1/n)$, their bound states that

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| \geq t) \leq \exp(-\Omega(nt^2)) \tag{15}$$

for a certain range of t and n. As noted by Rakhlin et al. (2005), the exponential decay assumption (14) is necessary in order for the Kutin-Niyogi method to yield exponential concentration. In contrast, the bounds we prove here

(i) do not require $|\varphi(X) - \varphi(\tilde{X})|$ to be everywhere bounded as in (12)

(ii) have a simple statement and proof, and generalize to non-iid processes with relative ease.

We defer the quantitative comparisons between (15) and our results until the latter are formally stated in Section 4.

The entropy method (Boucheron et al., 2003) may also be used to obtain concentration for unbounded functions but typically requires more detailed structural information. In a different line of work, Antonov (1979) gave inequalities for sums of independent random variables in the Orlicz spaces; these were recently improved by Rio (2013b). Bentkus (2008) considered an extension of

[1] Note, though, that McDiarmid's inequality is sharp in the sense that the constants in (1) cannot be improved in a distribution-free fashion.

Hoeffding's inequality to unbounded random variables. Rio (2013a) gave an L_p extension of McDiarmid's inequality. Kim and Vu (2000); Vu (2002) gave concentration inequalities for some classes of non-Lipschitz functions. An earlier notion of "effective" metric diameter in the context of concentration is that of *metric space length* (Schechtman, 1982). Another distribution-dependent refinement of diameter is the *spread constant* (Alon et al., 1998). Lecué and Mendelson (2013) gave minimax bounds for empirical risk minimization over subgaussian classes. More recently, Mendelson (2014) presented a framework for learning without concentration, which allows for unbounded loss functions and fat-tailed distributions. Cortes et al. (2013) gave relative bounds for unbound losses under moment assumptions. A result of van de Geer and Lederer (2013) interpolates between subgaussian and subexponential tails via a new Orlicz-type norm. Perhaps closest in spirit to the present work is the paper of Meir and Zhang (2003), whose Theorem 3 essentially expresses a subgaussian condition.

4 Concentration via subgaussian diameter

McDiarmid's inequality (1) may be stated in the notation of Section 2 as follows. Let $(\mathcal{X}_i, \rho_i, \mu_i)$, $i = 1, \ldots, n$ be a sequence of metric probability spaces and $\varphi : \mathcal{X}^n \to \mathbb{R}$ a 1-Lipschitz function. Then

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n \operatorname{diam}(\mathcal{X}_i)^2}\right) \tag{16}$$

(the equivalence of (1) and (16) is proved in the Appendix). We defined the subgaussian diameter $\Delta_{\text{SG}}(\mathcal{X}_i)$ in Section 2, having shown in (9) that it never exceeds the metric diameter. We also showed by example that the former can be finite when the latter is infinite. The main result of this section is that $\operatorname{diam}(\mathcal{X}_i)$ in (16) can essentially be replaced by $\Delta_{\text{SG}}(\mathcal{X}_i)$:

Theorem 1. *If $\varphi : \mathcal{X}^n \to \mathbb{R}$ is 1-Lipschitz and $\Delta_{\text{SG}}(\mathcal{X}_i) < \infty$ for all $i \in [n]$, then $\mathbb{E}\varphi < \infty$ and*

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > t) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^n \Delta_{\text{SG}}^2(\mathcal{X}_i)}\right).$$

Our constant in the exponent is worse than that of (16) by a factor of 4. This appears to be an inherent artifact of our method, and we do not know whether (16) holds verbatim with $\operatorname{diam}(\mathcal{X}_i)$ be replaced by $\Delta_{\text{SG}}(\mathcal{X}_i)$.

Proof. The strong integrability of φ — and in particular, finiteness of $\mathbb{E}\varphi$ — follow from exponential concentration (Ledoux, 2001). The rest of the proof will proceed via the Azuma-Hoeffding-McDiarmid method of martingale differences.

Define $V_i = \mathbb{E}[\varphi \,|\, X_1^i] - \mathbb{E}[\varphi \,|\, X_1^{i-1}]$ and expand

$$\begin{aligned}\mathbb{E}[\varphi \,|\, X_1^i] &= \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} \mathbb{P}(x_{i+1}^n)\varphi(X_1^i x_{i+1}^n) \\ \mathbb{E}[\varphi \,|\, X_1^{i-1}] &= \sum_{x_i^n \in \mathcal{X}_i^n} \mathbb{P}(x_i^n)\varphi(X_1^{i-1} x_i^n).\end{aligned}$$

Let us write $\tilde{V}_i$ to denote V_i as a function of X_1^{i-1} with X_i integrated out:

$$\tilde{V}_i = \sum_{x_{i+1}^n} \mathbb{P}(x_{i+1}^n) \sum_{x_i, x_i'} \mathbb{P}(x_i)\mathbb{P}(x_i') \left(\varphi(X_1^{i-1} x_i x_{i+1}^n) - \varphi(X_1^{i-1} x_i' x_{i+1}^n)\right).$$

Hence, by Jensen's inequality, we have

$$\mathbb{E}[e^{\lambda V_i} \,|\, X_1^{i-1}] \le \sum_{x_{i+1}^n} \mathbb{P}(x_{i+1}^n) \sum_{y,y'} \mathbb{P}(y)\mathbb{P}(y') e^{\lambda(\varphi(X_1^{i-1} y x_{i+1}^n) - \varphi(X_1^{i-1} y' x_{i+1}^n))}.$$

For fixed $X_1^{i-1} \in \mathcal{X}_1^{i-1}$ and $x_{i+1}^n \in \mathcal{X}_{i+1}^n$, define $F : \mathcal{X}_i \to \mathbb{R}$ by $F(y) = \varphi(X_1^{i-1} y x_{i+1}^n)$, and observe that F is 1-Lipschitz with respect to ρ_i. Since $e^t + e^{-t} = 2\cosh(t)$ and $\cosh(t) \le \cosh(s)$ for all $|t| \le s$, we have[2]

$$e^{\lambda(F(y) - F(y'))} + e^{\lambda(F(y') - F(y))} \le e^{\lambda \rho_i(y,y')} + e^{-\lambda \rho_i(y,y')}.$$

Now for every term in the sum of the form $\exp(\lambda(F(y) - F(y')))$ there is a matching term with the opposite sign in the exponent, and hence

$$\begin{aligned}&\sum_{y,y' \in \mathcal{X}_i} \mathbb{P}(y)\mathbb{P}(y') e^{\lambda(F(y) - F(y'))} \qquad (17) \\ &\le \tfrac{1}{2} \left[\sum_{y,y'} \mathbb{P}(y)\mathbb{P}(y') e^{\lambda \rho_i(y,y')} + \sum_{y,y'} \mathbb{P}(y)\mathbb{P}(y') e^{-\lambda \rho_i(y,y')}\right] \\ &= \mathbb{E} e^{\lambda \Xi(\mathcal{X}_i)} \le \exp(\lambda^2 \Delta_{\mathrm{SG}}^2(\mathcal{X}_i)/2),\end{aligned}$$

where $\Xi(\mathcal{X}_i)$ is the symmetrized distance (4) and the last inequality holds by definition of subgaussian diameter (6,7). It follows that

$$\mathbb{E}[e^{\lambda V_i} \,|\, X_1^{i-1}] \le \exp(\lambda^2 \Delta_{\mathrm{SG}}^2(\mathcal{X}_i)/2). \qquad (18)$$

Applying the standard Markov's inequality and exponential bounding argument,

[2] An analogous symmetrization technique is employed in Tao (2009) as a variant of the "square and rearrange" trick.

we have

$$\begin{aligned}
\mathbb{P}(\varphi - \mathbb{E}\varphi > t) &= \mathbb{P}\left(\sum_{i=1}^{n} V_i > t\right) \\
&\leq e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda V_i}\right] \\
&= e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^{n} \mathbb{E}[e^{\lambda V_i} \mid X_1^{i-1}]\right] \\
&\leq e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^{n} \exp(\lambda^2 \Delta_{\mathrm{SG}}^2(\mathcal{X}_i)/2)\right] \\
&= \exp\left(\frac{1}{2}\lambda^2 \sum_{i=1}^{n} \Delta_{\mathrm{SG}}^2(\mathcal{X}_i) - \lambda t\right). \qquad (19)
\end{aligned}$$

Optimizing over λ and applying the same argument to $\mathbb{E}\varphi - \varphi$ yields our claim. □

Let us see how Theorem 1 compares to previous results on some examples. Consider $\mathbb{R}^n$ equipped with the ℓ_1 metric $\rho^n(x, x') = \sum_{i\in[n]} |x_i - x_i'|$ and the standard Gaussian product measure $\mu^n = N(0, I_n)$. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be $1/n$-Lipschitz. Then Theorem 1 yields (recalling the calculation in (10))

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > \varepsilon) \leq 2\exp(-n\varepsilon^2/4), \qquad \varepsilon > 0, \qquad (20)$$

whereas the inequalities of McDiarmid (1) and Kutin-Niyogi (15) are both uninformative since the metric diameter is infinite.

For our next example, fix an $n \in \mathbb{N}$ and put $\mathcal{X}_i = \{\pm 1, \pm n\}$ with the metric $\rho_i(x, x') = |x - x'|$ and the distribution $\mu_i(x) \propto e^{-x^2}$. One may verify via a calculation analogous to (10) that $\Delta_{\mathrm{SG}}(\mathcal{X}_i) \leq \sqrt{2}$. For independent $X_i \sim \mu_i$, $i = 1, \ldots, n$, put $\varphi(X_1^n) = n^{-1}\sum_{i=1}^n X_i$. Then Theorem 1 implies that in this case the bound in (20) holds verbatim. On the other hand, φ is easily seen to be weakly difference-bounded by $(1, 1/n, e^{-\Omega(n)})$ and thus (15) also yields subgaussian concentration, albeit with worse constants. Applying (1) yields the much cruder estimate

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > \varepsilon) \leq 2\exp(-2\varepsilon^2).$$

5 Application to algorithmic stability

We refer the reader to (Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002; Rakhlin et al., 2005) for background on algorithmic stability and supervised learning. Our metric probability space $(\mathcal{Z}_i, \rho_i, \mu_i)$ will now have the structure $\mathcal{Z}_i = \mathcal{X}_i \times \mathcal{Y}_i$ where $\mathcal{X}_i$ and $\mathcal{Y}_i$ are, respectively, the *instance* and *label* space of the i^{th} example. Under the iid assumption, the $(\mathcal{Z}_i, \rho_i, \mu_i)$ are identical for all

$i \in \mathbb{N}$ (and so we will henceforth drop the subscript i from these). A training sample $S = Z_1^n \sim \mu^n$ is drawn and a *learning algorithm* $\mathcal{A}$ inputs S and outputs a *hypothesis* $f : \mathcal{X} \to \mathcal{Y}$. The hypothesis $f = \mathcal{A}(S)$ will be denoted by $\mathcal{A}_S$. In line with the previous literature, we assume that $\mathcal{A}$ is symmetric (i.e., invariant under permutations of S). The *loss* of a hypothesis f on an example $z = (x, y)$ is defined by

$$L(f, z) = \ell(f(x), y),$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is the *cost function*. To our knowledge, all previous work required the loss to be bounded by some constant $M < \infty$, which figures explicitly in the bounds; we make no such restriction. In the algorithmic stability setting, the *empirical risk* $\hat{R}_n(\mathcal{A}, S)$ is typically defined as

$$\hat{R}_n(\mathcal{A}, S) = \frac{1}{n} \sum_{i=1}^{n} L(\mathcal{A}_S, z_i) \tag{21}$$

and the *true risk* $R(\mathcal{A}, S)$ as

$$R(\mathcal{A}, S) = \mathbb{E}_{z \sim \mu}[L(\mathcal{A}_S, z)]. \tag{22}$$

The goal is to bound the true risk in terms of the empirical one. To this end, a myriad of notions of hypothesis stability have been proposed. A variant of *uniform stability* in the sense of Rakhlin et al. (2005) — which is slightly more general than the homonymous notion in Bousquet and Elisseeff (2002) — may be defined as follows. The algorithm $\mathcal{A}$ is said to be β-uniform stable if for all $\tilde{z} \in \mathcal{Z}$, the function $\varphi_{\tilde{z}} : \mathcal{Z}^n \to \mathbb{R}$ given by $\varphi_{\tilde{z}}(z) = L(\mathcal{A}_z, \tilde{z})$ is β-Lipschitz with respect to the Hamming metric on $\mathcal{Z}^n$:

$$\forall \tilde{z} \in \mathcal{Z}, \forall z, z' \in \mathcal{Z}^n : \ |\varphi_{\tilde{z}}(z) - \varphi_{\tilde{z}}(z')| \leq \beta \sum_{i=1}^{n} \mathbb{1}_{\{z_i \neq z_i'\}}.$$

We define the algorithm $\mathcal{A}$ to be *β-totally Lipschitz stable* if the function $\varphi : \mathcal{Z}^{n+1} \to \mathbb{R}$ given by $\varphi(z_1^{n+1}) = L(\mathcal{A}_{z_1^n}, z_{n+1})$ is β-Lipschitz with respect to the ℓ_1 product metric on $\mathcal{Z}^{n+1}$:

$$\forall z, z' \in \mathcal{Z}^{n+1} : \ |\varphi(z) - \varphi(z')| \leq \beta \sum_{i=1}^{n+1} \rho(z_i, z_i'). \tag{23}$$

Note that total Lipschitz stability is stronger than uniform stability since it requires the algorithm to respect the metric of $\mathcal{Z}$.

Let us bound the bias of stable algorithms.

Lemma 1. *Suppose $\mathcal{A}$ is a symmetric, β-totally Lipschitz stable learning algorithm over the metric probability space $(\mathcal{Z}, \rho, \mu)$ with $\Delta_{\text{SG}}(\mathcal{Z}) < \infty$. Then*

$$\mathbb{E}[R(\mathcal{A}, S) - \hat{R}_n(\mathcal{A}, S)] \quad \leq \quad \tfrac{1}{2}\beta^2 \Delta_{\text{SG}}^2(\mathcal{Z}).$$

We now turn to Lipschitz continuity.

Lemma 2. *Suppose $\mathcal{A}$ is a symmetric, β-totally Lipschitz stable learning algorithm and define the function $\varphi : \mathcal{Z}^n \to \mathbb{R}$ by $\varphi(z) = R(\mathcal{A}, z) - \hat{R}_n(\mathcal{A}, z)$. Then φ is 3β-Lipschitz.*

Combining Lemmas 1 and 2 with our concentration inequality in Theorem 1 yields the main result of this section:

Theorem 2. *Suppose $\mathcal{A}$ is a symmetric, β-totally Lipschitz stable learning algorithm over the metric probability space $(\mathcal{Z}, \rho, \mu)$ with $\Delta_{\mathrm{SG}}(\mathcal{Z}) < \infty$. Then, for training samples $S \sim \mu^n$ and $\varepsilon > 0$, we have*

$$\mathbb{P}\Big(R(\mathcal{A}, S) - \hat{R}_n(\mathcal{A}, S) > \tfrac{1}{2}\beta^2\Delta_{\mathrm{SG}}^2(\mathcal{Z}) + \varepsilon\Big) \leq \exp\left(-\frac{\varepsilon^2}{18\beta^2\Delta_{\mathrm{SG}}^2(\mathcal{Z})n}\right).$$

As in Bousquet and Elisseeff (2002) and related results on algorithmic stability, we require $\beta = o(n^{-1/2})$ for nontrivial decay. Bousquet and Elisseeff showed that this is indeed the case for some popular learning algorithms, albeit in their less restrictive definition of stability. Below we show that a natural metric regression algorithm is stable in our stronger sense.

5.1 Stability of regularized nearest-neighbor regression

In the *regression* setting, we take the label space $\mathcal{Y}$ to be all of $\mathbb{R}$ (and note that many existing approaches require $\mathcal{Y}$ to be a compact subst of $\mathbb{R}$). Gottlieb et al. (2013) proposed an efficient algorithm for regression in general metric spaces via Lipschitz extension, which is algorithmically realized by 1-nearest neighbors. Aside from efficiency, the nearest-neighbor approach also facilitates risk analysis. To any metric space $(\mathcal{X}, \rho)$ we associate the metric space $(\mathcal{Z}, \bar{\rho})$, where $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and $\bar{\rho}((x, y), (x', y')) = \rho(x, x') + |y - y'|$. Suppose that $(\mathcal{Z}, \bar{\rho})$ is endowed with a measure μ such that $\Delta_{\mathrm{SG}} = \Delta_{\mathrm{SG}}(\mathcal{Z}, \bar{\rho}, \mu) < \infty$. Write $\mathcal{F}_\lambda$ to denote the collection of all λ-Lipschitz hypotheses $f : \mathcal{X} \to \mathbb{R}$. The learning algorithm $\mathcal{A}$ maps the sample $S = Z_{i\in[n]}$, with $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, to the hypothesis $\hat{f} \in \mathcal{F}_\lambda$ by minimizing the empirical risk

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|$$

over all $f \in \mathcal{F}_\lambda$, where we have chosen the absolute loss $\ell(y, y') = |y - y'|$. We will give a heuristic argument for the stability of 1-NN regression regularized by Lipschitz continuity λ. This argument will be fleshed out formally in the full version of the paper. Since the value of a Lipschitz extension at a point is determined by its nearest neighbors (Bousquet and Elisseeff, 2002), it suffices to ensure that none of the $n + 1$ points (n sample and 1 test) is too isolated from the rest. The subgaussian assumption implies (see Rivasplata (2012, Theorem 3.1)) that with probability $1 - n\exp(-\Omega(n))$, each of the $n + 1$ points is within distance $O(\Delta_{\mathrm{SG}})$ of another point. Since a λ-Lipschitz function can vary by at

most $O(\lambda D)$ over a ball of diameter D, this implies that the regression algorithm is $\beta = O(\lambda \Delta_{\mathrm{SG}}/n)$-stable. Thus, Theorem 2 yields the risk bound

$$\begin{aligned}&\mathbb{P}\Big(R(\mathcal{A}, S) - \hat{R}_n(\mathcal{A}, S) > (\lambda \Delta_{\mathrm{SG}}/n)^2 + \varepsilon\Big)\\ &\leq \exp\left(-\Omega\left(\frac{\varepsilon^2 n}{\lambda^2 \Delta_{\mathrm{SG}}^2}\right)\right) + n\exp(-\Omega(n)).\end{aligned}$$

Note that the subgaussian assumption implies risk bounds not depending on any dimensions (doubling or otherwise) of the metric space (cf. Gottlieb et al. (2013)).

6 Relaxing the independence assumption

In this section we generalize Theorem 1 to strongly mixing processes. To this end, we require some standard facts concerning the probability-theoretic notions of *coupling* and *transportation* (Lindvall, 2002; Villani, 2003, 2009). Given the probability measures μ, μ' on a measurable space $\mathcal{X}$, a coupling π of μ, μ' is any probability measure on $\mathcal{X} \times \mathcal{X}$ with marginals μ and μ', respectively. Denoting by $\Pi = \Pi(\mu, \mu')$ the set of all couplings, we have

$$\begin{aligned}\inf_{\pi \in \Pi} \pi(\big\{(x, y) \in \mathcal{X}^2 : x \neq y\big\}) &= \tfrac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \mu'(x)|\\ &= \|\mu - \mu'\|_{\mathrm{TV}} \qquad (24)\end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the *total variation* norm. An *optimal* coupling is one that achieves the infimum in (24); one always exists, though it may not be unique. Another elementary property of couplings is that for any two $f, g : \mathcal{X} \to \mathbb{R}$ and any coupling $\pi \in \Pi(\mu, \mu')$, we have

$$\mathbb{E}_{\mu} f - \mathbb{E}_{\mu'} g = \mathbb{E}_{(X, X') \sim \pi}[f(X) - g(X')]. \qquad (25)$$

It is possible to refine the total variation distance (24) between μ and μ' so as to respect the metric of $\mathcal{X}$. Given a space equipped with probability measures μ, μ' and metric ρ, define the *transportation cost*[3] *distance* $T_\rho(\mu, \mu')$ by

$$T_\rho(\mu, \mu') = \inf_{\pi \in \Pi(\mu, \mu')} \mathbb{E}_{(X, X') \sim \pi} \rho(X, X').$$

It is easy to verify that T_ρ is a valid metric on probability measures and that for $\rho(x, x') = \mathbb{1}_{\{x \neq x'\}}$, we have $T_\rho(\mu, \mu') = \|\mu - \mu'\|_{\mathrm{TV}}$.

As in Section 4, we consider a sequence of metric spaces $(\mathcal{X}_i, \rho_i)$, $i = 1, \ldots, n$ and their ℓ_1 product $(\mathcal{X}^n, \rho^n)$. Unlike the independent case, we will allow non-product probability measures ν on $(\mathcal{X}^n, \rho^n)$. We will write $X_1^n \sim \nu$ to mean

[3] This fundamental notion is also known as the *Wasserstein, Monge-Kantorovich*, or *earth-mover* distance; see Villani (2003, 2009) for an encyclopedic treatment. The use of coupling and transportation techniques to obtain concentration for dependent random variables goes back to Marton (1996); Samson (2000); Chazottes et al. (2007).

that $\mathbb{P}(X_1^n \in A) = \nu(A)$ for all Borel $A \subset \mathcal{X}^n$. For $1 \le i \le j < k \le l \le n$, we will use the shorthand

$$\mathbb{P}(x_k^l \,|\, x_i^j) = \mathbb{P}\Big(X_k^l = x_k^l \,|\, X_i^j = x_i^j\Big).$$

The notation $\mathbb{P}(X_i^j)$ means the marginal distribution of X_i^j. Similarly, $\mathbb{P}(X_k^l \,|\, X_i^j = x_i^j)$ will denote the conditional distribution. For $1 \le i < n$, and $x_1^i \in \mathcal{X}_1^i$, $x_i' \in \mathcal{X}_i$ define

$$\tau_i(x_1^i, x_i') = T_{\rho_{i+1}^n}(\mathbb{P}(X_{i+1}^n \,|\, X_1^i = x_1^i), \mathbb{P}(X_{i+1}^n \,|\, X_1^i = x_1^{i-1} x_i')),$$

where ρ_{i+1}^n is the ℓ_1 product of $\rho_{i+1}, \dots \rho_n$ as in (3), and

$$\bar{\tau}_i = \sup_{x_1^i \in \mathcal{X}_1^i, x_i' \in \mathcal{X}_i} \tau_i(x_1^i, x_i'),$$

with $\bar{\tau}_n \equiv 0$. In words, $\tau_i(x_1^i, x_i')$ measures the transportation cost distance between the conditional distributions induced on the "tail" $\mathcal{X}_{i+1}^n$ given two prefixes that differ in the i^{th} coordinate, and $\bar{\tau}_i$ is the maximal value of this quantity. Kontorovich (2007); Kontorovich and Ramanan (2008) discuss how to handle conditioning on measure-zero sets and other technicalities. Note that for product measures the conditional distributions are identical and hence $\bar{\tau}_i = 0$.

We need one more definition before stating our main result. For the prefix x_1^{i-1}, define the conditional distribution

$$\nu_i(x_1^{i-1}) = \mathbb{P}\big(X_i \,|\, X_1^{i-1} = x_1^{i-1}\big)$$

and consider the corresponding metric probability space $(\mathcal{X}_i, \rho_i, \nu_i(x_1^{i-1}))$. Define its *conditional subgaussian diameter* by

$$\Delta_{\text{SG}}(\mathcal{X}_i \,|\, x_1^{i-1}) = \Delta_{\text{SG}}(\mathcal{X}_i, \rho_i, \nu_i(x_1^{i-1}))$$

and the *maximal subgaussian diameter* by

$$\bar{\Delta}_{\text{SG}}(\mathcal{X}_i) = \sup_{x_1^{i-1} \in \mathcal{X}_1^{i-1}} \Delta_{\text{SG}}(\mathcal{X}_i \,|\, x_1^{i-1}). \tag{26}$$

Note that for product measures, (26) reduces to the former definition (7). With these definitions, we may state the main result of this section.

Theorem 3. *If $\varphi : \mathcal{X}^n \to \mathbb{R}$ is 1-Lipschitz with respect to ρ^n, then*

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > t) \le 2 \exp\left(-\frac{(t - \sum_{i \le n} \bar{\tau}_i)^2}{2 \sum_{i \le n} \bar{\Delta}_{\text{SG}}^2(\mathcal{X}_i)}\right), \quad t > 0.$$

Observe that we recover Theorem 1 as a special case. Since typically we will take $t = \varepsilon n$, it suffices that $\sum_{i \le n} \bar{\tau}_i = o(n)$ and $\sum_{i \le n} \bar{\Delta}_{\text{SG}}^2(\mathcal{X}_i) = O(n)$ to ensure an exponential bound with decay rate $\exp(-\Omega(n\varepsilon^2))$. (Note that our functions are scaled to be $O(1)$-Lipschitz as opposed to the usual $O(1/n)$-Lipschiz condition.) The appearance of the mixing coefficients $\bar{\tau}_i$ in the numerator is nonstandard, and is mainly an artifact of our inability to obtain nontrivial bounds on this quantity. Elucidating its structure is an active research direction.

7 Other Orlicz diameters

Let us recall the notion of an *Orlicz norm* $\|X\|_\Psi$ of a real random variable X (see, e.g., Rao and Ren (1991)):

$$\|X\|_\Psi = \inf\left\{t > 0 : \mathbb{E}[\Psi(X/t)] \leq 1\right\},$$

where $\Psi : \mathbb{R} \to \mathbb{R}$ is a *Young function* — i.e., nonnegative, even, convex and vanishing at 0. In this section, we will consider the Young functions

$$\psi_p(x) = e^{|x|^p} - 1, \qquad p > 1,$$

and their induced Orlicz norms. A random variable X is subgaussian if and only if $\|X\|_{\psi_2} < \infty$ (Rivasplata, 2012). For $p \neq 2$, $\|X\|_{\psi_p} < \infty$ implies that

$$\mathbb{E}e^{\lambda X} \leq e^{(a|\lambda|)^p/p}, \qquad \lambda \in \mathbb{R}, \tag{27}$$

for some $a > 0$, but the converse implication need not hold. An immediate consequence of Markov's inequality is that any X for which (27) holds also satisfies

$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\frac{p-1}{p}\left(\frac{t}{a}\right)^{p/(p-1)}\right). \tag{28}$$

We define the *p-Orlicz diameter* of a metric probability space $(\mathcal{X}, \rho, \mu)$, denoted $\Delta_{\mathrm{OR}(p)}(\mathcal{X})$, as the smallest $a > 0$ that verifies (27) for the symmetrized distance $\Xi(\mathcal{X})$. In light of (28), Theorem 1 extends straightforwardly to finite p-Orlicz metric diameters:

Theorem 4. *Let $(\mathcal{X}_i, \rho_i, \mu_i)$, $i = 1, \ldots, n$ be a sequence of metric probability spaces and equip $\mathcal{X}^n$ with the usual product measure μ^n and ℓ_1 product metric ρ^n. Suppose that for some $p > 1$ and all $i \in [n]$ we have $\Delta_{\mathrm{OR}(p)}(\mathcal{X}_i) < \infty$, and define the vector $\boldsymbol{\Delta} \in \mathbb{R}^n$ by $\Delta_i = \Delta_{\mathrm{OR}(p)}(\mathcal{X}_i)$. If $\varphi : \mathcal{X}^n \to \mathbb{R}$ is 1-Lipschitz then for all $t > 0$,*

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > t) \leq 2\exp\left(-\frac{p-1}{p}\left(\frac{t}{\|\boldsymbol{\Delta}\|_p}\right)^{p/(p-1)}\right).$$

8 Discussion

We have given a concentration inequality for metric spaces with unbounded diameter, showed its applicability to algorithmic stability with unbounded losses, and gave an extension to non-independent sampling processes. Some fascinating questions remain:

(i) How tight is Theorem 1? First there is the vexing matter of having a worse constant in the exponent (i.e., 1/2) than McDiarmid's (optimal)

constant 2. Although this gap is not of critical importance, one would like a bound that recovers McDiarmid's in the finite-diameter case. More importantly, is it the case that finite subgaussian diameter is necessary for subgaussian concentration of all Lipschitz functions? That is, given the metric probability spaces $(\mathcal{X}_i, \rho_i, \mu_i)$, $i \in [n]$, can one always exhibit a 1-Lipschitz $\varphi : \mathcal{X}^n \to \mathbb{R}$ that achieves a nearly matching lower bound?

(ii) We would like to better understand how Theorem 1 compares to the Kutin-Niyogi bound (15). We conjecture that for any $(\mathcal{X}^n, \mu^n)$ and $\varphi : \mathcal{X}^n \to \mathbb{R}$ that satisfies (12) and (13), one can construct a product metric ρ^n for which $\sum_{i\in[n]} \Delta_{\mathrm{SG}}^2(\mathcal{X}_i) < \infty$ and φ is 1-Lipschitz. This would imply that whenever the Kutin-Niyogi bound is nontrivial, so is Theorem 1. We have already shown by example (20) that the reverse does not hold.

(iii) The quantity $\bar{\tau}_i$ defined in Section 6 is a rather complicated object; one desires a better handle on it in terms of the given distribution and metric.

(iv) We have argued in Section 5.1 that a natural regularized metric regression is totally Lipschitz stable under our definition (23). The next order of business would be to show that some other common learning algorithms, such as kernel SVM, are also stable in our strong sense.

Acknowledgements

John Lafferty encouraged me to seek a distribution-dependent refinement of McDiarmid's inequality. Thanks also to Gideon Schechtman, Shahar Mendelson, Assaf Naor, Iosif Pinelis and Csaba Szepesvári for helpful correspondence, and to Roi Weiss for carefully proofreading the manuscript.

References

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *J. Mach. Learn. Res.*, 10:441–474, June 2009. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1577069.1577085`.

N. Alon, R. Boppana, and J. Spencer. An asymptotic isoperimetric inequality. *Geometric & Functional Analysis GAFA*, 8(3):411–436, 1998. ISSN 1016-443X. doi: 10.1007/s000390050062. URL `http://dx.doi.org/10.1007/s000390050062`.

Sergei N. Antonov. Probability inequalities for a series of independent random variables. *Teor. Veroyatnost. i Primenen.*, 24(3):632–636, 1979. ISSN 0040-361X.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, pages 137–144, 2006.

Vidmantas Bentkus. An extension of the Hoeffding inequality to unbounded random variables. *Lith. Math. J.*, 48(2):137–157, 2008. ISSN 0363-1672. doi: 10.1007/s10986-008-9007-7. URL `http://dx.doi.org/10.1007/s10986-008-9007-7`.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614, 2003.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005. ISSN 1262-3318. doi: 10.1051/ps:2005018. URL `http://dx.doi.org/10.1051/ps:2005018`.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Jean-René Chazottes, Pierre Collet, Christof Külske, and Frank Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1-2):201–225, 2007.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theor. Comput. Sci.*, 519:103–126, 2014.

Corinna Cortes, Spencer Greenberg, and Mehryar Mohri. Relative deviation learning bounds and generalization with unbounded loss functions (arxiv:1310.5796). 2013.

Sanjoy Dasgupta and Philip M. Long. Boosting with diverse base classifiers. In *COLT*, pages 273–287, 2003.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

Miroslav Dudík, Robert E. Schapire, and Steven J. Phillips. Correcting sample selection bias in maximum entropy density estimation. In *NIPS*, 2005.

Ran El-Yaniv and Dmitry Pechyony. Stable transductive learning. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 35–49. Springer, Berlin, 2006. doi: 10.1007/11776420_6. URL `http://dx.doi.org/10.1007/11776420_6`.

David Gamarnik. Extension of the PAC framework to finite and countable Markov chains. *IEEE Trans. Inform. Theory*, 49(1):338–345, 2003.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate Lipschitz extension. In *SIMBAD*, pages 43–58, 2013.

Don Hush, Clint Scovel, and Ingo Steinwart. Stability of unstable learning algorithms. *Machine Learning*, 67(3):197–206, 2007. ISSN 0885-6125. doi: 10.1007/s10994-007-5004-z. URL `http://dx.doi.org/10.1007/s10994-007-5004-z`.

Rajeeva L. Karandikar and Mathukumalli Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Statist. Probab. Lett.*, 58(3):297–307, 2002. ISSN 0167-7152.

Jeong Han Kim and Van H. Vu. Concentration of multivariate polynomials and its applications. *Combinatorica*, 20(3):1439–6912, 2000.

Aryeh (Leonid) Kontorovich. *Measure Concentration of Strongly Mixing Processes with Applications*. PhD thesis, Carnegie Mellon University, 2007.

Leonid (Aryeh) Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.*, 36(6): 2126–2158, 2008.

Samuel Kutin. Extensions to McDiarmid's inequality when differences are bounded with high probability. Technical Report TR-2002-04, Department of Computer Science, University of Chicago, 2002.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *UAI*, pages 275–282, 2002.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds, arxiv:1305.4825. 2013.

Michel Ledoux. *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs Vol. 89*. American Mathematical Society, 2001.

Torgny Lindvall. *Lectures on the Coupling Method*. Dover Publications, 2002.

Ben London, Bert Huang, and Lise Getoor. Improved generalization bounds for large-scale structured prediction. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*, 2012.

Ben London, Bert Huang, Benjamin Taskar, and Lise Getoor. Collective stability in structured prediction: Generalization from one example. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

Katalin Marton. Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, 24(2):857–866, 1996.

Colin McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, volume 141 of LMS Lecture Notes Series*, pages 148–188. Morgan Kaufmann Publishers, San Mateo, CA, 1989.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *J. Mach. Learn. Res.*, 4:839–860, December 2003. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=945365.964297`.

Shahar Mendelson. Learning without concentration. In *COLT*, 2014.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary phi-mixing and beta-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Neural Information Processing Systems (NIPS)*, 2008.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006. ISSN 1019-7168. URL `http://dx.doi.org/10.1007/s10444-004-7634-z`.

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Anal. Appl. (Singap.)*, 3(4):397–417, 2005. ISSN 0219-5305. doi: 10.1142/S0219530505000650. URL http://dx.doi.org/10.1142/S0219530505000650.

Malempati Madhusudana Rao and Zhong Dao Ren. *Theory of Orlicz spaces*, volume 146 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker Inc., New York, 1991. ISBN 0-8247-8478-2.

Emmanuel Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(10):905–908, 2000.

Emmanuel Rio. On McDiarmid's concentration inequality. *Electron. Commun. Probab.*, 18:no. 44, 11, 2013a. ISSN 1083-589X. doi: 10.1214/ECP.v18-2659.

Emmanuel Rio. Extensions of the Hoeffding-Azuma inequalities. *Electron. Commun. Probab.*, 18:no. 54, 6, 2013b. ISSN 1083-589X. doi: 10.1214/ECP.v18-2690. URL http://dx.doi.org/10.1214/ECP.v18-2690.

Omar Rivasplata. Subgaussian random variables: An expository note. 2012. URL http://www.math.ualberta.ca/~orivasplata/publications/subgaussians.pdf.

Afshin Rostamizadeh and Mehryar Mohri. Stability bounds for non-i.i.d. processes. In *Neural Information Processing Systems (NIPS)*, 2007.

Benjamin I. P. Rubinstein and Aleksandr Simma. On the stability of empirical risk minimization in the presence of multiple risk minimizers. *Information Theory, IEEE Transactions on*, 58(7):4160–4163, 2012. ISSN 0018-9448. doi: 10.1109/TIT.2012.2191681.

Paul-Marie Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

Gideon Schechtman. Lévy type inequality for a class of finite metric spaces. In Jia-Arng Chao and Wojbor A. Woyczyński, editors, *Martingale Theory in Harmonic Analysis and Banach Spaces*, volume 939 of *Lecture Notes in Mathematics*, pages 211–215. Springer Berlin Heidelberg, 1982. ISBN 978-3-540-11569-4. doi: 10.1007/BFb0096270. URL http://dx.doi.org/10.1007/BFb0096270.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, 2010. ISSN 1532-4435.

Cosma Rohilla Shalizi and Aryeh Kontorovich. Predictive PAC learning and process decompositions. In *Neural Information Processing Systems (NIPS)*, 2013.

Ingo Steinwart and Andreas Christmann. Fast learning from non-i.i.d. observations. In *NIPS*, pages 1768–1776, 2009.

Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175 – 194, 2009. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2008.04.001. URL http://www.sciencedirect.com/science/article/pii/S0047259X08001097.

Terrence Tao. Talagrand's concentration inequality, 2009. URL http://terrytao.wordpress.com/2009/06/09/talagrands-concentration-inequality/.

Antonio Torralba, Robert Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.

Sara van de Geer and Johannes Lederer. The bernstein-orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1-2):225–250, 2013. ISSN 0178-8051. doi: 10.1007/s00440-012-0455-y. URL http://dx.doi.org/10.1007/s00440-012-0455-y.

Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 0-8218-3312-X.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3.

Van H. Vu. Concentration of non-Lipschitz functions and applications. *Random Struct. Algorithms*, 20(3):262–316, 2002. ISSN 1042-9832. doi: http://dx.doi.org/10.1002/rsa.10032.

Bin Zou, Zong-ben Xu, and Jie Xu. Generalization bounds of ERM algorithm with Markov chain samples. *Acta Mathematicae Applicatae Sinica (English Series)*, pages 1–16. ISSN 0168-9673. URL http://dx.doi.org/10.1007/s10255-011-0096-4. 10.1007/s10255-011-0096-4.