High-dimensional p-norms

Gérard Biau, David D.M. Mason

Gérard Biau, David D.M. Mason. High-dimensional p-norms. 2013. hal-00879436

HAL Id: hal-00879436
https://hal.science/hal-00879436

Preprint submitted on 3 Nov 2013

High-Dimensional p-Norms

Gérard Biau and David M. Mason

Abstract Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a $\mathbb{R}^d$-valued random vector with i.i.d. components, and let $\|\mathbf{X}\|_p = (\sum_{j=1}^d |X_j|^p)^{1/p}$ be its p-norm, for $p > 0$. The impact of letting d go to infinity on $\|\mathbf{X}\|_p$ has surprising consequences, which may dramatically affect high-dimensional data processing. This effect is usually referred to as the *distance concentration phenomenon* in the computational learning literature. Despite a growing interest in this important question, previous work has essentially characterized the problem in terms of numerical experiments and incomplete mathematical statements. In the present paper, we solidify some of the arguments which previously appeared in the literature and offer new insights into the phenomenon.

1 Introduction

In what follows, for $\mathbf{x} = (x_1, \ldots, x_d)$ a vector of $\mathbb{R}^d$ and $0 < p < \infty$, we set

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^d |x_j|^p \right)^{1/p}. \tag{1}$$

It is recalled that for $p \geq 1$, $\|.\|_p$ is a norm on $\mathbb{R}^d$ (the L^p-norm) but for $0 < p < 1$, the triangle inequality does not hold and $\|.\|_p$ is sometimes called a prenorm. In the sequel, we take the liberty to call p-norm a norm or prenorm of the form (1), with $p > 0$.

Gérard Biau

Université Pierre et Marie Curie, Ecole Normale Supérieure & Institut universitaire de France, e-mail: gerard.biau@upmc.fr

David M. Mason
University of Delaware, e-mail: davidm@udel.edu

Now, let $\mathbf{X} = (X_1, \ldots, X_d)$ be a $\mathbb{R}^d$-valued random vector with i.i.d. components. The study of the probabilistic properties of $\|\mathbf{X}\|_p$ as the dimension d tends to infinity has recently witnessed an important research effort in the computational learning community (see, e.g., François et al., 2007, for a review). This activity is easily explained by the central role played by the quantity $\|\mathbf{X}\|_p$ in the analysis of nearest neighbor search algorithms, which are currently widely used in data management and database mining. Indeed, finding the closest matching object in an L^p-sense is of significant importance for numerous applications, including pattern recognition, multimedia content retrieving (images, videos, etc.), data mining, fraud detection and DNA sequence analysis, just to name a few. Most of these real applications involve very high-dimensional data (for example, pictures taken by a standard camera consist of several million pixels) and the curse of dimensionality (when $d \to \infty$) tends to be a major obstacle in the development of nearest neighbor-based techniques.

The effect on $\|\mathbf{X}\|_p$ of letting d go large is usually referred to as the *distance concentration phenomenon* in the computational learning literature. It is in fact a quite vague term that encompasses several interpretations. For example, it has been observed by several authors (e.g., François et al., 2007) that, under appropriate moment assumptions, the so-called *relative standard deviation* $\sqrt{\mathrm{Var}\|\mathbf{X}\|_p}/\mathbb{E}\|\mathbf{X}\|_p$ tends to zero as d tends to infinity. Consequently, by Chebyshev's inequality (this will be rigorously established in Section 2), for all $\varepsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{\|\mathbf{X}\|_p}{\mathbb{E}\|\mathbf{X}\|_p} - 1\right| \geq \varepsilon\right\} \to 0, \text{ as } d \to \infty.$$

This simple result reveals that the relative error made as considering $\mathbb{E}\|\mathbf{X}\|_p$ instead of the random value $\|\mathbf{X}\|_p$ becomes asymptotically negligible. Therefore, high-dimensional vectors $\mathbf{X}$ appear to be distributed on a sphere of radius $\mathbb{E}\|\mathbf{X}\|_p$.

The distance concentration phenomenon is also often expressed by considering an i.i.d. $\mathbf{X}$ sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and observing that, under certain conditions, the *relative contrast*

$$\frac{\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_p - \min_{1 \leq i \leq n} \|\mathbf{X}_i\|_p}{\min_{1 \leq i \leq n} \|\mathbf{X}_i\|_p}$$

vanishes in probability as d tends to infinity, whereas the *contrast*

$$\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_p - \min_{1 \leq i \leq n} \|\mathbf{X}_i\|_p$$

behaves in expectation as $d^{1/p-1/2}$ (Beyer et al., 1999; Hinneburg et al., 2000; Aggarwal et al., 2001; Kabán, 2012). Thus, assuming that the query point is located at the origin, the ratio between the largest and smallest p-distances from the sample to the query point becomes negligible as the dimension increases, and all points seem to be located at approximately the same distance from the origin. This phenomenon may dramatically affect high-dimensional data processing, analysis, retrieval and indexing, insofar these procedures rely on some notion of p-norm. Accordingly, serious questions are raised as to the validity of many nearest neighbor search heuristics

in high dimension, a problem that can be further exacerbated by techniques that find approximate neighbors in order to improve algorithmic performance (Beyer et al., 1999).

Even if people have now a better understanding of the distance concentration phenomenon and its practical implications, it is however our belief that there is still a serious need to solidify its mathematical background. Indeed, previous work has essentially characterized the problem in terms of numerical experiments and (often) incomplete probabilistic statements, with missing assumptions and (sometimes) defective proofs. Thus, our objective in the present paper is to solidify some of the statements which previously appeared in the computational learning literature. We start in Section 2 by offering a thorough analysis of the behavior of the p-norm $\|\mathbf{X}\|_p$ (as a function of p and the properties of the distribution of $\mathbf{X}$) as $d \to \infty$. Section 3 is devoted to the investigation of some new asymptotic properties of the *contrast* $\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_p - \min_{1 \leq i \leq n} \|\mathbf{X}_i\|_p$, both as $d \to \infty$ and $n \to \infty$. For the sake of clarity, most technical proofs are gathered in Section 4.

2 Asymptotic behavior of p-norms

2.1 Consistency

Throughout the document, the notation $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{\mathscr{D}}$ stand for convergence in probability and in distribution, respectively. The notation $u_n = \mathrm{o}(v_n)$ and $u_n = \mathrm{O}(v_n)$ mean, respectively, that $u_n/v_n \to 0$ and $u_n \leq C v_n$ for some constant C, as $n \to \infty$. The symbols $\mathrm{o}_{\mathbb{P}}(v_n)$ and $\mathrm{O}_{\mathbb{P}}(v_n)$ denote, respectively, a sequence of random variables $\{Y_n\}_{n \geq 1}$ such that $Y_n/v_n \xrightarrow{\mathbb{P}} 0$ and Y_n/v_n is bounded in probability, as $n \to \infty$.

We start this section with a general proposition that plays a key role in the analysis.

Proposition 1. *Let $\{U_d\}_{d \geq 1}$ be a sequence of random variables such that $U_d \xrightarrow{\mathbb{P}} a$, and let φ be a real-valued measurable function which is continuous at a. Assume that:*

(i) φ is bounded on $[-M, M]$ for some $M > |a|$;
(ii) $\mathbb{E}|\varphi(U_d)| < \infty$ for all $d \geq 1$.

Then, as $d \to \infty$,

$$\mathbb{E}\varphi(U_d) \to \varphi(a)$$

if and only if

$$\mathbb{E}\left(\varphi(U_d)\mathbf{1}\left\{|U_d| > M\right\}\right) \to 0. \tag{2}$$

Proof. The proof is easy. Condition (i) and continuity of φ at a allow us to apply the bounded convergence theorem to get

$$\mathbb{E}\left(\varphi(U_d)\mathbf{1}\left\{|U_d| \leq M\right\}\right) \to \varphi(a).$$

Since

$$\mathbb{E}\varphi(U_d) = \mathbb{E}\left(\varphi(U_d)\mathbf{1}\{|U_d| \le M\}\right) + \mathbb{E}\left(\varphi(U_d)\mathbf{1}\{|U_d| > M\}\right),$$

the rest of the proof is obvious. □

We shall now specialize the result of Proposition 1 to the case when

$$U_d = d^{-1}\sum_{j=1}^{d} Y_j := \overline{Y}_d,$$

where $\{Y_j\}_{j\ge 1}$ is a sequence of i.i.d. Y random variables with finite mean μ. In this case, by the strong law of large numbers, $U_d \to \mu$ almost surely. The following lemma gives two sufficient conditions for (2) to hold when $U_d = \overline{Y}_d$.

Lemma 1. *let φ be a real-valued measurable function. Assume that one of the following two conditions is satisfied:*

Condition 1 *The function $|\varphi|$ is convex on $\mathbb{R}$ and $\mathbb{E}|\varphi(Y)| < \infty$.*

Condition 2 *For some $s > 1$,*

$$\limsup_{d\to\infty}\mathbb{E}\left|\varphi(\overline{Y}_d)\right|^s < \infty.$$

Then (2) is satisfied for the sequence $\{\overline{Y}_d\}_{d\ge 1}$ with $a = \mu$ and $M > |\mu|$.

Proof. Suppose that **Condition 1** is satisfied. Then note that by the convexity assumption

$$\mathbb{E}\left(\left|\varphi(\overline{Y}_d)\right|\mathbf{1}\{|\overline{Y}_d| > M\}\right) \le d^{-1}\sum_{j=1}^{d}\mathbb{E}\left(\left|\varphi(Y_j)\right|\mathbf{1}\{|\overline{Y}_d| > M\}\right)$$
$$= \mathbb{E}\left(\left|\varphi(Y)\right|\mathbf{1}\{|\overline{Y}_d| > M\}\right).$$

Since $M > |\mu|$, we conclude that with probability one, $|\varphi(Y)|\mathbf{1}\{|\overline{Y}_d| > M\} \to 0$. Also $|\varphi(Y)|\mathbf{1}\{|\overline{Y}_d| > M\} \le |\varphi(Y)|$. Therefore, by the dominated convergence theorem, (2) holds.

Next, notice by Hölder's inequality with $1/r = 1 - 1/s$ that

$$\mathbb{E}\left(\left|\varphi(\overline{Y}_d)\right|\mathbf{1}\{|\overline{Y}_d| > M\}\right) \le \left(\mathbb{E}\left|\varphi(\overline{Y}_d)\right|^s\right)^{1/s}\left(\mathbb{P}\{|\overline{Y}_d| > M\}\right)^{1/r}.$$

Since $\mathbb{P}\{|\overline{Y}_d| > M\} \to 0$, (2) immediately follows from **Condition 2**. □

Let us now return to the distance concentration problem, which has been discussed in the introduction. Recall that we denote by $\mathbf{X} = (X_1,\dots,X_d)$ a $\mathbb{R}^d$-valued random vector with i.i.d. X components. Whenever for $p > 0$ $\mathbb{E}|X|^p < \infty$, we set $\mu_p = \mathbb{E}|X|^p$. Also when $\mathrm{Var}|X|^p < \infty$, we shall write $\sigma_p^2 = \mathrm{Var}|X|^p$. Proposition 1 and Lemma 1 yield the following corollary:

Corollary 1. *Fix $p > 0$ and $r > 0$.*

(*i*) *Whenever $r/p < 1$ and $\mathbb{E}|X|^p < \infty$,*

$$\frac{\mathbb{E}\|\mathbf{X}\|_p^r}{d^{r/p}} \to \mu_p^{r/p}, \text{ as } d \to \infty,$$

whereas if $\mathbb{E}|X|^p = \infty$, then

$$\lim_{d \to \infty} \frac{\mathbb{E}\|\mathbf{X}\|_p^r}{d^{r/p}} = \infty.$$

(*ii*) *Whenever $r/p \geq 1$ and $\mathbb{E}|X|^r < \infty$,*

$$\frac{\mathbb{E}\|\mathbf{X}\|_p^r}{d^{r/p}} \to \mu_p^{r/p}, \text{ as } d \to \infty,$$

whereas if $\mathbb{E}|X|^r = \infty$, then, for all $d \geq 1$,

$$\frac{\mathbb{E}\|\mathbf{X}\|_p^r}{d^{r/p}} = \infty.$$

Proof. We shall apply Proposition 1 and Lemma 1 to $Y = |X|^p$, $Y_j = |X_j|^p$, $j \geq 1$, and $\varphi(u) = |u|^{r/p}$.

Proof of (*i*)

For the first part of (*i*), notice that with $s = p/r > 1$

$$\mathbb{E}\left|\varphi\left(\frac{\sum_{j=1}^d |X_j|^p}{d}\right)\right|^s = \frac{\sum_{j=1}^d \mathbb{E}|X_j|^p}{d} = \mathbb{E}|X|^p < \infty.$$

This shows that sufficient **Condition 2** of Lemma 1 holds, which by Proposition 1 gives the result.

For the second part of (*i*) observe that for any $K > 0$

$$\mathbb{E}\left(\frac{\sum_{j=1}^d |X_j|^p}{d}\right)^{r/p} \geq \mathbb{E}\left(\frac{\sum_{j=1}^d |X_j|^p \mathbf{1}\{|X_j| \leq K\}}{d}\right)^{r/p}.$$

Observing that the right-hand side of the inequality converges to $(\mathbb{E}|X|^p \mathbf{1}\{|X| \leq K\})^{r/p}$ as $d \to \infty$, we get for any $K > 0$

$$\liminf_{d \to \infty} \mathbb{E}\left(\frac{\sum_{j=1}^d |X_j|^p}{d}\right)^{r/p} \geq \mathbb{E}(|X|^p \mathbf{1}\{|X| \leq K\})^{r/p}.$$

Since K can be chosen arbitrarily large and we assume that $\mathbb{E}|X|^p = \infty$, we see that the conclusion holds.

Proof of (ii)

For the first part of (ii), note that in this case $r/p \geq 1$, so φ is convex. Moreover, note that

$$\mathbb{E}\left|\varphi\left(\frac{\sum_{j=1}^d |X_j|^p}{d}\right)\right| = \mathbb{E}\left(\frac{\sum_{j=1}^d |X_j|^p}{d}\right)^{r/p}$$
$$\leq d^{-1}\mathbb{E}|X|^r$$
$$\text{(by Jensen's inequality)}$$
$$< \infty.$$

Thus sufficient **Condition 1** of Lemma 1 holds, which by Proposition 1 leads to the result.

For the second part of (ii), observe that if $\mathbb{E}|X|^r = \infty$, then, for all $d \geq 1$,

$$\mathbb{E}\left(\frac{\sum_{j=1}^d |X_j|^p}{d}\right)^{r/p} \geq d^{-r/p}\mathbb{E}|X|^r = \infty.$$

$\square$

Applying Corollary 1 with $p > 0$ and $r = 2$ yields the following important result:

Proposition 2. *Fix $p > 0$ and assume that $0 < \mathbb{E}|X|^m < \infty$ for $m = \max(2, p)$. Then, as $d \to \infty$,*

$$\frac{\mathbb{E}\|\mathbf{X}\|_p}{d^{1/p}} \to \mu_p^{1/p}$$

and

$$\frac{\mathbb{E}\|\mathbf{X}\|_p^2}{d^{2/p}} \to \mu_p^{2/p},$$

which implies

$$\frac{\sqrt{\mathrm{Var}\|\mathbf{X}\|_p}}{\mathbb{E}\|\mathbf{X}\|_p} \to 0, \text{ as } d \to \infty.$$

This result, when correctly stated, corresponds to Theorem 5 of François et al. (2007). It expresses the fact that the *relative standard deviation* converges towards zero when the dimension grows. It is known in the computational learning literature as the *p*-norm concentration in high-dimensional spaces. It is noteworthy that, by Chebyshev's inequality, for all $\varepsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{\|\mathbf{X}\|_p}{\mathbb{E}\|\mathbf{X}\|_p}-1\right|\geq\varepsilon\right\}=\mathbb{P}\left\{\left|\|\mathbf{X}\|_p-\mathbb{E}\|\mathbf{X}\|_p\right|\geq\varepsilon\mathbb{E}\|\mathbf{X}\|_p\right\}$$

$$\leq\frac{\mathrm{Var}\|\mathbf{X}\|_p}{\varepsilon^2\mathbb{E}^2\|\mathbf{X}\|_p}\to0,\text{ as }d\to\infty. \quad (3)$$

That is, $\|\mathbf{X}\|_p/\mathbb{E}\|\mathbf{X}\|_p\overset{\mathbb{P}}{\to}1$ or, in other words, the sequence $\{\|\mathbf{X}\|_p\}_{d\geq1}$ is relatively stable (Boucheron et al., 2013). This property guarantees that the random fluctuations of $\|\mathbf{X}\|_p$ around its expectation are of negligible size when compared to the expectation, and therefore most information about the size of $\|\mathbf{X}\|_p$ is given by $\mathbb{E}\|\mathbf{X}\|_p$ as d becomes large.

2.2 Rates of convergence

The asymptotic concentration statement of Corollary 1 can be made more precise by means of rates of convergence, at the price of stronger moment assumptions. To reach this objective, we first need a general result to control the behavior of a function of an i.i.d. empirical mean around its true value. Thus, assume that $\{Y_j\}_{j\geq1}$ are i.i.d. Y with mean μ and variance σ^2. As before, we define

$$\overline{Y}_d=d^{-1}\sum_{j=1}^d Y_j.$$

Let φ be a real-valued function with derivatives φ' and φ''. Khan (2004) provides sufficient conditions for

$$\mathbb{E}\varphi(\overline{Y}_d)=\varphi(\mu)+\frac{\varphi''(\mu)\sigma^2}{2d}+\mathrm{o}(d^{-2})$$

to hold. The following lemma, whose assumptions are less restrictive, can be used in place of Khan's result (2004). For the sake of clarity, its proof is postponed to Section 4.

Lemma 2. *Let $\{Y_j\}_{j\geq1}$ be a sequence of i.i.d. Y random variables with mean μ and variance σ^2, and φ be a real-valued function with continuous derivatives φ' and φ'' in a neighborhood of μ. Assume that for some $r>1$,*

$$\mathbb{E}|Y|^{r+1}<\infty \quad (4)$$

and, with $1/s=1-1/r$,

$$\limsup_{d\to\infty}\mathbb{E}\left|\varphi(\overline{Y}_d)\right|^s<\infty. \quad (5)$$

Then, as $d\to\infty$,

$$\mathbb{E}\varphi(\overline{Y}_d)=\varphi(\mu)+\frac{\varphi''(\mu)\sigma^2}{2d}+\mathrm{o}(d^{-1}).$$

The consequences of Lemma 2 in terms of p-norm concentration are summarized in the following proposition:

Proposition 3. *Fix $p > 0$ and assume that $0 < \mathbb{E}|X|^m < \infty$ for $m = \max(4, 3p)$. Then, as $d \to \infty$,*

$$\mathbb{E}\|\mathbf{X}\|_p = d^{1/p}\mu_p^{1/p} + \mathrm{O}(d^{1/p-1})$$

and

$$\mathrm{Var}\|\mathbf{X}\|_p = \frac{\mu_p^{2/p-2}\sigma_p^2}{d^{1-2/p}p^2} + \mathrm{o}(d^{-1+2/p}),$$

which implies

$$\frac{\sqrt{d\,\mathrm{Var}\|\mathbf{X}\|_p}}{\mathbb{E}\|\mathbf{X}\|_p} \to \frac{\sigma_p}{p\mu_p}, \ as\ d \to \infty.$$

Proposition 3 shows that for a fixed large d, the *relative standard deviation* evolves with p as the ratio $\sigma_p/(p\mu_p)$. For instance, when the distribution of X is uniform,

$$\mu_p = \frac{1}{p+1} \quad \text{and} \quad \sigma_p = \frac{p}{p+1}\sqrt{\frac{1}{2p+1}}.$$

In that case, we conclude that

$$\frac{\sqrt{d\,\mathrm{Var}\|\mathbf{X}\|_p}}{\mathbb{E}\|\mathbf{X}\|_p} \to \sqrt{\frac{1}{2p+1}}.$$

Thus, in the uniform setting, the limiting *relative standard deviation* is a strictly decreasing function of p. This observation is often interpreted by saying that p-norms are more concentrated for larger values of p. There are however distributions for which this is not the case. A counterexample is given by a balanced mixture of two standard Gaussian random variables with mean 1 and -1, respectively (see François et al., 2007, page 881). In that case, it can be seen that the asymptotic *relative standard deviation* with $p \le 1$ is smaller than for values of $p \in [8, 30]$, making fractional norms more concentrated.

Proof (Proposition 3). Fix $p > 0$ and introduce the functions on $\mathbb{R}$

$$\varphi_1(u) = |u|^{1/p} \quad \text{and} \quad \varphi_2(u) = |u|^{2/p}.$$

Assume that $\mathbb{E}|X|^{\max(4,p)} < \infty$. Applying Corollary 1 we get that, as $d \to \infty$,

$$\mathbb{E}\left(\frac{\sum_{j=1}^d |X_j|^p}{d}\right)^{2/p} \to \mu_p^{2/p}$$

and

$$\mathbb{E}\left(\frac{\sum_{j=1}^d |X_j|^p}{d}\right)^{4/p} \to \mu_p^{4/p}.$$

This says that with $r = 2$ and $s = 2$, for $i = 1, 2$,

$$\limsup_{d \to \infty} \mathbb{E}\left| \varphi_i \left(\frac{\sum_{j=1}^d |X_j|^p}{d} \right) \right|^s < \infty.$$

Now, let $Y = |X|^p$. If we also assume that $\mathbb{E}|Y|^{r+1} = \mathbb{E}|Y|^3 = \mathbb{E}|X|^{3p} < \infty$, we get by applying Lemma 2 to φ_1 and φ_2 that for $i = 1, 2$

$$\mathbb{E}\varphi_i(\overline{Y}_d) = \varphi_i(\mu_p) + \frac{\varphi_i''(\mu_p)\sigma_p^2}{2d} + \mathrm{o}(d^{-1}).$$

Thus, whenever $\mathbb{E}|X|^m < \infty$, where $m = \max(4, 3p)$,

$$\mathbb{E}|\overline{Y}_d|^{1/p} = \mu_p^{1/p} + \frac{1}{p}\left(\frac{1-p}{p} \right) \frac{\mu_p^{1/p-2}\sigma_p^2}{2d} + \mathrm{o}(d^{-1})$$

and

$$\mathbb{E}|\overline{Y}_d|^{2/p} = \mu_p^{2/p} + \frac{1}{p}\left(\frac{2-p}{p} \right) \frac{\mu_p^{2/p-2}\sigma_p^2}{d} + \mathrm{o}\left(d^{-1}\right).$$

Therefore, we see that

$$\mathrm{Var}|\overline{Y}_d|^{1/p} = \mathbb{E}|\overline{Y}_d|^{2/p} - \mathbb{E}^2|\overline{Y}_d|^{1/p}$$
$$= \frac{\mu_p^{2/p-2}\sigma_p^2}{dp^2} + \mathrm{o}\left(d^{-1}\right).$$

The identity $\overline{Y}_d = d^{-1}\sum_{j=1}^d |X_j|^p$ yields the desired results. $\quad\square$

We conclude the section with a corollary, which specifies inequality (3).

Corollary 2. *Fix $p > 0$.*

(i) If $0 < \mathbb{E}|X|^m < \infty$ for $m = \max(4, 3p)$, then, for all $\varepsilon > 0$,

$$\mathbb{P}\left\{ \left| \frac{\|\mathbf{X}\|_p}{\mathbb{E}\|\mathbf{X}\|_p} - 1 \right| \geq \varepsilon \right\} \leq \frac{\sigma_p^2}{\varepsilon^2 dp^2 \mu_p^2} + \mathrm{o}(d^{-1}).$$

(ii) If for some positive constant C, $0 < |X| \leq C$ almost surely, then, for $p \geq 1$ and all $\varepsilon > 0$,

$$\mathbb{P}\left\{ \left| \frac{\|\mathbf{X}\|_p}{\mathbb{E}\|\mathbf{X}\|_p} - 1 \right| \geq \varepsilon \right\} \leq 2\exp\left(-\frac{\varepsilon^2 d^{2/p-1}\mu_p^{2/p}}{2C^2} + \mathrm{o}(d^{2/p-1}) \right).$$

Proof. Statement (i) is an immediate consequence of Proposition 3 and Chebyshev's inequality. Now, assume that $p \geq 1$, and let $A = [-C, C]$. For $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, let $g : A^d \to \mathbb{R}$ be defined by

$$g(\mathbf{x}) = \|\mathbf{x}\|_p = \left(\sum_{j=1}^{d} |x_j|^p \right)^{1/p}.$$

Clearly, for each $1 \leq j \leq d$,

$$\sup_{\substack{(x_1,\ldots,x_d) \in A^d \\ x_j' \in A}} \left| g(x_1,\ldots,x_d) - g(x_1,\ldots,x_{j-1},x_j',x_{j+1},\ldots,x_d) \right|$$

$$= \sup_{\mathbf{x} \in A^d, x_j' \in A} \left| \|\mathbf{x}\|_p - \|\mathbf{x}'\|_p \right|,$$

where $\mathbf{x}'$ is identical to $\mathbf{x}$, except on the j-th coordinate where it takes the value x_j'. It follows, by Minkowski inequality (which is valid here since $p \geq 1$), that

$$\sup_{\substack{(x_1,\ldots,x_d) \in A^d \\ x_j' \in A}} \left| g(x_1,\ldots,x_d) - g(x_1,\ldots,x_{j-1},x_j',x_{j+1},\ldots,x_d) \right|$$

$$\leq \sup_{\substack{\mathbf{x} \in A^d \\ x_j' \in A}} \|\mathbf{x} - \mathbf{x}'\|_p$$

$$= \sup_{(x,x') \in A^2} |x - x'| \leq 2C.$$

Consequently, using the bounded difference inequality (McDiarmid, 1989), we obtain

$$\mathbb{P}\left\{ \left| \frac{\|\mathbf{X}\|_p}{\mathbb{E}\|\mathbf{X}\|_p} - 1 \right| \geq \varepsilon \right\} = \mathbb{P}\left\{ \left| \|\mathbf{X}\|_p - \mathbb{E}\|\mathbf{X}\|_p \right| \geq \varepsilon \mathbb{E}\|\mathbf{X}\|_p \right\}$$

$$\leq 2 \exp\left(-\frac{2(\varepsilon \mathbb{E}\|\mathbf{X}\|_p)^2}{4dC^2} \right)$$

$$= 2 \exp\left(-\frac{\varepsilon^2 d^{2/p-1} \mu_p^{2/p}}{2C^2} + \mathrm{o}(d^{2/p-1}) \right),$$

where, in the last inequality, we used Proposition 3. This concludes the proof. $\square$

3 Minima and maxima

Another important question arising in high-dimensional nearest neighbor search analysis concerns the relative asymptotic behavior of the minimum and maximum p-distances to the origin within a random sample. To be precise, let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be an i.i.d. $\mathbf{X}$ sample, where $\mathbf{X} = (X_1,\ldots,X_d)$ is as usual a $\mathbb{R}^d$-valued random vector with i.i.d. X components. We will be primarily interested in this section in the asymptotic properties of the difference (the *contrast*) $\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_p - \min_{1 \leq i \leq d} \|\mathbf{X}_i\|_p$.

In other words, given a data set and a fixed query point located—without loss of generality—at the origin, we seek to analyze how much the distances to the farthest and nearest neighbors differ.

Assume, to start with, that n is fixed and only d is allowed to grow. Then an immediate application of the law of large numbers shows that, whenever $\mu_p = \mathbb{E}|X|^p < \infty$, almost surely as $d \to \infty$,

$$d^{-1/p} \left(\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_p - \min_{1 \leq i \leq n} \|\mathbf{X}_i\|_p \right) \xrightarrow{\mathbb{P}} 0.$$

Moreover, if $0 < \mu_p < \infty$, then

$$\frac{\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_p}{\min_{1 \leq i \leq n} \|\mathbf{X}_i\|_p} \xrightarrow{\mathbb{P}} 1.$$

The above ratio is sometimes called the *relative contrast* in the computational learning literature. Thus, as d becomes large, all observations seem to be distributed at approximately the same p-distance from the query point. The concept of nearest neighbor (measured by p-norms) in high dimension is therefore less clear than in small dimension, with resulting computational difficulties and algorithmic inefficiencies.

These consistency results can be specified by means of asymptotic distributions. Recall that if $Z_1, \ldots, Z_n$ are i.i.d standard normal random variables, the sample range is defined to be

$$M_n = \max_{1 \leq i \leq n} Z_i - \min_{1 \leq i \leq n} Z_i.$$

The asymptotic distribution of M_n is well known (see, e.g., David, 1981). Namely, for any x one has

$$\lim_{n \to \infty} \mathbb{P} \left\{ \sqrt{2 \log n} \left(M_n - 2\sqrt{2 \log n} + \frac{\log \log n + \log 4\pi}{2\sqrt{2 \log n}} \right) \leq x \right\}$$
$$= \int_{-\infty}^{\infty} \exp\left(-t - e^{-t} - e^{-(x-t)} \right) dt.$$

For future reference, we shall sketch the proof of this fact here. It is well known that with

$$a_n = \sqrt{2 \log n} \quad \text{and} \quad b_n = \sqrt{2 \log n} - \frac{1}{2} \frac{(\log \log n + \log 4\pi)}{\sqrt{2 \log n}} \tag{6}$$

we have

$$\left(a_n (\max_{1 \leq i \leq n} Z_i - b_n), a_n (\min_{1 \leq i \leq n} Z_i + b_n) \right) \to (E, -E'), \tag{7}$$

where E and E' are independent, $E = E'$ and $\mathbb{P}\{E \leq x\} = \exp(-\exp(-x))$, $-\infty < x < \infty$. (The asymptotic independence of the maximum and minimum part can be inferred from Theorem 4.2.8 of Reiss, 1989, and the asymptotic distribution part from Example 2 on page 71 of Resnick, 1987.) From (7) we get

$$a_n(\max_{1 \le i \le n} Z_i - \min_{1 \le i \le n} Z_i) - 2a_n b_n \overset{\mathscr{D}}{\to} E + E'.$$

Clearly,

$$\mathbb{P}\{E + E' \le x\} = \int_{-\infty}^{\infty} \exp\left(-e^{-(x-t)}\right) \exp(-e^{-t}) e^{-t} dt$$

$$= \int_{-\infty}^{\infty} \exp\left(-t - e^{-t} - e^{-(x-t)}\right) dt.$$

Our first result treats the case when n is fixed and $d \to \infty$.

Proposition 4. *Fix* $p > 0$ *and assume that* $0 < \mathbb{E}|X|^p < \infty$. *Then, for fixed* n, *as* $d \to \infty$,

$$d^{1/2 - 1/p}\left(\max_{1 \le i \le n} \|\mathbf{X}_i\|_p - \min_{1 \le i \le n} \|\mathbf{X}_i\|_p\right) \overset{\mathscr{D}}{\to} \frac{\sigma_p \mu_p^{1/p-1}}{p} M_n.$$

To our knowledge, this is the first statement of this type in the analysis of high-dimensional nearest neighbor problems. In fact, most of the existing results merely bound the asymptotic expectation of the (normalized) difference and ratio between the max and the min, but with bounds which are unfortunately not of the same order in n as soon as $n \ge 3$ (see, e.g., Theorem 3 in Hinneburg et al., 2000).

One of the consequences of Proposition 4 is that, for fixed n, the difference between the farthest and nearest neighbors does not necessarily go to zero in probability as d tends to infinity. Indeed, we see that the size of

$$\max_{1 \le i \le n} \|\mathbf{X}_i\|_p - \min_{1 \le i \le n} \|\mathbf{X}_i\|_p$$

grows as $d^{1/p - 1/2}$. For example, this difference increases with dimensionality as $\sqrt{d}$ for the L^1 (Manhattan) metric and remains stable in distribution for the L^2 (Euclidean) metric. It tends to infinity in probability for $p < 2$ and to zero for $p > 2$. This observation is in line with the conclusions of Hinneburg et al. (2000), who argue that nearest neighbor search in a high-dimensional space tends to be meaningless for norms with larger exponents, since the maximum observed distance tends towards the minimum one. It should be noted, however, that the variance of the limiting distribution depends on the value of p.

Remark 1. Let $Z_1, \ldots, Z_n$ be i.i.d standard normal random variables, and let

$$R_n = \frac{\max_{1 \le i \le n} Z_i}{\min_{1 \le i \le n} Z_i}.$$

Assuming $\mu_p > 0$, one can prove, using the same technique, that

$$\frac{\max_{1 \le i \le n} \|\mathbf{X}_i\|_p - d^{1/p} \mu_p}{\min_{1 \le i \le n} \|\mathbf{X}_i\|_p - d^{1/p} \mu_p} \overset{\mathscr{D}}{\to} R_n.$$

Proof (Proposition 4). Denote by $\mathbf{Z}_n$ a centered Gaussian random vector in $\mathbb{R}^n$, with identity covariance matrix. By the central limit theorem, as $d \to \infty$,

$$\sqrt{d}\left[\left(\frac{\|\mathbf{X}_1\|_p^p}{d},\ldots,\frac{\|\mathbf{X}_n\|_p^p}{d}\right) - (\mu_p,\ldots,\mu_p)\right] \xrightarrow{\mathscr{D}} \sigma_p \mathbf{Z}_n.$$

Applying the delta method with the mapping $f(x_1,\ldots,x_n) = (x_1^{1/p},\ldots,x_n^{1/p})$ (which is differentiable at $(\mu_p,\ldots,\mu_p)$ since $\mu_p > 0$), we obtain

$$\sqrt{d}\left[\left(\frac{\|\mathbf{X}_1\|_p}{d^{1/p}},\ldots,\frac{\|\mathbf{X}_n\|_p}{d^{1/p}}\right) - (\mu_p^{1/p},\ldots,\mu_p^{1/p})\right] \xrightarrow{\mathscr{D}} \frac{\sigma_p \mu_p^{1/p-1}}{p}\mathbf{Z}_n.$$

Thus, by continuity of the maximum and minimum functions,

$$d^{1/2-1/p}\left(\max_{1\leq i\leq n}\|\mathbf{X}_i\|_p - \min_{1\leq i\leq n}\|\mathbf{X}_i\|_p\right) \xrightarrow{\mathscr{D}} \frac{\sigma_p \mu_p^{1/p-1}}{p}M_n.$$

$\square$

In the previous analysis, n (the sample size) was fixed whereas d (the dimension) was allowed to grow to infinity. A natural question that arises concerns the impact of letting n be a function of d such that n tends to infinity as $d \to \infty$ (Mallows, 1972). Proposition 5 below offers a first answer.

Proposition 5. *Fix $p \geq 1$, and assume that $0 < \mathbb{E}|X|^{3p} < \infty$ and $\sigma_p > 0$. For any sequence of positive integers $\{n(d)\}_{d\geq 1}$ converging to infinity and satisfying*

$$n(d) = \mathrm{o}\left(\frac{d^{1/5}}{\log^{6/5} d}\right), \text{ as } d \to \infty, \tag{8}$$

we have

$$\frac{p a_{n(d)} d^{1/2-1/p}}{\mu_p^{1/p-1}\sigma_p}\left(\max_{1\leq i\leq n(d)}\|\mathbf{X}_i\|_p - \min_{1\leq i\leq n(d)}\|\mathbf{X}_i\|_p\right) - 2a_{n(d)}b_{n(d)} \xrightarrow{\mathscr{D}} E + E',$$

where a_n and b_n are as in (6), and E and E' are as in (7).

Proof. In the following, we let $\delta(d) = 1/\log d$. For future use note that

$$\delta^2(d)\log n(d) \to 0 \quad \text{and} \quad \frac{n^5(d)}{d\delta^6(d)} \to 0, \text{ as } d \to \infty. \tag{9}$$

In the proof we shall often suppress the dependence of n and δ on d. For $1 \leq i \leq n$, we set

$$\mathbf{X}_i = (X_{1,i},\ldots,X_{d,i}) \quad \text{and} \quad \|\mathbf{X}_i\|_p^p = \sum_{j=1}^{d}|X_{j,i}|^p.$$

We see that for $n \geq 1$,

$$
\left(\frac{\|\mathbf{X}_1\|_p^p - d\mu_p}{\sqrt{d}\sigma_p}, \ldots, \frac{\|\mathbf{X}_n\|_p^p - d\mu_p}{\sqrt{d}\sigma_p} \right)
$$

$$
= \left(\frac{\sum_{j=1}^d |X_{j,1}|^p - d\mu_p}{\sqrt{d}\sigma_p}, \ldots, \frac{\sum_{j=1}^d |X_{j,n}|^p - d\mu_p}{\sqrt{d}\sigma_p} \right)
$$

$$
:= (Y_1, \ldots, Y_n) = \mathbf{Y}_n \in \mathbb{R}^n.
$$

As above, let $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$ be a centered Gaussian random vector in $\mathbb{R}^n$, with identity covariance matrix. Write, for $1 \leq j \leq d$,

$$
\boldsymbol{\xi}_j = \left(\frac{|X_{j,1}|^p - \mu_p}{\sqrt{d}\sigma_p}, \ldots, \frac{|X_{j,n}|^p - \mu_p}{\sqrt{d}\sigma_p} \right)
$$

and note that $\sum_{j=1}^d \boldsymbol{\xi}_j = \mathbf{Y}_n$. Set $\beta = \sum_{j=1}^d \|\boldsymbol{\xi}_j\|_2^3$. Then, by Jensen's inequality,

$$
\mathbb{E}\|\boldsymbol{\xi}_j\|_2^3 = \mathbb{E}\left(\frac{\sum_{i=1}^n \left(|X_{j,i}|^p - \mu_p\right)^2}{d\sigma_p^2} \right)^{3/2} \leq \left(\frac{n}{d\sigma_p^2} \right)^{3/2} \mathbb{E}\left| |X|^p - \mu_p \right|^3.
$$

This gives that for any $\delta > 0$, possibly depending upon n,

$$
B := \beta n \delta^{-3} \leq \frac{n^{5/2}}{\sqrt{d}\sigma_p^3} \mathbb{E}\left| |X|^p - \mu_p \right|^3 \delta^{-3}.
$$

Applying a result of Yurinskiĭ (1977) as formulated in Section 4 of Chapter 10 of Pollard (2001) we get, on a suitable probability space depending on $\delta > 0$, there exist random vectors $\mathbf{Y}_n$ and $\mathbf{Z}_n$ satisfying

$$
\mathbb{P}\left\{ \|\mathbf{Y}_n - \mathbf{Z}_n\|_2 > 3\delta \right\} \leq CB\left(1 + \frac{|\log(B)|}{n} \right), \tag{10}
$$

where C is a universal constant. Using the fact that

$$
\left| \max_{1 \leq i \leq n} x_i - \max_{1 \leq i \leq n} y_i \right| \leq \sqrt{\sum_{i=1}^n (x_i - y_i)^2},
$$

we get, for all $\varepsilon > 0$,

$$
\mathbb{P}\left\{ a_n \left| \max_{1 \leq i \leq n} Y_i - \max_{1 \leq i \leq n} Z_i \right| > \varepsilon \right\} \leq \mathbb{P}\left\{ \sqrt{2\log n}\, \|\mathbf{Y}_n - \mathbf{Z}_n\|_2 > \varepsilon \right\}.
$$

Thus, for all d large enough,

$$\mathbb{P}\left\{a_n|\max_{1\le i\le n} Y_i - \max_{1\le i\le n} Z_i| > \varepsilon\right\} \le \mathbb{P}\left\{\sqrt{2\log n}\,\|\mathbf{Y}_n - \mathbf{Z}_n\|_2 > 3\delta\sqrt{2\log n}\right\}$$

$$\text{(since } \delta\sqrt{\log n} \to 0 \text{ as } d \to \infty)$$

$$= \mathbb{P}\left\{\|\mathbf{Y}_n - \mathbf{Z}_n\|_2 > 3\delta\right\}.$$

From (10), we deduce that for all $\varepsilon > 0$ and all d large enough,

$$\mathbb{P}\left\{a_n|\max_{1\le i\le n} Y_i - \max_{1\le i\le n} Z_i| > \varepsilon\right\} \le CB\left(1 + \frac{|\log(B)|}{n}\right).$$

But, by our choice of $\delta(d)$ and (9),

$$B\left(1 + \frac{|\log(B)|}{n}\right) \to 0,$$

so that

$$a_n|\max_{1\le i\le n} Y_i - \max_{1\le i\le n} Z_i| = o_{\mathbb{P}}(1).$$

Similarly, one proves that

$$a_n|\min_{1\le i\le n} Y_i - \min_{1\le i\le n} Z_i| = o_{\mathbb{P}}(1).$$

Thus, by (7), we conclude that

$$\left(a_n(\max_{1\le i\le n} Y_i - b_n), a_n(\min_{1\le i\le n} Y_i + b_n)\right) \xrightarrow{\mathscr{D}} (E, -E'). \tag{11}$$

Next, we have

$$\left(a_n(\max_{1\le i\le n} Y_i - b_n), a_n(\min_{1\le i\le n} Y_i + b_n)\right)$$

$$= \left(a_n\left(\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p} - \frac{\sqrt{d}\mu_p}{\sigma_p} - b_n\right), a_n\left(\frac{\min_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p} - \frac{\sqrt{d}\mu_p}{\sigma_p} + b_n\right)\right)$$

$$= \left(a_n\left(\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p} - \beta_n\right), a_n\left(\frac{\min_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p} - \beta_n'\right)\right),$$

where $\beta_n = \frac{\sqrt{d}\mu_p}{\sigma_p} + b_n$ and $\beta_n' = \frac{\sqrt{d}\mu_p}{\sigma_p} - b_n$. Note that $a_n \to \infty$ and (11) imply that both

$$\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p} - \beta_n \xrightarrow{\mathbb{P}} 0 \quad \text{and} \quad \frac{\min_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p} - \beta_n' \xrightarrow{\mathbb{P}} 0. \tag{12}$$

Observe also that by a two term Taylor expansion, for a suitable $\widetilde{\beta}_n$ between β_n and $(\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p)/(\sqrt{d}\sigma_p)$,

$$\frac{pa_n}{\beta_n^{1/p-1}}\left(\left(\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p}\right)^{1/p}-\beta_n^{1/p}\right)$$

$$=a_n\left(\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p}-\beta_n\right)$$

$$+\frac{a_n}{\beta_n^{1/p-1}}\frac{1-p}{2p}\widetilde{\beta}_n^{1/p-2}\left(\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p}-\beta_n\right)^2.$$

We obtain by (11) and (12) that

$$a_n^2\left(\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p}-\beta_n\right)^2\frac{\widetilde{\beta}_n^{1/p-2}}{a_n\beta_n^{1/p-1}}=\mathrm{O}_{\mathbb{P}}\left(\frac{1}{a_n\beta_n}\right)=\mathrm{o}_{\mathbb{P}}(1).$$

Similarly,

$$\frac{pa_n}{(\beta_n')^{1/p-1}}\left(\left(\frac{\min_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p}\right)^{1/p}-(\beta_n')^{1/p}\right)$$

$$=a_n\left(\frac{\min_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p}-\beta_n'\right)+\mathrm{o}_{\mathbb{P}}(1).$$

Keeping in mind that $\beta_n/\beta_n'\to 1$, we get

$$\frac{pa_n}{\beta_n^{1/p-1}}\left(\left(\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p}\right)^{1/p}-\beta_n^{1/p},\left(\frac{\min_{1\le i\le n}\|\mathbf{X}_i\|_p^p}{\sqrt{d}\sigma_p}\right)^{1/p}-(\beta_n')^{1/p}\right)$$

$$\xrightarrow{\mathscr{D}}(E,-E')$$

and hence

$$\frac{pa_n}{\beta_n^{1/p-1}}\left(\frac{\max_{1\le i\le n}\|\mathbf{X}_i\|_p}{(\sqrt{d}\sigma_p)^{1/p}}-\frac{\min_{1\le i\le n}\|\mathbf{X}_i\|_p}{(\sqrt{d}\sigma_p)^{1/p}}-\beta_n^{1/p}+(\beta_n')^{1/p}\right)\xrightarrow{\mathscr{D}}E+E'.$$

Next notice that (8) implies that $b_n/\sqrt{d}\to 0$, as $d\to\infty$. Thus, recalling

$$\frac{\beta_n}{\sqrt{d}u_p/\sigma_p}=1+\frac{b_n}{\sqrt{d}\mu_p/\sigma_p}\quad\text{and}\quad\frac{\beta_n'}{\sqrt{d}u_p/\sigma_p}=1-\frac{b_n}{\sqrt{d}\mu_p/\sigma_p},$$

we are led to

$$\frac{pa_n}{\beta_n^{1/p-1}}\left(\beta_n^{1/p}-(\beta_n')^{1/p}\right)=2a_nb_n+\mathrm{O}(a_nb_n^2\beta_n^{-1})=2a_nb_n+\mathrm{o}(1).$$

Therefore we get

$$\frac{pa_{n(d)}d^{1/2-1/p}}{\mu_p^{1/p-1}\sigma_p}\left(\max_{1\le i\le n(d)}\|\mathbf{X}_i\|_p - \min_{1\le i\le n(d)}\|\mathbf{X}_i\|_p\right) - 2a_{n(d)}b_{n(d)} \xrightarrow{\mathscr{D}} E + E'.$$

□

4 Proof of Lemma 2

In the sequel, to lighten notation a bit, we set $\overline{Y} = \overline{Y}_d$. Choose any $\varepsilon > 0$ and $\delta > 0$ such that φ has continuous derivatives φ' and φ'' on $I_\delta = [\mu - \delta, \mu + \delta]$ and $|\varphi''(\mu) - \varphi''(x)| \le \varepsilon$ for all $x \in I_\delta$. We see that by Taylor's theorem that for $\overline{Y} \in I_\delta$

$$\varphi(\overline{Y}) = \varphi(\mu) + \varphi'(\mu)(\overline{Y} - \mu) + 2^{-1}\varphi''(\widetilde{\mu})(\overline{Y} - \mu)^2, \qquad (13)$$

where $\widetilde{\mu}$ lies between $\overline{Y}$ and μ. Clearly,

$$\left|\mathbb{E}\varphi(\overline{Y}) - \varphi(\mu) - \frac{\sigma^2\varphi''(\mu)}{2d}\right|$$
$$= \left|\mathbb{E}\left(\varphi(\overline{Y}) - \left(\varphi(\mu) + \varphi'(\mu)(\overline{Y} - \mu) + 2^{-1}\varphi''(\mu)(\overline{Y} - \mu)^2\right)\right)\right|$$
$$\le \left|\mathbb{E}\left(\left\{\varphi(\overline{Y}) - \left(\varphi(\mu) + \varphi'(\mu)(\overline{Y} - \mu) + 2^{-1}\varphi''(\mu)(\overline{Y} - \mu)^2\right)\right\}\mathbf{1}\{\overline{Y} \in I_\delta\}\right)\right|$$
$$+ \mathbb{E}\left(\left|\varphi(\overline{Y})\right|\mathbf{1}\{\overline{Y} \notin I_\delta\}\right) + \mathbb{E}\left(\left|P(\overline{Y})\right|\mathbf{1}\{\overline{Y} \notin I_\delta\}\right),$$

where

$$P(y) = \varphi(\mu) + \varphi'(\mu)(y - \mu) + 2^{-1}\varphi''(\mu)(y - \mu)^2.$$

Now using (13) and $|\varphi''(\mu) - \varphi''(x)| \le \varepsilon$ for all $x \in I_\delta$, we may write

$$\left|\mathbb{E}\left(\left\{\varphi(\overline{Y}) - \left(\varphi(\mu) + \varphi'(\mu)(\overline{Y} - \mu) + 2^{-1}\varphi''(\mu)(\overline{Y} - \mu)^2\right)\right\}\mathbf{1}\{\overline{Y} \in I_\delta\}\right)\right|$$
$$\le \frac{\varepsilon}{2}\mathbb{E}(\overline{Y} - \mu)^2 = \frac{\varepsilon\sigma^2}{2d}.$$

Next, we shall bound

$$\mathbb{E}\left(\left|\varphi(\overline{Y})\right|\mathbf{1}\{\overline{Y} \notin I_\delta\}\right) + \mathbb{E}\left(\left|P(\overline{Y})\right|\mathbf{1}\{\overline{Y} \notin I_\delta\}\right) := \Delta_d^{(1)} + \Delta_d^{(2)}.$$

Recall that we assume that for some $r > 1$, condition (4) holds. In this case, by Theorem 28 on page 286 of Petrov (1975) applied with "r" replaced by "$r + 1$", for all $\delta > 0$,

$$\mathbb{P}\{|\overline{Y} - \mu| \ge \delta\} = \mathrm{o}(d^{-r}). \qquad (14)$$

Then, by using Hölder's inequality and (5), we get

$$\Delta_d^{(1)} \le \left(\mathbb{E}\left|\varphi(\overline{Y})\right|^s\right)^{1/s}\left(\mathbb{P}\{\overline{Y} \notin I_\delta\}\right)^{1/r} = \mathrm{o}(d^{-1}).$$

We shall next bound $\Delta_d^{(2)}$. Obviously from (14)

$$|\varphi(\mu)|\mathbb{P}\{\overline{Y} \notin I_\delta\} = \mathrm{o}(d^{-1}).$$

Furthermore, by Cauchy-Schwarz inequality and (14),

$$\mathbb{E}\left|\varphi'(\mu)(\overline{Y} - \mu)\mathbf{1}\{\overline{Y} \notin I_\delta\}\right| \leq \left|\varphi'(\mu)\right|\sigma d^{-1/2}\mathrm{o}(d^{-r/2}) = \mathrm{o}(d^{-1}),$$

and by Hölder's inequality with $p = (r+1)/2$ and

$$q^{-1} = 1 - p^{-1} = 1 - 2/(r+1) = (r-1)/(r+1),$$

we have

$$2^{-1}\left|\varphi''(\mu)\right|\mathbb{E}\left((\overline{Y} - \mu)^2\mathbf{1}\{\overline{Y} \notin I_\delta\}\right)$$
$$\leq 2^{-1}\left|\varphi''(\mu)\right|\left(\mathbb{E}|\overline{Y} - \mu|^{r+1}\right)^{2/(r+1)}\left(\mathbb{P}\{\overline{Y} \notin I_\delta\}\right)^{1/q}.$$

Applying Rosenthal's inequality (see equation (2.3) in Giné et al., 2003) we obtain

$$\mathbb{E}|\overline{Y} - \mu|^{r+1} = \mathbb{E}\left|d^{-1}\sum_{i=1}^{d}(Y_i - \mu)\right|^{r+1}$$
$$\leq \left(\frac{15(r+1)}{\log(r+1)}\right)^{r+1}\max\left(d^{-(r+1)/2}\left(\mathbb{E}Y^2\right)^{(r+1)/2}, d^{-r}\mathbb{E}|Y|^{r+1}\right).$$

Thus

$$\left(\mathbb{E}|\overline{Y} - \mu|^{r+1}\right)^{2/(r+1)} = \mathrm{O}(d^{-1}),$$

which when combined with (14) gives

$$2^{-1}\left|\varphi''(\mu)\right|\left(\mathbb{E}|\overline{Y} - \mu|^{r+1}\right)^{2/(r+1)}\left(\mathbb{P}\{\overline{Y} \notin I_\delta\}\right)^{(r-1)/(r+1)} = \mathrm{o}(d^{-1}).$$

Thus

$$\Delta_d^{(2)} = \mathrm{o}(d^{-1}).$$

Putting everything together, we conclude that for any $\varepsilon > 0$

$$\limsup_{d\to\infty} d\left|\mathbb{E}\varphi(\overline{Y}_d) - \varphi(\mu) - \frac{\sigma^2\varphi''(\mu)}{2d}\right| \leq \frac{\varepsilon\sigma^2}{2}.$$

Since $\varepsilon > 0$ can be chosen arbitrarily small, this completes the proof.

References

C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Proceedings of the 8th International Conference on Database Theory*, pages 420–434, Berlin, 2001. Springer.

K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Proceedings of the 7th International Conference on Database Theory*, pages 217–235, Berlin, 1999. Springer.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

H. David. *Order Statistics. 2nd Edition*. Wiley, New York, 1981.

D. François, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19:873–886, 2007.

E. Giné, D.M. Mason, and A.Yu. Zaitsev. The L_1-norm density estimator process. *The Annals of Probability*, 31:719–768, 2003.

A. Hinneburg, C.C. Aggarwal, and D.A. Keim. What is the nearest neighbor in high dimensional spaces. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 506–515, San Francisco, 2000. Morgan Kaufmann.

A. Kabán. Non-parametric detection of meaningless distances in high dimensional data. *Statistics and Computing*, 22:375–385, 2012.

R.A. Khan. Approximation for the expectation of a function of the sample mean. *Statistics*, 38:117–122, 2004.

C.L. Mallows. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 43:508–515, 1972.

C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, 1989*, London Mathematical Society Lecture Note Series 141, pages 148–188. Cambridge University Press, 1989.

V.V. Petrov. *Sums of Independent Random Variables*, volume 82 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, New York, 1975.

D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, Cambridge, 2001.

R.-D. Reiss. *Approximate Distributions of Order Statistics. With Applications to Nonparametric Statistics*. Springer, New York, 1989.

S.I. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer, New York, 1987.

V.V. Yurinskiĭ. On the error of the Gaussian approximation for convolutions. *Teoriya Veroyatnostei i ee Primeneniya*, 22:242–253, 1977.