Probability Inequalities for Kernel Embeddings in Sampling without Replacement

Markus Schneider
University of Ulm, Institute of Neural Information Processing
89069 Ulm, Germany
Ravensburg-Weingarten University of Applied Sciences
88250 Weingarten, Germany

Abstract

The *kernel embedding of distributions* is a popular machine learning technique to manipulate probability distributions and is an integral part of numerous applications. Its empirical counterpart is an estimate from a finite set of samples from the distribution under consideration. However, for large-scale learning problems the empirical kernel embedding becomes infeasible to compute and approximate, constant time solutions are necessary. One can use a random subset of smaller size as a proxy for the exhaustive set of samples to calculate the empirical kernel embedding which is known as *sampling without replacement*. In this work we generalize the results of Serfling (1974) to quantify the difference between the full empirical kernel embedding and the one estimated from random subsets. Furthermore, we derive probability inequalities for Banach space valued martingales in the setting of sampling without replacement.

1 INTRODUCTION

The *kernel embedding of distributions* or *kernel mean map* (Smola et al., 2007; Sriperumbudur et al., 2008) became a popular technique to handle probability measures. The key idea is to embed a distribution into a reproducing kernel Hilbert space (RKHS) where the distribution is then in a more accessible form and can be manipulated efficiently. Most often we have to estimate the kernel embedding empirically from samples and consequently we are concerned with *concentration inequalities* for such sample estimates which are introduced in the next section.

Kernel embedding has been successfully applied to a wide variety of applications. Gretton et al. (2012a) used the kernel mean map to perform statistical hypotheses testing in high-dimensional spaces. They formulated a two-sample test using the difference between the kernel embeddings of two distributions as a test statistic which is called *maximum mean discrepancy* (MMD) (Gretton et al., 2005, 2012b). Gretton et al. (2005) applied the MMD to the problem of independence testing, with the associated test statistic being the *Hilbert-Schmidt independence criterion* (HSIC). The HSIC can also be applied to feature selection by choosing a subset of features which maximizes the dependence between data and labels (Song et al., 2012) or to derive a non-parametric independence test for random processes (Chwialkowski and Gretton, 2014). A new representation of *Hidden Markov Models* (HMM) for structured and non-Gaussian continuous observation distributions can be defined in a way such that the model updates can be performed entirely in the RKHS in which these distributions are embedded (Song et al., 2010). Representing prior and conditional probabilities through the kernel mean map allows the *kernel Bayes' rule* to perform inference of the posterior distribution in the same RKHS without an explicit parametric from (Fukumizu et al., 2013). Song et al. (2011) derived a kernelized belief propagation algorithm using the kernel embedding of conditional distributions (Song et al., 2009).

In the following, we introduce concepts and notation required for the understanding of reproducing kernel Hilbert spaces and the kernel embedding.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

1.1 The Kernel Mean Embedding

Let X be a random variable taking values in a measurable space $(\mathcal{X}, \mathscr{X})$ with distribution $\mathbb{P}$. A Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ of functions $f\colon \mathcal{X} \to \mathbb{R}$ is said to be a *reproducing kernel Hilbert space* (RKHS) if the evaluation functional $\bar{\delta}_x\colon f \mapsto f(x)$ is continuous. The function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which satisfies the reproducing property

$$\begin{aligned} \langle f, k(x,\cdot)\rangle &= f(x) \quad \text{and in particular} \\ \langle k(x,\cdot), k(y,\cdot)\rangle &= k(x,y) \end{aligned}$$

is called the *reproducing kernel* of $\mathcal{H}$ (Steinwart and Christmann, 2008).

The map $\phi : \mathcal{X} \to \mathcal{H}$, $\phi\colon x \mapsto k(x,\cdot)$ with the property that

$$k(x,y) = \langle \phi(x), \phi(y)\rangle$$

is called *feature map*. Given a reproducing kernel, the *mean map* embeds a probability measure into an RKHS and is defined as follows.

Definition 1 (Kernel Embedding)**.** *The* kernel embedding *or* kernel mean map *of a measure* $\mathbb{P}$ *associated with the kernel function* $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is*

$$\begin{aligned} \mu\colon \mathcal{Q}(\mathcal{X}) &\to \mathcal{H} \\ \mu[\mathbb{P}] &= \int_{\mathcal{X}} k(x,\cdot)\,\mathrm{d}\mathbb{P}(x), \end{aligned}$$

where $\mathcal{Q}(\mathcal{X})$ *is the set of all finite Borel measures on* $\mathcal{X}$*.*

In the following, we assume that $\mathcal{X}$ is a separable measurable space and that the kernel k is continuous and bounded in expectation such that $\mu[\mathbb{P}]$ exists for all $\mathbb{P} \in \mathcal{Q}(\mathcal{X})$ (Sriperumbudur et al., 2011). We see that μ maps every distribution (measure) to a single point in the RKHS $\mathcal{H}$. This mapping is injective if k is characteristic (Sriperumbudur et al., 2008; Fukumizu et al., 2009; Sriperumbudur et al., 2011). However, we emphasize that the characteristic property is not relevant for this work.

Often, the underlying distribution $\mathbb{P}$ is unknown, but it is possible to draw independent samples $X_1, X_2, X_3, \ldots$ from $\mathbb{P}$. The empirical measure after n draws

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

where δ_x is the Dirac measure, can act as a proxy for $\mathbb{P}$. Hereby $\mathbb{P}_n$ is used to construct an approximation $\mu[\mathbb{P}_n]$ of $\mu[\mathbb{P}]$ as

$$\mu[\mathbb{P}] \approx \mu[\mathbb{P}_n] = \int_{\mathcal{X}} \phi(t)\; d\mathbb{P}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)$$

which is called *empirical kernel embedding* (Smola et al., 2007). This is reasonable a choice since the law of large numbers implies that $\mu[\mathbb{P}_n]$ converges almost surely to $\mu[\mathbb{P}]$ for $n \to \infty$ as $\mathbb{E}\big[\|\phi(X)\|\big]$ is bounded.

We may ask how well $\mu[\mathbb{P}_n]$ approximates $\mu[\mathbb{P}]$. An answer to this question is given by the following *concentration inequality*. We will see in Remark 1 that

$$\Pr\left(\big\|\mu[\mathbb{P}] - \mu[\mathbb{P}_n]\big\| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8d^2}\right) \qquad (1)$$

holds if $\|\phi(X)\| \leq d$ almost surely, bounding the probability that the empirical kernel embedding $\mu[\mathbb{P}_n]$ differs from the actual embedding $\mu[\mathbb{P}]$ using n independent samples. This is (up to some constants) a Banach space version of Hoeffding's inequality. Here $\|\cdot\|$ denotes the norm on $\mathcal{H}$ induced by the inner product $\langle\cdot,\cdot\rangle$.

Concentration inequalities are of wide interest in various applications.

> "The problem of providing exponential bounds for the probabilities $\Pr(\|S_n\| \geq \epsilon)$ $(\epsilon > 0)$ is of paramount importance, both in Probability and Statistics. From a statistical viewpoint, such inequalities can be used, among other things, for the purpose of providing rates of convergence (both in the probability sense and almost surely) for estimates of various quantities." (Roussas, 1996)

1.2 Large-Scale Kernel Embedding

In most machine learning and data-mining applications, the samples from $\mathbb{P}$ are given in form of a fixed dataset $(x_1, \ldots, x_{\mathrm{N}})$ of size $\mathrm{N} \in \mathbb{N}$. The empirical kernel embedding $\mu[\mathbb{P}_{\mathrm{N}}]$ is then typically estimated using all N samples from this set. However, for large-scale learning problems, $\mu[\mathbb{P}_{\mathrm{N}}]$ becomes infeasible to compute due to the linear computational complexity in N. As a consequence one might to consider to use only a subset of n samples from the dataset, where $1 \leq n \leq \mathrm{N}$.

It is therefore of central interest how much error we introduce into the empirical kernel embedding by subsampling the dataset. The approximation accuracy increases as more samples are used and obviously $\mu[\mathbb{P}_n] \to \mu[\mathbb{P}_{\mathrm{N}}]$ as $n \to \mathrm{N}$. The theory developed in this work can be applied to state probabilistic bounds for the difference

$$\|\mu[\mathbb{P}_n] - \mu[\mathbb{P}_{\mathrm{N}}]\|$$

for any $1 \leq n \leq \mathrm{N}$. The estimation of this quantity is well known in probability theory as *sampling without replacement* which is formally defined as follows.

Definition 2 (Sampling without Replacement)**.** *Let $(x_1, \dots, x_{\mathrm{N}})$ be a finite population of N elements. For any $n \in \mathbb{N}$ with $n \leq \mathrm{N}$, the realization $(x_{I_1}, \dots, x_{I_n})$ of the random variables $(X_1, \dots, X_n)$ with*

$$\Pr\left[(I_1, \dots, I_n) = (i_1, \dots, i_n)\right] = \left[\mathrm{N}(\mathrm{N}-1)\cdots(\mathrm{N}-n+1)\right]^{-1}$$

is said to be drawn without replacement *(Serfling, 1974).*

Following Hoeffding's argumentation (Hoeffding, 1963), it can be shown that Eq. (1) does also hold in the case of sampling without replacement and hence we can use it to bound the error $\|\mu[\mathbb{P}_n] - \mu[\mathbb{P}_{\mathrm{N}}]\|$ by

$$\Pr\left(\left\|\mu[\mathbb{P}_n] - \mu[\mathbb{P}_{\mathrm{N}}]\right\| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8d^2}\right), \qquad (2)$$

however we will see that much tighter bounds can be found if the finite size of the dataset is taken into account.

1.3 Contributions and Outline

Serfling (1974) already derived bounds in the case of sampling without replacement superior to Hoeffding's inequality, but only for real-valued random variables. However the random variables $Z = \phi(X)$ used to estimate the kernel mean map are elements of the reproducing kernel Hilbert space $\mathcal{H}$ which requires a more general theory of concentration inequalities.

The main contributions of this work are the following two theorems. In the former we derive probabilistic bounds in the setting of sampling without replacement for the sum $S_n = Z_1 + \cdots + Z_n$ of independent random variables $Z_1, Z_2, \dots$ taking values in a 2-smooth Banach space.

Theorem 1. *Let $(\mathcal{B}, \|\cdot\|)$ be a $(2, D)$-smooth*[1] *separable Banach space and the variables $(z_1, \dots, z_{\mathrm{N}})$ be elements of $\mathcal{B}$ with $(Z_1, \dots, Z_n)$ being a random sample without replacement. We assume that all $\|z_i\| \leq d$ almost surely with constant $d > 0$. Then*

$$\Lambda_n \leq 2\exp\left(-\frac{n\epsilon^2}{8D^2d^2\left(1 - \frac{n-1}{\mathrm{N}}\right)}\right),$$

where

$$\Lambda_n = \Pr\left(\max_{1\leq k\leq n}\left\|\frac{\mathrm{N}-n}{\mathrm{N}-k}\sum_{i=1}^{k}(Z_i - \bar{\mu})\right\| \geq n\epsilon\right)$$

for all $\epsilon > 0$, where $\bar{\mu} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} z_i$ is the mean of over all realizations.

[1] See Definition 3 for an explanation of $(2, D)$-smooth.

This concentration inequality is very general and can be applied to numerous problems in machine learning, probability theory and statistics. We will use Theorem 1 to derive probabilistic bounds for the difference of empirical kernel mean maps $\|\mu[\mathbb{P}_n] - \mu[\mathbb{P}_{\mathrm{N}}]\|$ which is subject of the second theorem.

Theorem 2. *Let $(x_1, \dots, x_{\mathrm{N}})$ be a realization with elements in $\mathcal{X}$ of the random sample without replacement $(X_1, \dots, X_n)$. Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the reproducing kernel of the RKHS $\mathcal{H}$ with $\|\phi(x)\| \leq d$ almost surely for all $(x_1, \dots, x_{\mathrm{N}})$ with constant $d > 0$. Then for all $\epsilon > 0$*

$$\Pr\left(\left\|\mu[\mathbb{P}_n] - \mu[\mathbb{P}_{\mathrm{N}}]\right\| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8d^2\left(1 - \frac{n-1}{\mathrm{N}}\right)}\right)$$

where

$$\mu[\mathbb{P}_n] = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i) \quad \textit{and} \quad \mu[\mathbb{P}_{\mathrm{N}}] = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\phi(x_i)$$

are elements of $\mathcal{H}$.

This is a probabilistic bound to quantify the error introduced by $\mu[\mathbb{P}_n]$ as a surrogate for $\mu[\mathbb{P}_{\mathrm{N}}]$. We see that the factor $(1 - \frac{n-1}{\mathrm{N}})$ in the equation above leads to a significant improvement over Eq. (2) in the setting of sampling without replacement.

Sampling without replacement is itself subject of various application such as survey sampling (Kish, 1965), Markov chain Monte Carlo algorithms (Bardenet et al., 2014) and computational learning theory (Cannon et al., 2002) to name a few. We emphasize that the kernel mean embedding is just one of many applications of Theorem 1.

The remainder of this paper is structured as follows:

Even though $\mu[\mathbb{P}_n]$ is the empirical mean of Hilbert space valued random variables $Z = \phi(X)$, we will derive more general bounds for Banach space valued martingales established in Theorem 1 and then show that Theorem 2 is only a special case thereof.

- Section 2 introduces the inequalities derived by Serfling for sampling without replacement.
- In Section 3 we generalize Serfling's inequality to Banach space valued martingale sequences which proves Theorem 1.
- The subsequent Section 4 presents the proof for Theorem 2 which is the concentration inequality for the kernel embedding $\mu[\mathbb{P}_n]$.

2 SERFLING'S INEQUALITY

Hoeffding's inequality (Hoeffding, 1963) provides an upper bound on the probability that a sum of independent, real-valued random variables $X_1 + \cdots + X_{\mathrm{N}}$ deviates from its expectation $\mathbb{E}[X_1 + \cdots + X_{\mathrm{N}}]$. The same upper bound also holds for the setting of sampling without replacement, where $X_1, \ldots, X_n$ is a random sample without replacement from the sample $(x_1, \ldots, x_{\mathrm{N}})$. Hoeffding's inequality is then

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} x_i\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right),$$

where $a = \min_{1\leq i\leq \mathrm{N}} x_i$ and $b = \max_{1\leq i\leq \mathrm{N}} x_i$.

Serfling (1974) introduced a tighter Hoeffding bound for this inequality replacing the factor n by $\frac{n}{(1-(n-1)/\mathrm{N})}$ in the bound on the right hand side of the equation above.

Theorem 3 (Serfling)**.** *Let $(x_1, \ldots, x_{\mathrm{N}}) \in \mathbb{R}^{\mathrm{N}}$, $\mathrm{N} \in \mathbb{N}$ be a realization of the random sample without replacement $(X_1, \ldots, X_n)$. Defining*

$$a = \min_{1\leq i\leq \mathrm{N}} x_i \quad \text{and} \quad b = \max_{1\leq i\leq \mathrm{N}} x_i$$

it holds for all $\epsilon > 0$ that

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} x_i \geq \epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{(1-\frac{n-1}{\mathrm{N}})(b-a)^2}\right).$$

Serfling did not provide a two-sided variant of this inequality, but this can easily be derived using symmetry and by adding a factor of 2 to the right hand side of the equation above. Bardenet and Maillard (2015) (Bardenet and Maillard, 2015) derived a Bernstein-type Serfling inequality for real-valued random variables which is based on the *variances* of $X_1, X_2, \ldots$ instead of their *bounds d*. A generalization of their work to Hilbert- or Banach-valued random variables is not straightforward and out of scope of this work.

In the following section we generalize Serfling's bound to random variables which take values in a separable Banach space instead of $\mathbb{R}$.

3 A BANACH SPACE INEQUALITY FOR SAMPLING WITHOUT REPLACEMENT

For the deviation of the concentration inequality in Theorem 1 we will exploit bounds for certain martingale structures derived by (Serfling, 1974). However, in contrast to Serfling we will not bound the moment generating function, but utilize Pinelis' theorem for Banach space valued martingales, which we will state next together with some preliminaries.

Let $\mathbb{N}_0 = \mathbb{N} \cup 0$ and $\{M_n\} = \{M_n\}_{n\in\mathbb{N}_0}$, $M_0 = 0$ be a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume the martingale takes values in a separable Banach space $(\mathcal{B}, \|\cdot\|)$ and denote the set of all such martingales with $\mathcal{M}(\mathcal{B})$ (Kallenberg, 2006).

Definition 3 ((r, D)-smooth Banach spaces)**.** *A Banach space $(\mathcal{B}, \|\cdot\|)$ is called (r, D)-*smooth *for some $1 < r \leq 2$, if there exists a $D > 0$ such that*

$$\|x + y\|^r + \|x - y\|^r \leq 2\|x\|^r + 2D^r\|y\|^r$$

for all $x, y \in \mathcal{B}$ and is simply called 2-smooth if it is $(2, D)$-smooth for some $D > 0$.

The notion 2-smooth Banach spaces is the analogue of the notion of a Banach space of type 2 in the case of martingale differences. For example any $\mathcal{L}^p$ space, $p > 1$ (of $\mathbb{R}$-valued functions), associated with a σ-finite measure is r-smooth for $r = \min(2, p)$. Then $D^2 = p - 1$ if $p \geq 2$ and $D^2 = 2$ if $1 \leq p < 2$. It can be shown that any Hilbert space is $(2, 1)$-smooth (Cuny, 2015).

Pinelis derived an inequality for the probability that the martingale $\{M_n\}$ escapes a ball with radius ϵ as described in (Pinelis, 1994, Theorem 3.5) and (Pinelis, 1992, Theorem 3).

Theorem 4 (Pinelis)**.** *Suppose that $\{M_n\} \in \mathcal{M}(\mathcal{B})$, where $\mathcal{B}$ is a $(2, D)$-smooth separable Banach space and $\sum_{i=1}^{\infty} \operatorname{ess\,sup}\|M_i - M_{i-1}\|^2 \leq c^2$ for some $c > 0$. Then*

$$\Pr\left(\sup\left\|\{M_n\}\right\| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2c^2D^2}\right)$$

for all $\epsilon > 0$.

The challenge is to choose a martingale sequence such that Pinelis theorem can be applied to the difference $\left\|\mu[\mathbb{P}_n] - \mu[\mathbb{P}_{\mathrm{N}}]\right\|$ where c then yield the desired bound with a dependence on n.

3.1 A Serfling Type Inequality for Banach Spaces

The remainder of this section is concerned with a bound for the martingale difference

$$\sum_{i=1}^{\infty} \operatorname{ess\,sup}\|M_i - M_{i-1}\|^2 \leq c^2$$

used at a later stage to prove Theorem 1. We start with the definition of a stochastic process and show that it is a martingale.

Lemma 1. *Let $(Z_1, Z_2, \dots)$ be random variables in a $(2, D)$-smooth separable Banach space. The stochastic process $\{M_k\}_{k\in\mathbb{N}_0}$ defined as*

$$M_k = \frac{1}{\mathrm{N}-k}\sum_{i=1}^{k}(Z_i - \bar{\mu}) \quad (1 \le k \le \mathrm{N})$$

with $M_0 = 0$ is a martingale.

Proof. Since Z_k is uniformly drawn from the remaining $\mathrm{N}-k+1$ points, it holds that for $1 \le k \le \mathrm{N}$

$$\begin{aligned}\mathbb{E}[Z_k \mid Z_1, \dots, Z_{k-1}] &= \frac{\sum_{i=k}^{\mathrm{N}} Z_i}{\mathrm{N}-k+1} \\ &= \frac{\mathrm{N}\bar{\mu} - \sum_{i=1}^{k-1} Z_i}{\mathrm{N}-k+1} \\ &= \bar{\mu} - \frac{\sum_{i=1}^{k-1} Z_i - (k-1)\bar{\mu}}{\mathrm{N}-k+1} \\ &= \bar{\mu} - M_{k-1}, \end{aligned} \tag{3}$$

where the expectation is taken with respect to the remaining elements $Z_k, \dots, Z_{\mathrm{N}}$. By definition we have

$$\begin{aligned}M_k &= \frac{1}{\mathrm{N}-k}\sum_{i=1}^{k}(Z_i - \bar{\mu}) \\ &= \frac{1}{\mathrm{N}-k}\sum_{i=1}^{k-1}(Z_i - \bar{\mu}) + \frac{Z_k - \bar{\mu}}{\mathrm{N}-k} \\ &= \frac{\mathrm{N}-k+1}{\mathrm{N}-k}M_{k-1} + \frac{Z_k - \bar{\mu}}{\mathrm{N}-k} \end{aligned} \tag{4}$$

and hence

$$\begin{aligned}\mathbb{E}[M_k \mid M_1, \dots, M_{k-1}] &= \frac{\mathrm{N}-k+1}{\mathrm{N}-k}M_{k-1} \\ &\quad + \frac{\mathbb{E}[Z_k \mid M_1, \dots, M_{k-1}] - \bar{\mu}}{\mathrm{N}-k} \\ &= \frac{\mathrm{N}-k+1}{\mathrm{N}-k}M_{k-1} \\ &\quad + \frac{\bar{\mu} - M_{k-1} - \bar{\mu}}{\mathrm{N}-k} \\ &= M_{k-1}\end{aligned}$$

where we used Eq. (3) in the second step. □

We can now bound the quantity $\|M_k - M_{k-1}\|$ using Eq. (4) from which we get

$$M_k = M_{k-1} + \frac{Z_k - \bar{\mu} + M_{k-1}}{\mathrm{N}-k}$$

and hence for $1 \le k \le \mathrm{N}$ it holds that

$$\begin{aligned}\|M_k - M_{k-1}\| &= \frac{1}{\mathrm{N}-k}\big\|Z_k - \bar{\mu} + M_{k-1}\big\| \\ &\le \frac{2d}{\mathrm{N}-k}.\end{aligned} \tag{5}$$

We now follow a similar analysis as (Serfling, 1974, Lemma 2.1) using the inequality

$$\sum_{r=k+1}^{m}\frac{1}{r^2} \le \frac{m-k}{k(m+1)}, \quad 1 \le k \le m \in \mathbb{N}, \tag{6}$$

and are now in a position to prove Theorem 1.

Proof of Theorem 1. Using the previous derivations we observe that

$$\begin{aligned}\frac{(N-n)^2}{4d^2}\sum_{k=1}^{n}\|M_k - M_{k-1}\|^2 &\le \sum_{k=1}^{n}\frac{(\mathrm{N}-n)^2}{(\mathrm{N}-k)^2} \\ &= 1 + (\mathrm{N}-n)^2\sum_{k=\mathrm{N}-n+1}^{\mathrm{N}-1}\frac{1}{k^2} \\ &\le 1 + \frac{(\mathrm{N}-n)^2(n-1)}{\mathrm{N}(\mathrm{N}-n)} \\ &= \frac{\mathrm{N} + (\mathrm{N}-n)(n-1)}{\mathrm{N}} \\ &= n\Big(1 - \frac{n-1}{\mathrm{N}}\Big)\end{aligned}$$

where we used Eq. (6) in the second last step. Hence the bound on the martingale difference is

$$\sum_{k=1}^{n}\big\|M_k - M_{k-1}\big\|^2 \le c^2$$

with

$$c^2 = \frac{\alpha^2 n}{(\mathrm{N}-n)^2} \quad \text{and} \quad \alpha^2 = 4d^2\Big(1 - \frac{n-1}{\mathrm{N}}\Big).$$

We now apply Pinelis' inequality (Theorem 4) and get

$$\Pr\left(\sup\big\|\{M_n\}\big\| \ge \lambda\right) \le 2\exp\left(-\frac{\lambda^2}{2D^2c^2}\right)$$

$$\Pr\left(\max_{1\le k\le n}\Big\|\frac{S_k - k\bar{\mu}}{\mathrm{N}-k}\Big\| \ge \lambda\right) \le 2\exp\left(-\frac{\lambda^2(\mathrm{N}-n)^2}{2D^2\alpha^2 n}\right).$$

Substituting $(\mathrm{N}-n)\lambda = n\epsilon$ in the equation above yields

$$\begin{aligned}\Lambda_n &\le 2\exp\left(-\frac{\epsilon^2}{2D^2\alpha^2 n}\right) \\ &\le 2\exp\left(-\frac{n\epsilon^2}{8D^2d^2\left(1 - \frac{n-1}{\mathrm{N}}\right)}\right),\end{aligned}$$

which concludes the proof. □

We will now show that Theorem 2 is a consequence of Theorem 1 if an appropriate martingale structure is applied.

4 KERNEL MEAN EMBEDDING INEQUALITIES

Using the results from the previous derivation we can now proof Theorem 2. We first remind that if $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ is a Hilbert space, then it is $(2,1)$-smooth.

Proof of Theorem 2. Let $Z_1, Z_2, \ldots$ be independent random variables in the Hilbert space $\mathcal{H}$, bounded by $\|Z_i\| \leq d$ almost surely. Then $\{M_k\} = \sum_{i=1}^{k} Z_i$ is obviously a martingale and $\|M_i - M_{i-1}\|^2 = \|Z_i\|^2 \leq d^2$ a.s.

For any fixed $n \in \mathbb{N}$ we now consider the martingale difference

$$M_i - M_{i-1} = \begin{cases} Z_i & \text{if } 1 \leq i \leq n \\ 0 & \text{otherwise,} \end{cases}$$

for which the bound

$$\sum_{i=1}^{\infty} \operatorname{ess\,sup} \|M_i - M_{i-1}\|^2 \leq nd^2$$

holds. From the elementary relation

$$\Pr\left(\left\| \frac{\mathrm{N}-n}{\mathrm{N}-n} \sum_{i=1}^{n} (Z_i - \bar{\mu}) \right\| \geq n\epsilon \right) \leq \Lambda_n$$

it follows that

$$\Pr\left(\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i - \bar{\mu} \right\| \geq \epsilon \right) \leq 2 \exp\left(- \frac{n\epsilon^2}{8d^2 \left(1 - \frac{n-1}{\mathrm{N}}\right)} \right)$$

by means of Theorem 1. Substituting

$$Z_i = \phi(X_i) \quad \text{and}$$
$$\bar{\mu} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \phi(x_i)$$

concludes the proof for Theorem 2. □

Remark 1. *Notice that from the discussion above it is also clear that Eq.* (1) *holds using Pinelis's bound (Theorem 4) with* $c^2 = nd^2$ *and* $D = 1$. *A bound similar to Eq.* (1) *can be found in (Smola et al., 2007) which states the following.*

> *Assume that* $\|f\|_\infty \leq d$ *for all* $f \in \mathcal{H}$ *with* $\|f\| \leq 1$. *Then*
>
> $$\left\| \mu[\mathbb{P}_n] - \mu[\mathbb{P}] \right\| \leq 2R_n(\mathcal{H}, \mathbb{P}) + d\sqrt{-n^{-1} \log \delta}$$
>
> *with probability at least* $1 - \delta$.

A rearrangement of variables yields an exponential inequality similar to Eq. (1). *However, this bound involves the Rademacher average* $R_n(\mathcal{H}, \mathbb{P})$ *(Bartlett and Mendelson, 2002) associated with the Hilbert space* $\mathcal{H}$ *and the measure* $\mathbb{P}$.

5 EMPIRICAL EVALUATION

In this section we empirically evaluate the deviation between the kernel embeddings $\mu[\mathbb{P}_n]$ and $\mu[\mathbb{P}_\mathrm{N}]$ as $n \to \mathrm{N}$. Since the equation

$$\begin{aligned} \left\| \mu[\mathbb{P}_n] - \mu[\mathbb{P}_\mathrm{N}] \right\|^2 &= \langle \mu[\mathbb{P}_n], \mu[\mathbb{P}_n] \rangle - 2\langle \mu[\mathbb{P}_n], \mu[\mathbb{P}_\mathrm{N}] \rangle \\ &\quad + \langle \mu[\mathbb{P}_\mathrm{N}], \mu[\mathbb{P}_\mathrm{N}] \rangle \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(X_i, X_j) \\ &\quad + \frac{1}{\mathrm{N}^2} \sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{N}} k(x_i, x_j) - \frac{2}{n\mathrm{N}} \sum_{i=1}^{n} \sum_{j=1}^{\mathrm{N}} k(X_i, x_j) \end{aligned}$$

holds, it is not necessary to explicitly represent $\phi(x)$ and we can calculate the error $\left\| \mu[\mathbb{P}_n] - \mu[\mathbb{P}_\mathrm{N}] \right\|$ exactly.

As data we take random instances from the MNIST (LeCun et al., 1998) and ImageNet (Russakovsky et al., 2015) datasets. The former contains images of handwritten digits and the latter is a database containing various images. We randomly sample $\mathrm{N} = 1000$ instances from each dataset and calculate $\left\| \mu[\mathbb{P}_n] - \mu[\mathbb{P}_\mathrm{N}] \right\|$ for $1 \leq n \leq \mathrm{N}$ using the squared exponential and Matérn kernel.

We use two different kernel functions. First, the *squared exponential* kernel given by

$$k(x, y) = \alpha \exp\left(\frac{\|x - y\|^2}{2\sigma^2} \right)$$

with *scaling factor* α and *length scale* σ^2, where $\alpha, \sigma^2 > 0$. The second kernel is the Matérn (Minasny and McBratney, 2005) defined as

$$k(x, y) = \frac{\alpha}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{2\sqrt{\nu}\|x - y\|}{\sigma^2} \right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu}\|x - y\|}{\sigma^2} \right),$$

where Γ is the gamma function and K_ν is the modified Bessel function of the second kind. In the following we set $\nu = \frac{3}{2}$, $\sigma^2 = 10$ (MNIST) and $\sigma^2 = 100$ (ImageNet) to obtain reasonable kernel values. For both kernels, we vary the scaling factor α such that $\|\phi(x)\| \leq d$, for $d = 1, 3, 5$ to gain insight how the convergence is influenced by this bound. A variation of the bandwidth σ^2 does not have any influence on the bound d and hence does not affect our inequalities.

The results of 10 independent experiments are illustrated in Fig. 1, where we report the mean $\pm 2\times$ variance. As reflected in Theorem 2, the convergence rate becomes slower as $\|\phi(x)\|$ grows and we also observe an increased variance. The influence of this bound dominates over the influence of the kernel characteristics and the dataset properties.

6 CONCLUSION

We provide concentration inequalities for Banach space valued martingales in *sampling without replacement*. These probabilistic bounds are a generalization of Serflings work, which has numerous applications in probability, statistics and machine learning. Our inequalities do not use any knowledge about the underlying distribution other than the requirement of bounded martingale differences and are therefore applicable to a wide class of problems.

We apply our new theory to estimate errors for the empirical kernel embedding of distributions which is itself a Hilbert space valued random variable. This embedding is done with the kernel mean map and used to manipulate probability distributions without the need of an explicit analytical representation. We focus on the error introduced when only a subset of the data is used as a proxy for the full dataset. This is of special interest for very large-scale applications where it is, due to computational limitations, not possible to process all available data. This is just one of various potential applications of these new concentration inequalities.

References

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning*, pages 405–413, 2014.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3:463–482, 2002. ISSN 15324435.

Adam Cannon, J Mark Ettinger, Don Hush, and Clint Scovel. Machine learning with data dependent hypothesis classes. *The Journal of Machine Learning Research*, 2:335–358, 2002.

Kacper Chwialkowski and Arthur Gretton. A Kernel Independence Test for Random Processes. In *Proceedings of 31st International Conference on Machine Learning*, 2014. URL `http://arxiv.org/abs/1402.4501`.

Christophe Cuny. A compact LIL for martingales in 2-smooth Banach spaces with applications. *Bernoulli*, 21(1):374–400, 2015.

Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Bharath K Sriperumbudur. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems*, pages 473–480, 2009.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14:3753–3783, 2013. URL `http://jmlr.org/papers/v14/fukumizu13a.html`.

Arthur Gretton, Olivier Bousquet, Alexander Johannes Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012a.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012b.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

Leslie Kish. *Survey sampling*. John Wiley and Sons, New York, 1965.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11):2278–2324, 1998.

Budiman Minasny and Alex B McBratney. The Matérn function as a general model for soil variograms. *Geoderma*, 128(3):192–207, 2005.

Iosif Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. In *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference*, pages 128–134. Springer, 1992.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

George G Roussas. Exponential probability inequalities with some applications. *Lecture Notes-Monograph Series*, pages 303–319, 1996.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge.

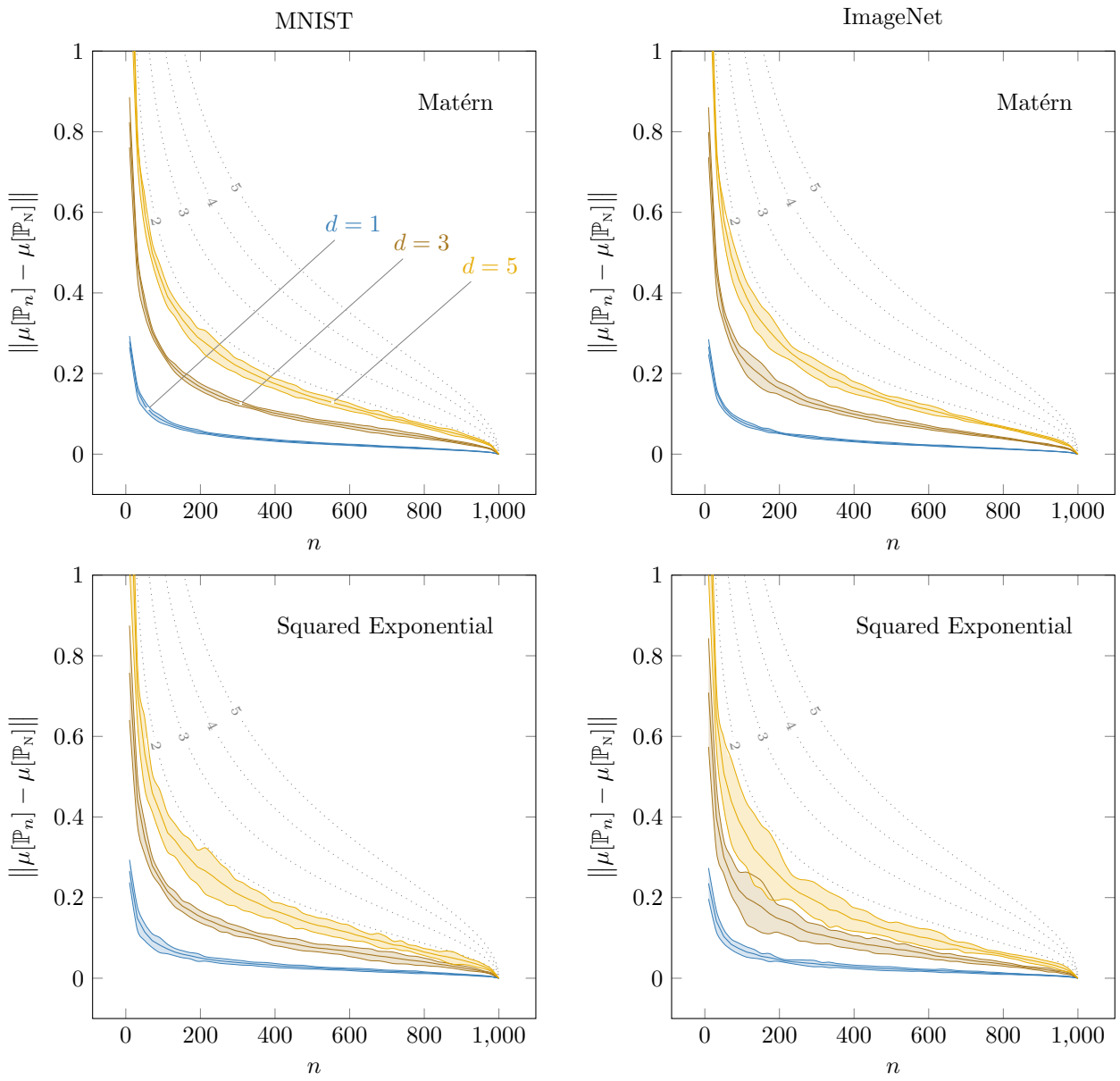

Figure 1: This figure shows plots for the deviation between $\mu[\mathbb{P}_n]$ and $\mu[\mathbb{P}_N]$ as $n \to N$ for different datasets and different kernels. Each kernel is scaled such that $\|\phi(x)\| \leq d$, for $d = 1, 3, 5$. We repeated each experiment several times and plot the mean (solid line) $\pm 2\times$ variance (shaded region). The theoretical upper bounds for various d are plotted in gray (dotted line) such that $\left\|\mu[\mathbb{P}_n] - \mu[\mathbb{P}_N]\right\| \leq \epsilon$ with 95% probability. Each column shows a different dataset (MNIST and ImageNet) and each row a different kernel (Matérn and Squared Exponential).

International Journal of Computer Vision, pages 1–42, apr 2015.

Robert J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2:39–48, 1974. ISSN 0090-5364.

Alex J Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

Le Song, Jonathan Huang, Alexander Johannes Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009.

Le Song, Byron Boots, Sajid M Siddiqi, Geoffrey J Gordon, and Alex J Smola. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th international conference on machine learning*, pages 991–998, 2010.

Le Song, Arthur Gretton, Danny Bickson, Yucheng Low, and Carlos Guestrin. Kernel belief propagation. *International Conference on Artificial Intelligence and Statistics*, pages 707–715, 2011.

Le Song, Alexander Johannes Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. *Proceedings of the 21st Annual Conference on Learning Theory*, pages 111–122, 2008.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert R G Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *The Journal of Machine Learning Research*, 12:2389–2410, 2011.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer, 2008.