Statistics and Probability Letters 173 (2021) 109071

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On concentration inequalities for vector-valued Lipschitz functions

Dimitrios Katselis [a,*], Xiaotian Xie [b], Carolyn L. Beck [b], R. Srikant [c]

[a] ECE Department, University of Illinois at Urbana-Champaign, USA
[b] ISE Department and Coordinated Science Lab, University of Illinois at Urbana-Champaign, USA
[c] ECE Department and Coordinated Science Lab, University of Illinois at Urbana-Champaign, USA

ARTICLE INFO

Article history:
Received 14 July 2020
Received in revised form 11 February 2021
Accepted 17 February 2021
Available online 3 March 2021

MSC:
00-01
99-00

Keywords:
Theorem of Bobkov and Götze
Concentration
Markov chain
Transportation cost inequality

ABSTRACT

We derive two upper bounds for the probability of deviation of a vector-valued Lipschitz function of a collection of random variables from its expected value. The resulting upper bounds can be tighter than bounds obtained by a direct application of a classical theorem due to Bobkov and Götze.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

In many statistical settings, vector-valued estimators naturally arise and determining their statistical rates is essential. As an example, assume that $X = (X_1, \ldots, X_n)$ corresponds to a random sample drawn from a product measure $\nu_\theta^{\otimes n}$, denoted by $X \sim \nu_\theta^{\otimes n}$, where θ is a parameter vector taking values in a compact set $\Theta \subset \mathbb{R}^k$ with a fixed dimension k independent of n. In this case, $\hat{\theta}_n = f(X)$ is a candidate vector-valued estimator of θ, where f is a measurable function with respect to X. In such settings, we are interested in knowing how close $f(X)$ is to its expectation $E_\nu[f(X)]$. Concentration inequalities for $f(X)$ when the deviation from $E_\nu[f(X)]$ is measured in terms of a metric, often norm-induced, are important.

The derivation of concentration bounds relies on imposing smoothness conditions on f, which guarantee that f is not very sensitive to any particular coordinate variable (Boucheron et al., 2013; Raginsky and Sason, 2014; Van Handel, 2016). This sensitivity is quantified either locally via gradients or globally via Lipschitz properties of f. Marton introduced the *transportation method* to establish concentration of measure for product measures and Markov chains (Marton, 1986, 1996) by showing that transportation cost inequalities can be used to deduce concentration. In Bobkov and Götze (1999), Bobkov and Götze extended Marton's argument into an *equivalence* by showing that Wasserstein distances and relative entropies are comparable only when the moment-generating functions of *real-valued* Lipschitz functions defined with respect to the underlying metric can be controlled and vice versa. The connection between moment-generating functions and the two

* Corresponding author.
E-mail addresses: katselis@illinois.edu (D. Katselis), xx5@illinois.edu (X. Xie), beck3@illinois.edu (C.L. Beck), rsrikant@illinois.edu (R. Srikant).

https://doi.org/10.1016/j.spl.2021.109071
0167-7152/© 2021 Elsevier B.V. All rights reserved.

aforementioned indices of closeness of probability measures is established via the Gibbs variational principle (Van Handel, 2016, Lemma 4.10), (Dembo and Zeitouni, 1998) or alternatively, via the Donsker–Varadhan lemma (Rezakhanlou, 2015).

A key example in showing the connection between concentration of measure and Lipschitz functions is McDiarmid's or *bounded-difference* inequality (McDiarmid, 1997), traditionally viewed as a result of the *martingale approach* in establishing concentration (Boucheron et al., 2013; Raginsky and Sason, 2014; Van Handel, 2016). Let each X_i take values in a measurable space $\mathcal{X}_i$ and equip the product space $\mathbb{X} = \mathcal{X}_1 \times \cdot \times \mathcal{X}_n$ with the weighted Hamming metric $d_c(x, y) = \sum_{i=1}^n c_i \mathbb{1}\{x_i \neq y_i\}$, where $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{X}$. For $X \in \mathbb{X}$ with independent entries, $f(X)$ is a sub-Gaussian random variable with parameter $\sigma^2 = \sum_{i=1}^n c_i^2/4$ for every $f : \mathbb{X} \to \mathbb{R}$ which is 1-Lipschitz with respect to d_c by McDiarmid's inequality. Motivated by this example, a question of interest is for which measures ν on a metric space $(\mathbb{X}, d)$ such that $X \sim \nu$, the random variable $f(X)$ is σ^2-sub-Gaussian for every *real-valued* 1-Lipschitz function $f : \mathbb{X} \to \mathbb{R}$. The answer to this question is given by the aforementioned theorem of Bobkov and Götze (1999). In this paper, we focus on Lipschitz mappings $f : \mathbb{X} \to \mathbb{Y}$ between metric spaces $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$, where $\mathbb{Y} \subseteq \mathbb{R}^k$ and $d_{\mathbb{Y}}$ is any ℓ_p-metric for $p \geq 1$. In the spirit of Marton (1986, 1996), Bobkov and Götze (1999), we prove a concentration inequality when a transportation cost inequality can be shown to hold. We then provide a simple stationary measure estimation example for Markov chains demonstrating that the derived inequality gives better results than directly applying the theorem of Bobkov and Götze or by combining the aforementioned theorem with Boole's inequality. An interesting observation regarding the role of the particular ℓ_p-norm is also highlighted via this example.

2. Preliminaries and main result

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a Polish space and ρ be a (Borel) probability measure on $(\mathbb{X}, d_{\mathbb{X}})$. The triplet $(\mathbb{X}, d_{\mathbb{X}}, \rho)$ defines a metric probability space in the sense of Gromov (Raginsky and Sason, 2014, p. 103), (Gromov, 2007). Given two metric spaces $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$, let $f : \mathbb{X} \to \mathbb{Y}$ be a Lipschitz mapping with Lipschitz constant $\|f\|_{\text{Lip}}$, i.e., $d_{\mathbb{Y}}(f(x), f(\tilde{x})) \leq \|f\|_{\text{Lip}} d_{\mathbb{X}}(x, \tilde{x})$, $\forall x, \tilde{x} \in \mathbb{X}$. In the following, the set of all such mappings will be denoted by $\text{Lip}(\mathbb{X}, \mathbb{Y}, d_{\mathbb{X}}, d_{\mathbb{Y}})$. Moreover, let $\mathbb{P}_1(\mathbb{X})$ denote the set of all probability measures ρ on $\mathbb{X}$ such that $E_\rho[d_{\mathbb{X}}(X, x_0)] < \infty$ holds for an arbitrary (and therefore for all) $x_0 \in \mathbb{X}$. The L^1 Wasserstein distance between $\rho, \tilde{\rho} \in \mathbb{P}_1(\mathbb{X})$ is defined as

$$W_1(\rho, \tilde{\rho}) = \inf_{X \sim \rho, \tilde{X} \sim \tilde{\rho}} E[d_{\mathbb{X}}(X, \tilde{X})] = \inf_{\gamma \in \Pi(\rho, \tilde{\rho})} \int_{\mathbb{X} \times \mathbb{X}} d_{\mathbb{X}}(x, \tilde{x}) \gamma(dx, d\tilde{x}). \tag{1}$$

The infimum in the first part is taken over all jointly distributed pairs $(X, \tilde{X})$ on the product space $\mathbb{X}^2 = \mathbb{X} \times \mathbb{X}$ with marginals $\rho, \tilde{\rho}$, respectively. In the last part, $\Pi(\rho, \tilde{\rho})$ denotes the set of all possible couplings of $\rho, \tilde{\rho}$. An optimal coupling $\gamma^* \in \Pi(\rho, \tilde{\rho})$ achieving the infimum exists (Raginsky and Sason, 2014, p. 109), (Villani, 2008, Ch. 6). Additionally, a different measure gauging the dissimilarity between two probability measures $\rho, \tilde{\rho}$ is the relative entropy or Kullback–Leibler divergence $D(\tilde{\rho} \parallel \rho) = E_{\tilde{\rho}}\left[\log \frac{d\tilde{\rho}}{d\rho}\right] = E_\rho\left[\frac{d\tilde{\rho}}{d\rho} \log \frac{d\tilde{\rho}}{d\rho}\right]$ for $\tilde{\rho} \ll \rho$ and $D(\tilde{\rho} \parallel \rho) = \infty$ otherwise, where $d\tilde{\rho}/d\rho$ is the Radon–Nikodym derivative of $\tilde{\rho}$ with respect to ρ and $\tilde{\rho} \ll \rho$ denotes that $\tilde{\rho}$ is absolutely continuous with respect to ρ.

The following theorem provides a concentration inequality for $f(X)$ when the deviation from $E_\nu[f(X)]$ is measured in terms of the ℓ_2-metric.

Theorem 1. *Let $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$ be two Polish spaces, where $\mathbb{Y} \subseteq \mathbb{R}^k$ and $d_{\mathbb{Y}}(y, \tilde{y}) = \|y - \tilde{y}\|_2$ with $\|\cdot\|_2$ being the Euclidean norm. Let X be a random variable taking values in $\mathbb{X}$ and assume that $X \sim \nu$, where ν is a probability measure on $(\mathbb{X}, d_{\mathbb{X}})$. Then, the inequality $W_1(\mu, \nu) \leq \sqrt{2\sigma^2 D(\mu \parallel \nu)}, \forall \mu$ implies that $\forall \epsilon > 0$, $\forall \varepsilon \in (0, 1]$ and for any $f \in \text{Lip}(\mathbb{X}, \mathbb{Y}, d_{\mathbb{X}}, \|\cdot\|_2)$ such that $\|E_\nu[f]\|_\infty < \infty$,*

$$P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) \leq \min\left\{\left(1 + \frac{2}{\varepsilon}\right)^k e^{-\frac{\epsilon^2(1-\varepsilon)^2}{2\sigma^2\|f\|_{\text{Lip}}^2}}, 2^{\frac{k}{2}} e^{-\frac{\epsilon^2}{4\sigma^2\|f\|_{\text{Lip}}^2}}\right\}. \tag{2}$$

The proof of this theorem is provided in Section 3. This result can be straightforwardly extended to any ℓ_p-metric for $p \geq 1, p \neq 2$. Define

$$\tau_p = \sup_{y \in f(\mathbb{X})} \frac{\|y - E_\nu[f(X)]\|_p}{\|y - E_\nu[f(X)]\|_2}. \tag{3}$$

Here, y is assumed different from $E_\nu[f(X)]$ if $E_\nu[f(X)] \in f(\mathbb{X})$, since for $y = E_\nu[f(X)]$ (if $E_\nu[f(X)] \in f(\mathbb{X})$) the inequality $\|y - E_\nu[f(X)]\|_p \leq \tau_p \|y - E_\nu[f(X)]\|_2$ trivially holds for any $\tau_p > 0$. Then, (2) can be replaced by

$$P(\|f(X) - E_\nu[f(X)]\|_p \geq \epsilon) \leq \min\left\{\left(1 + \frac{2}{\varepsilon}\right)^k e^{-\frac{\epsilon^2(1-\varepsilon)^2}{2\tau_p^2\sigma^2\|f\|_{\text{Lip}}^2}}, 2^{\frac{k}{2}} e^{-\frac{\epsilon^2}{4\sigma^2\tau_p^2\|f\|_{\text{Lip}}^2}}\right\} \tag{4}$$

due to $\{\|f(X) - E_\nu[f(X)]\|_p \geq \epsilon\} \subseteq \{\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon/\tau_p\}$ by (3).

3. Proof of Theorem 1 and Additional Results

Proof of the first bound in (2). The proof of the first bound relies on covering arguments; see Vershynin (2018), Wainwright (2019), Lattimore and Szepesvári (2020) and references therein for results based on such arguments. For some ε in the interval $(0, 1]$, let $\mathcal{N}(\varepsilon)$ be an ε-net of the unit Euclidean sphere $\mathbb{S}^{k-1}$ in $\mathbb{R}^k$ with cardinality $|\mathcal{N}(\varepsilon)| \leq \left(1+\frac{2}{\varepsilon}\right)^k$ (Vershynin, 2018, Corollary 4.2.13). Additionally, by Exercise 4.4.2 in Vershynin (2018),

$$\|f(X) - E_\nu[f(X)]\|_2 \leq \frac{1}{1-\varepsilon} \sup_{w\in\mathcal{N}(\varepsilon)} \langle w, f(X) - E_\nu[f(X)]\rangle \quad \text{a.s.} \tag{5}$$

By (5) and by Boole's inequality (union bound) we have that

$$\begin{aligned} P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) &\leq P\left(\sup_{w\in\mathcal{N}(\varepsilon)} \langle w, f(X) - E_\nu[f(X)]\rangle \geq \epsilon(1-\varepsilon)\right) \\ &\leq \sum_{w\in\mathcal{N}(\varepsilon)} P\left(\langle w, f(X) - E_\nu[f(X)]\rangle \geq \epsilon(1-\varepsilon)\right) \\ &\leq |\mathcal{N}(\varepsilon)| \inf_{\lambda>0} e^{-\lambda\epsilon(1-\varepsilon)} E_\nu\left[e^{\lambda\langle w_*, f(X)-E_\nu[f(X)]\rangle}\right], \end{aligned} \tag{6}$$

where the last inequality follows from the Chernoff bound (Raginsky and Sason, 2014, p. 14) and

$$w_* = \arg\max_{w\in\mathcal{N}(\varepsilon)} P\left(\langle w, f(X) - E_\nu[f(X)]\rangle \geq \epsilon(1-\varepsilon)\right).$$

We now note that the function $\langle w_*, f(x) - E_\nu[f(X)]\rangle$ is $\|f\|_{\text{Lip}}$-Lipschitz by the Cauchy–Schwarz inequality and $\langle w_*, f(X) - E_\nu[f(X)]\rangle$ is mean zero. Assuming that $W_1(\mu, \nu) \leq \sqrt{2\sigma^2 D(\mu \parallel \nu)}$, $\forall \mu$ holds, an application of the theorem of Bobkov and Götze implies that

$$P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) \leq |\mathcal{N}(\varepsilon)| \inf_{\lambda>0} e^{-\lambda\epsilon(1-\varepsilon) + \frac{\lambda^2\sigma^2\|f\|_{\text{Lip}}^2}{2}}. \tag{7}$$

The exponent in the right-hand side of (7) is minimized for $\lambda_* = \frac{\epsilon(1-\varepsilon)}{\sigma^2\|f\|_{\text{Lip}}^2} > 0$ leading to

$$P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) \leq |\mathcal{N}(\varepsilon)| e^{-\frac{\epsilon^2(1-\varepsilon)^2}{2\sigma^2\|f\|_{\text{Lip}}^2}}. \tag{8}$$

By employing the bound on $|\mathcal{N}(\varepsilon)|$ the desired result follows.

Proof of the second bound in (2). Using a similar argument as in Hsu et al. (2012, Theorem 2.1), let $Z \sim \mathcal{N}(0, I_k)$ be a standard Gaussian random vector, which is independent of X. Recall that $E[e^{\langle Z, q\rangle}] = e^{\frac{\|q\|_2^2}{2}}$, $\forall q \in \mathbb{R}^k$. For any $\lambda \in \mathbb{R}$ we note that

$$\begin{aligned} &E\left[e^{\lambda\langle Z, f(X)-E_\nu[f]\rangle}\right] \geq \\ &E\left[e^{\lambda\langle Z, f(X)-E_\nu[f]\rangle} \,\middle|\, \|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon\right] P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) = \\ &E_\nu\left[E_Z\left[e^{\lambda\langle Z, f(X)-E_\nu[f]\rangle}\right] \,\middle|\, \|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon\right] P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) = \\ &E_\nu\left[e^{\frac{\lambda^2\|f(X)-E_\nu[f]\|_2^2}{2}} \,\middle|\, \|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon\right] P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) \geq \\ &e^{\frac{\lambda^2\epsilon^2}{2}} P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) \end{aligned}$$

or

$$P(\|f(X) - E_\nu[f(X)]\|_2 \geq \epsilon) \leq e^{-\frac{\lambda^2\epsilon^2}{2}} E\left[e^{\lambda\langle Z, f(X)-E_\nu[f]\rangle}\right]. \tag{9}$$

We now note that the function $\langle Z, f(x) - E_\nu[f(X)]\rangle$ is $\|Z\|_2\|f\|_{\text{Lip}}$-Lipschitz when Z is fixed (by the Cauchy–Schwarz inequality) and $\langle Z, f(X) - E_\nu[f(X)]\rangle$ is mean zero. Assuming that $W_1(\mu, \nu) \leq \sqrt{2\sigma^2 D(\mu \parallel \nu)}$, $\forall \mu$ holds, an application of the theorem of Bobkov and Götze implies that

$$E\left[e^{\lambda\langle Z, f(X)-E_\nu[f]\rangle}\right] = E_Z\left[E_\nu\left[e^{\lambda\langle Z, f(X)-E_\nu[f]\rangle}\right]\right] \leq E_Z\left[e^{\frac{\lambda^2\|Z\|_2^2\sigma^2\|f\|_{\text{Lip}}^2}{2}}\right].$$

Moreover, $\|Z\|_2^2$ is a chi-squared random variable with k degrees of freedom and for such a random variable $E\left[e^{t\|Z\|_2^2}\right] = \frac{1}{(1-2t)^{k/2}}, \quad t < \frac{1}{2}$. Therefore, we conclude that

$$E\left[e^{\lambda\langle Z, f(X)-E_\nu[f]\rangle}\right] \leq E_Z\left[e^{\frac{\lambda^2\|Z\|_2^2\sigma^2\|f\|_{\text{Lip}}^2}{2}}\right] = \frac{1}{\left(1-\lambda^2\sigma^2\|f\|_{\text{Lip}}^2\right)^{k/2}}, \tag{10}$$

where the last equality holds for any $\lambda \in \mathbb{R}$ such that $\lambda^2\sigma^2\|f\|_{\text{Lip}}^2 < 1$. By combining (9) and (10) we obtain

$$P(\|f(X)-E_\nu[f(X)]\|_2 \geq \epsilon) \leq \frac{e^{-\frac{\lambda^2\epsilon^2}{2}}}{\left(1-\lambda^2\sigma^2\|f\|_{\text{Lip}}^2\right)^{k/2}}, \quad |\lambda| < \frac{1}{\sigma\|f\|_{\text{Lip}}}. \tag{11}$$

Choosing $\lambda = \frac{1}{\sqrt{2}\sigma\|f\|_{\text{Lip}}}$, we conclude that

$$P(\|f(X)-E_\nu[f(X)]\|_2 \geq \epsilon) \leq 2^{\frac{k}{2}} e^{-\frac{\epsilon^2}{4\sigma^2\|f\|_{\text{Lip}}^2}}. \quad \square \tag{12}$$

In the previous proof and more specifically in (7) and (10), the theorem of Bobkov and Götze for the mean zero, vector-valued function $f(X) - E_\nu[f(X)]$ has been applied via a real-valued function of the form $\langle g, f(X) - E_\nu[f(X)]\rangle$ with $g = w_*$ and $g = Z$, respectively. This suggests the following extension of the theorem of Bobkov and Götze for vector-valued Lipschitz functions:

Proposition 1 (*Theorem of Bobkov and Götze for Vector-Valued Functions*)**.** *Let $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$ be two Polish spaces, where $\mathbb{Y} \subseteq \mathbb{R}^k$ and $d_{\mathbb{Y}}$ is an ℓ_p-metric for some $p \geq 1$. Let X be a random variable taking values in $\mathbb{X}$ and assume that $X \sim \nu$, where ν is a probability measure on $(\mathbb{X}, d_{\mathbb{X}})$. Then, the following statements are equivalent:*

1. $W_1(\mu, \nu) \leq \sqrt{2\sigma^2 D(\mu \parallel \nu)}, \forall \mu$
2. *$f(X) - E_\nu[f(X)]$, for every function $f \in \text{Lip}(\mathbb{X}, \mathbb{Y}, d_{\mathbb{X}}, d_{\mathbb{Y}})$ such that $\|E_\nu[f]\|_\infty < \infty$, is a sub-Gaussian vector with $\sigma^2\|h\|_q^2\|f\|_{\text{Lip}}^2$-sub-Gaussian one-dimensional marginals $\langle h, f(X) - E_\nu[f(X)]\rangle$, i.e.,*

$$E_\nu\left[e^{\lambda\langle h, f(X)-E_\nu[f(X)]\rangle}\right] \leq e^{\frac{\lambda^2\|h\|_q^2\sigma^2\|f\|_{\text{Lip}}^2}{2}}, \forall h \in \mathbb{R}^k, \forall \lambda \in \mathbb{R}. \tag{13}$$

Here, $\|\cdot\|_q$ corresponds to the dual norm of $\|\cdot\|_p$.

The direction "1. **implies** 2." can be obtained by the theorem of Bobkov and Götze for the real-valued function $\langle h, f(x) - E_\nu[f(X)]\rangle$ by invoking Hölder's inequality to show that the corresponding Lipschitz constant is at most $\|h\|_q\|f\|_{\text{Lip}}$. The direction "2. **implies** 1." is a direct consequence of the theorem of Bobkov and Götze by choosing f to have only one nonzero coordinate, e.g., $f = \tilde{f}e_1$ and $h = e_1$. Here, $\tilde{f} : \mathbb{X} \to \mathbb{R}$ is any real-valued $\|f\|_{\text{Lip}}$-Lipschitz function and e_1 is the first element of the canonical basis in $\mathbb{R}^k$.

4. Example

Consider a sample $X_{0:n-1} = \{X_0, X_1, \ldots, X_{n-1}\}$ of size n drawn from an ergodic, discrete-time, finite-state Markov chain $(X_k)_{k\geq 0}$ with state space $\mathbb{E} = [K] = \{1, 2, \ldots, K\}$, transition matrix $\mathrm{P} = [\mathrm{P}_{ij}]$ and $X_0 \sim \varrho$, where ϱ denotes the initial measure of the chain. We denote such a Markov chain by (P, ϱ) and the corresponding stationary chain by (P, π), where π is the underlying invariant measure. Let the chain be r-contractive with Dobrushin coefficient $r < 1$. Consider the natural plug-in estimators for the stationary probabilities:

$$\hat{\pi}_i(X_{0:n-1}) = \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}(X_k = i).$$

By the Ergodic Theorem for Markov chains (Brémaud, 2013), $\hat{\pi}_i \to \pi_i, \forall i \in \mathbb{E}$ with probability 1 as $n \to \infty$ for any initial measure ϱ.

The distance of (P, ϱ) from stationarity can be quantified by the (nonstationarity) index (Paulin, 2015, p. 11)

$$\left\|\frac{\varrho}{\pi}\right\|_{2,\pi}^2 = E_\pi\left[\left(\frac{d\varrho}{d\pi}\right)^2\right] = \sum_{i\in\mathbb{E}}\left[\frac{\varrho(i)}{\sqrt{\pi(i)}}\right]^2, \tag{14}$$

where the first equality corresponds to the general definition of the index for $\varrho \ll \pi$ and the second equality is the specialization of this definition to our setting. Furthermore, $1 \leq \|\varrho/\pi\|_{2,\pi} \leq \infty$ and $\|\cdot\|_{2,\pi}$ is the norm induced by the inner product $\langle f, g\rangle_\pi = \sum_{i\in\mathbb{E}} f(i)g(i)\pi(i)$ in $\ell^2(\pi)$ (Levin and Peres, 2017). Due to ergodicity, $\min_{i\in\mathbb{E}}\pi(i) > 0$ and also $\|\varrho/\pi\|_{2,\pi} \leq 1/\sqrt{\min_{i\in\mathbb{E}}\pi(i)}$. Additionally, $\|\varrho/\pi\|_{2,\pi} = 1$ for $\varrho = \pi$ and $\|\varrho/\pi\|_{2,\pi} = \infty$ if ϱ is not absolutely continuous with respect to π.

The index in (14) is useful in our context due to the following theorem (Paulin, 2015, Proposition 3.10):

Theorem 2. *Let $X_{0:n-1}$ be a sample drawn from a time-homogeneous Markov chain (P, ϱ) with state space $\mathbb{E}$ and stationary measure π. Then for any measurable function $g : \mathbb{E}^n \to \mathbb{R}$ and $\forall \epsilon > 0$,*

$$P_\varrho\left(g(X_{0:n-1}) \geq \epsilon\right) \leq \left\| \frac{\varrho}{\pi} \right\|_{2,\pi} \sqrt{P_\pi(g(X_{0:n-1}) \geq \epsilon)},$$

where P_ϱ is the law of (P, ϱ) and P_π is the law of (P, π).

Our goal is to bound $P_\varrho(\|\hat{\pi} - \pi\|_p \geq \epsilon) = P_\varrho(\|\hat{\pi} - E_\pi[\hat{\pi}]\|_p \geq \epsilon)$ for $p \geq 1$. To tackle the problem within the transportation method framework, we will use the following theorem due to Marton (Marton, 1996, Proposition 1) adapted to our setting:

Theorem 3. *Consider a Markov chain $(X_k)_{k\geq 0}$ with a finite state space $\mathcal{X}$, transition matrix $P = [P_{ij}]$ and Dobrushin coefficient $r < 1$. For $x_{0:n-1}, \tilde{x}_{0:n-1} \in \mathcal{X}^n$ let $\tilde{d}_{1,\mathcal{X}^n}(x_{0:n-1}, \tilde{x}_{0:n-1}) = \sum_{k=0}^{n-1} \mathbb{1}(x_k \neq \tilde{x}_k)$. Then,*

$$W_1(\mu, \nu) \leq \left[\frac{n}{2(1-r)^2} D(\mu \parallel \nu) \right]^{1/2},$$

where $\nu = P_\varrho$ is the measure on $(\mathbb{X} = \mathcal{X}^n, d_{\mathbb{X}} = \tilde{d}_{1,\mathcal{X}^n})$ due to the Markov chain starting at some arbitrary initial measure ϱ and μ is any measure on $(\mathbb{X}, d_{\mathbb{X}})$.

We now compare different approaches for bounding $P_\varrho(\|\hat{\pi} - \pi\|_p \geq \epsilon)$ and show that the bound obtained in Theorem 1 gives better results than other, more direct applications of the theorem of Bobkov and Götze.

Approach 1: Direct application of the theorem of Bobkov and Götze. Let $f(x) = f(x_{0:n-1}) = \|\hat{\pi}(x_{0:n-1}) - \pi\|_p$. Assume that $x = x_{0:n-1}$ and $\tilde{x} = \tilde{x}_{0:n-1}$ are two realizations of the random sequence $X_{0:n-1}$, which differ at a single element. By employing the reverse triangle inequality for the ℓ_p-norm we obtain

$$\begin{aligned} \left|f(x) - f(\tilde{x})\right| &\leq \|\hat{\pi}(x_{0:n-1}) - \hat{\pi}(\tilde{x}_{0:n-1})\|_p \\ &\leq \frac{\sqrt[p]{2}}{n} = \frac{\sqrt[p]{2}}{n} \tilde{d}_{1,\mathbb{E}^n}(x_{0:n-1}, \tilde{x}_{0:n-1}) = \frac{\sqrt[p]{2}}{n} \sum_{k=0}^{n-1} \mathbb{1}(x_k \neq \tilde{x}_k). \end{aligned} \tag{15}$$

Clearly, (15) implies that $\|f\|_{\mathrm{Lip}} = \sqrt[p]{2}/n$ for any $x_{0:n-1}, \tilde{x}_{0:n-1}$ (not necessarily different at a single element). This can be easily seen by expressing $\hat{\pi}(x_{0:n-1}) = (1/n) \sum_{k=0}^{n-1} \sum_{i=1}^{K} \mathbb{1}(x_k = i) e_i$, where $\{e_1, \ldots, e_K\}$ is the canonical basis in $\mathbb{R}^K$.

Consider the stationary chain (P, π). An application of the theorem of Bobkov and Götze (one-sided version) combined with Theorem 3 gives

$$P_\pi(\|\hat{\pi}(X_{0:n-1}) - \pi\|_p \geq E_\pi[\|\hat{\pi}(X_{0:n-1}) - \pi\|_p] + \epsilon) \leq e^{-2^{1-2/p} n \epsilon^2 (1-r)^2}, \quad \forall \epsilon > 0$$

or equivalently, $\forall \epsilon > E_\pi[\|\hat{\pi}(X_{0:n-1}) - \pi\|_p]$,

$$P_\pi(\|\hat{\pi}(X_{0:n-1}) - \pi\|_p \geq \epsilon) \leq e^{-2^{1-2/p} n \left(\epsilon - E_\pi[\|\hat{\pi}(X_{0:n-1}) - \pi\|_p]\right)^2 (1-r)^2}.$$

Theorem 2 now implies that $\forall \epsilon > E_\pi[\|\hat{\pi}(X_{0:n-1}) - \pi\|_p]$,

$$P_\varrho(\|\hat{\pi}(X_{0:n-1}) - \pi\|_p \geq \epsilon) \leq \left\| \frac{\varrho}{\pi} \right\|_{2,\pi} e^{-2^{1-2/p} n \left(\epsilon - E_\pi[\|\hat{\pi}(X_{0:n-1}) - \pi\|_p]\right)^2 (1-r)^2}. \tag{16}$$

Finally, for any $\delta \in (0, 1)$ and any $\epsilon > E_\pi[\|\hat{\pi}(X_{0:n-1}) - \pi\|_p]$, $P_\varrho(\|\hat{\pi}(X_{0:n-1}) - \pi\|_p \geq \epsilon) \leq \delta$ for any n such that

$$n \geq \frac{2^{\frac{2}{p}} \log\left(\frac{\|\frac{\varrho}{\pi}\|_{2,\pi}}{\delta} \right)}{\left(\epsilon - E_\pi[\|\hat{\pi}(X_{0:n-1}) - \pi\|_p]\right)^2 (1-r)^2}. \tag{17}$$

Approach 2: A union bound approach

We may try to eliminate the problem of ϵ being *bounded away from zero* by using the observation that any ℓ_p-norm is separable in the corresponding coordinates. We have

$$P_\pi(\|\hat{\pi} - \pi\|_p \geq \epsilon) = P_\pi\left(\sum_{i=1}^{K} |\hat{\pi}_i - \pi_i|^p \geq \epsilon^p \right) \leq \sum_{i=1}^{K} P_\pi\left(|\hat{\pi}_i - E_\pi[\hat{\pi}_i]| \geq \frac{\epsilon}{\sqrt[p]{K}} \right),$$

where the union bound and the fact that $\hat{\pi}_i$ are unbiased estimators $\forall i$ have been used. In this case $f(x) = f(x_{0:n-1}) = \hat{\pi}_i(x_{0:n-1})$, therefore

$$|f(x) - f(\tilde{x})| \leq \frac{1}{n} \tilde{d}_{1,\mathbb{E}^n}(x_{0:n-1}, \tilde{x}_{0:n-1}), \quad \forall x_{0:n-1}, \tilde{x}_{0:n-1} \in \mathbb{E}^n.$$

By combining the theorem of Bobkov and Götze with Theorems 2 and 3 we obtain:

$$P_{\varrho}(\|\hat{\pi}(X_{0:n-1}) - \pi\|_p \geq \epsilon) \leq \sqrt{2K} \left\| \frac{\varrho}{\pi} \right\|_{2,\pi} e^{-K^{-2/p} n \epsilon^2 (1-r)^2}. \tag{18}$$

Finally, for any $\delta \in (0, 1)$ and any $\epsilon > 0$, $P_{\varrho}(\|\hat{\pi}(X_{0:n-1}) - \pi\|_p \geq \epsilon) \leq \delta$ for any n such that

$$n \geq \frac{K^{\frac{2}{p}} \log\left(\frac{\sqrt{2K} \| \frac{\varrho}{\pi} \|_{2,\pi}}{\delta} \right)}{\epsilon^2 (1-r)^2}. \tag{19}$$

Approach 3: Application of Theorem 1

For simplicity, we will work with the first bound in (2). Note that by working with both bounds in (2) we can only obtain an improvement of the derived sample complexity.

We observe that by the usual norm equivalence constants, the definition of τ_p in (3) and the hierarchy of ℓ_p-norms in $\mathbb{R}^k$ we have that $\tau_1 \leq \sqrt{k}$, $\tau_2 = 1$, $\tau_p \leq 1$ for any $p > 2$ and $\tau_{p_1} \leq \tau_{p_2}$ for $p_1 \geq p_2$. By Theorems 1–3 we obtain that $\forall \epsilon > 0$ and $\forall \varepsilon \in (0, 1]$,

$$P_{\varrho}(\|\hat{\pi}(X_{0:n-1}) - \pi\|_p \geq \epsilon) \leq \left\| \frac{\varrho}{\pi} \right\|_{2,\pi} \left(1 + \frac{2}{\varepsilon}\right)^{\frac{K}{2}} e^{-\frac{n\epsilon^2 (1-\varepsilon)^2 (1-r)^2}{2\tau_p^2}}. \tag{20}$$

Therefore, for any $\delta \in (0, 1)$, any $\epsilon > 0$ and any $\varepsilon \in (0, 1]$, $P_{\varrho}(\|\hat{\pi}(X_{0:n-1}) - \pi\|_p \geq \epsilon) \leq \delta$ for any n such that

$$n \geq \frac{2\tau_p^2}{\epsilon^2 (1-\varepsilon)^2 (1-r)^2} \left[\frac{K}{2} \log\left(1 + \frac{2}{\varepsilon}\right) + \log\left(\left\| \frac{\varrho}{\pi} \right\|_{2,\pi} \right) + \log\left(\frac{1}{\delta}\right) \right]. \tag{21}$$

Sample Complexity Comparisons

We first note that $\|\varrho/\pi\|_{2,\pi} \leq 1/\sqrt{\min_{i \in \mathbb{E}} \pi(i)}$ and often $\min_{i \in \mathbb{E}} \pi(i) \asymp 1/K^m$ for some $m \geq 1$. For a rough complexity comparison between (17) and (21) consider for simplicity the special case of i.i.d. random variables and $p = 1$. In this setting, we correspondingly work with $P_{\pi}(\cdot)$ only (only $P_{\pi}(\cdot)$ is meaningful). It turns out that $E_{\pi}[\|\hat{\pi} - \pi\|_1] \lesssim K/\sqrt{n}$. Then, (17) and (21) are orderwise the same, but without the problem of ϵ being bounded away from zero in (21). Further, (21) is better by a **logarithmic in** K **factor** over (19) for $p \in \{1, 2\}$. For chains such that $\min_{i \in \mathbb{E}} \pi(i) \asymp 1/e^K$, (21) is better by a K **factor** over (19) for $p \in \{1, 2\}$. We also note that depending on the geometry of $f(\mathbb{X})$, f and ν, τ_p may or may not have a favorable value for a particular p. More specifically, it might be possible that the last approach is orderwise better than the union bound approach for some choices of p, primarily for $p \in \{1, 2\}$, while it is worse for other values of p, depending on the particular problem at hand.

Acknowledgment

This work was supported by the ONR, USA N00014-19-1-2566.

References

Bobkov, S.G., Götze, F., 1999. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. J. Funct. Anal. 163 (1), 1–28.
Boucheron, S., Lugosi, G., Massart, P., 2013. Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press.
Brémaud, P., 2013. Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues, Vol. 31. Springer Science & Business Media.
Dembo, A., Zeitouni, O., 1998. Large Deviations Techniques and Applications. Springer-Verlag, New York.
Gromov, M., 2007. Metric Structures for Riemannian and Non-Riemannian Spaces. Springer Science & Business Media.
Hsu, D., Kakade, S., Zhang, T., 2012. A tail inequality for quadratic forms of subgaussian random vectors. Electron. Commun. Probab. 17 (52), 1–6.
Lattimore, T., Szepesvári, C., 2020. Bandit Algorithms. Cambridge University Press.
Levin, D.A., Peres, Y., 2017. Markov Chains and Mixing Times, Vol. 107. American Mathematical Soc.
Marton, K., 1986. A simple proof of the blowing-up lemma (corresp.). IEEE Trans. Inform. Theory 32 (3), 445–446.
Marton, K., 1996. Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration. Ann. Probab. 24 (2), 857–866.
McDiarmid, C., 1997. Centering sequences with bounded differences. Comb. Probab. Comput. 6 (1), 79–86.
Paulin, D., 2015. Concentration inequalities for Markov chains by Marton couplings and spectral methods. Electron. J. Probab. 20 (79), 1–32.
Raginsky, M., Sason, I., 2014. Concentration of measure inequalities in information theory, communications, and coding. pp. 1–180, arXiv:1212.4663.
Rezakhanlou, F., 2015. Lectures on the Large Deviation Principle. Lecture Notes, Math UC Berkeley.
Van Handel, R., 2016. Probability in High Dimension. APC 550 Lecture Notes, Princeton University.
Vershynin, R., 2018. High-Dimensional Probability: An Introduction with Applications in Data Science, Vol. 47. Cambridge University Press.
Villani, C., 2008. Optimal Transport: Old and New, Vol. 338. Springer Science & Business Media.
Wainwright, M.J., 2019. High-Dimensional Statistics: A Non-Asymptotic Viewpoint, Vol. 48. Cambridge University Press.