A Measure of Difference between Discrete Sample Sets

Debejyo Chakraborty[†*] and Narayan Kovvali[‡]
[†]Global Research & Development, General Motors Company, Warren, Michigan
[‡]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona

***Abstract*—The estimation of statistical distance between populations is a task of importance for many applications. Conventional methods often rely on the use of a maximum-likelihood (ML) estimator, usually due to its analytical and computational simplicity. However, the ML point estimate provides no information about the uncertainty in the parameters and distance estimated, which grows with lesser amounts of observed data. In this paper, a new measure is developed for statistical difference between finite sized sample sets of discrete observations. The measure is defined as the expected distance between probability mass functions (pmfs), with the expectation carried out over Dirichlet posteriors on the pmfs given the observed samples. In contrast to conventional ML estimates of distance, this approach by-design accounts for the uncertainty due to the finite size of the observation sets. In the limit of infinite number of observation samples, the expected distance simplifies to the ML estimate. For finite and small sized sample sets, the expected distance yields a more reliable measure of statistical difference.**

I. Introduction

The estimation of statistical distance between populations is a task of importance for many applications, for example, classification [1], dimensionality reduction [2], and region-of-interest based tracking [3]. Conventional methods often rely on the use of a maximum-likelihood (ML) estimator [4, 5], usually due to its analytical and computational simplicity. However, the ML point estimate provides little information about the uncertainty in the parameters and distance estimated, which is greater for lesser amounts of observed data. For data sets with very few observation samples, the ML point estimate may be completely unrepresentative of the true distance. Therefore, while the ML estimator may suffice in situations where plenty of data is available, its use is undesirable for many real problems which suffer from insufficient data because of the difficulty and cost associated with the collection and management of large amounts of data.

In this paper, we develop a new measure of statistical difference between finite sized sample sets of discrete observations. The measure is defined as the expected distance between probability mass functions (pmfs), with the expectation carried out over Dirichlet posteriors on the pmfs given the observed samples. In contrast to conventional ML estimates of distance, this approach by-design accounts for the uncertainty due to the finite size of the observation sets. In particular, by using the expected squared deviation of the distance, probabilistic bounds can be defined for the distance measure in order to quantify the statistical difference between the sample sets of finite size. In the limit of infinite number of observation samples, the expected distance simplifies to the ML estimate. For finite and small sized sample sets, the expected distance yields a more reliable measure of statistical difference.

The distance measure can be used in a variety of applications. For instance, the measure can be used to implement an effective and efficient convergence diagnostic [6–8] for (discrete) single-run Markov chain Monte Carlo (MCMC) [9] simulations with known or unknown target distribution. This measure can also be used in statistical hypothesis tests, detection theory [10], and for classification problems.

The remainder of this paper is organized as follows. In Section II, we discuss the new distance measure and provide closed-form analytical expressions for its computation. In Section III, results are presented comparing the analytical calculations with Monte Carlo integration, and some example applications of the distance measure are given. This is followed by conclusion in Section IV.

II. Theory

The proposed method considers the uncertainty in the underlying pmfs manifested by the finite size of the observed data sets, and leverages it in order to estimate and quantify the resulting uncertainty in the distance measure.

Consider two discrete data sets of size N, $\mathbb{Y}_N^1 = \{y_n^1\}_{n=1}^N \sim \mathbf{p}_1$ and $\mathbb{Y}_N^2 = \{y_n^2\}_{n=1}^N \sim \mathbf{p}_2$, distributed according to underlying M-state pmfs $\mathbf{p}_1$ and $\mathbf{p}_2$, respectively. Let $\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2)$ denote a statistical measure of difference between $\mathbf{p}_1$ and $\mathbf{p}_2$, also referred to as *statistical distance*. Some common examples of statistical distance measures include Kullback-Leibler (KL) divergence [5, 11], Bhattacharyya distance [12, 13], and total variation distance; appropriate choice of distance measure is made based on the application.

In the ML approach, pmf estimates $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$ are computed using the frequency of occurrence of the states in the observed data samples:

$$\hat{p}_l[i] = \frac{|\{n : y_n^l \in \text{state } i\}|}{N},\ i = 1, \ldots, M,\ l = 1, 2,$$

and these estimates are used to approximate the statistical distance as $\mathcal{D}(\hat{\mathbf{p}}_1||\hat{\mathbf{p}}_2)$.

In this paper, the expected distance measure is defined as

$$\overline{\mathcal{D}}(\mathbf{p}_1||\mathbf{p}_2) \triangleq \mathbf{E}\left[\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2)\right] = \iint \mathcal{D}(\mathbf{p}_1||\mathbf{p}_2)\, P\left(\mathbf{p}_1|\mathbb{Y}_N^1\right)\, P\left(\mathbf{p}_2|\mathbb{Y}_N^2\right)\, d\mathbf{p}_1\, d\mathbf{p}_2, \quad \text{(1a)}$$

*Point of contact. *Email*: debejyo@gmail.com
Web: http://www.debejyo.com

where $P\left(\mathbf{p}_1|\mathbb{Y}_N^1\right)$ and $P\left(\mathbf{p}_2|\mathbb{Y}_N^2\right)$ are the posterior distributions over $\mathbf{p}_1$ and $\mathbf{p}_2$ given the observation sets $\mathbb{Y}_N^1$ and $\mathbb{Y}_N^2$, respectively. The variance can be written as

$$\mathbf{Var}\left[\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2)\right] = \iint \left[\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2) - \overline{\mathcal{D}}(\mathbf{p}_1||\mathbf{p}_2)\right]^2 P\left(\mathbf{p}_1|\mathbb{Y}_N^1\right) P\left(\mathbf{p}_2|\mathbb{Y}_N^2\right)\, d\mathbf{p}_1\, d\mathbf{p}_2. \quad \text{(1b)}$$

In particular, the posteriors $P\left(\mathbf{p}_1|\mathbb{Y}_N^1\right) = \mathrm{Dir}(\boldsymbol{\alpha}_1)$ and $P\left(\mathbf{p}_2|\mathbb{Y}_N^2\right) = \mathrm{Dir}(\boldsymbol{\alpha}_2)$ are Dirichlet distributions with parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ determined by the frequency of occurance of the states in the observed data samples:

$$\alpha_l[i] = 1 + |\{n : y_n^l \in \text{state } i\}|,\ i = 1, \ldots, M,\ l = 1, 2.$$

Probabilistic bounds can be obtained for the distance measure using Chebyshev's inequality with the expected distance and its variance:

$$P(\left|\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2) - \overline{\mathcal{D}}(\mathbf{p}_1||\mathbf{p}_2)\right| \geq \epsilon) \leq \frac{\mathbf{Var}\left[\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2)\right]}{\epsilon^2}, \quad (2)$$

i.e., with probability at least γ, the statistical distance $\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2)$ lies within $\overline{\mathcal{D}}(\mathbf{p}_1||\mathbf{p}_2) \pm \sqrt{\frac{\mathbf{Var}[\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2)]}{1-\gamma}}$ (with 0 a trivial lower bound).

The quantities in Eq. (1) can be computed for general distance measures $\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2)$ by using Monte Carlo integration. In this work, we provide closed-form analytical expressions for the case when $\mathcal{D}(\mathbf{p}_1||\mathbf{p}_2) = \mathcal{D}_{\mathrm{KL}}(\mathbf{p}_1||\mathbf{p}_2)$ is the KL divergence:

$$\mathcal{D}_{\mathrm{KL}}(\mathbf{p}_1||\mathbf{p}_2) \triangleq \sum_{i=1}^{M} p_1[i]\, \log \frac{p_1[i]}{p_2[i]}. \quad (3)$$

The specific expressions for the KL distance mean and variance are as follows.

$$\overline{\mathcal{D}}_{\mathrm{KL}}(\mathbf{p}_1||\mathbf{p}_2) = \sum_{i=1}^{M} \frac{\alpha_1[i]}{\alpha_1^0}\Big[\psi(\alpha_1[i]+1) - \psi(\alpha_1^0+1) - \psi(\alpha_2[i]) + \psi(\alpha_2^0)\Big], \quad \text{(4a)}$$

$$\mathbf{Var}\left[\mathcal{D}_{\mathrm{KL}}(\mathbf{p}_1||\mathbf{p}_2)\right] = \mathbf{E}\left[\mathcal{D}_{\mathrm{KL}}(\mathbf{p}_1||\mathbf{p}_2)^2\right] - \overline{\mathcal{D}}_{\mathrm{KL}}(\mathbf{p}_1||\mathbf{p}_2)^2, \quad \text{(4b)}$$

with the second moment

$$\mathbf{E}\left[\mathcal{D}_{\mathrm{KL}}(\mathbf{p}_1||\mathbf{p}_2)^2\right] = \sum_{i=1}^{M}\sum_{j=1}^{M}\Big[\mathfrak{I}_1 - \mathfrak{I}_2\mathfrak{I}_3 - \mathfrak{I}_4\mathfrak{I}_5 + \mathfrak{I}_6\mathfrak{I}_7\Big],$$

where $\alpha_l^0 = \sum_{i=1}^{M} \alpha_l[i]$ for $l = 1, 2$, and $\mathfrak{I}_1$ through $\mathfrak{I}_7$ are given by

$$\mathfrak{I}_1 = \frac{\alpha_1[i]\,(\alpha_1[j]+\delta_{i,j})}{\alpha_1^0\,(\alpha_1^0+1)}\Big[\psi_1(\alpha_1[i]+2)\,\delta_{i,j} - \psi_1(\alpha_1^0+2) + \big(\psi(\alpha_1[i]+1+\delta_{i,j}) - \psi(\alpha_1^0+2)\big) \cdot \big(\psi(\alpha_1[j]+1+\delta_{i,j}) - \psi(\alpha_1^0+2)\big)\Big],$$

$$\mathfrak{I}_2 = \frac{\alpha_1[i]\,(\alpha_1[j]+\delta_{i,j})}{\alpha_1^0\,(\alpha_1^0+1)}\Big[\psi(\alpha_1[j]+1+\delta_{i,j}) - \psi(\alpha_1^0+2)\Big],$$

$$\mathfrak{I}_3 = \psi(\alpha_2[i]) - \psi(\alpha_2^0),$$

$$\mathfrak{I}_4 = \frac{\alpha_1[i]\,(\alpha_1[j]+\delta_{i,j})}{\alpha_1^0\,(\alpha_1^0+1)}\Big[\psi(\alpha_1[i]+1+\delta_{i,j}) - \psi(\alpha_1^0+2)\Big],$$

$$\mathfrak{I}_5 = \psi(\alpha_2[j]) - \psi(\alpha_2^0),$$

$$\mathfrak{I}_6 = \frac{\alpha_1[i]\,(\alpha_1[j]+\delta_{i,j})}{\alpha_1^0\,(\alpha_1^0+1)},$$

$$\mathfrak{I}_7 = \psi_1(\alpha_2[i])\,\delta_{i,j} - \psi_1(\alpha_2^0) + \big(\psi(\alpha_2[i]) - \psi(\alpha_2^0)\big)\big(\psi(\alpha_2[j]) - \psi(\alpha_2^0)\big).$$

Here $\psi(\cdot)$ is the digamma function, $\psi_1(\cdot)$ is the trigamma function, and δ denotes the Kronecker delta function.

The method requires $O(M^2)$ computational effort for evaluating the statistical KL distance. The result is a reliable and efficiently-computed KL distance measure which can be deployed for real-time applications.

Note that in scenarios where one of the pmf is known (i.e., when computing the statistical distance of a given discrete sample set to a known pmf), the corresponding integral in (1) vanishes. For example, if the pmf $\mathbf{p}_2 = \mathbf{p}_*$ is known, then

$$\overline{\mathcal{D}}(\mathbf{p}_1||\mathbf{p}_*) = \int \mathcal{D}(\mathbf{p}_1||\mathbf{p}_*)\, P\left(\mathbf{p}_1|\mathbb{Y}_N^1\right)\, d\mathbf{p}_1, \quad \text{(5a)}$$

$$\mathbf{Var}\left[\mathcal{D}(\mathbf{p}_1||\mathbf{p}_*)\right] = \int \left[\mathcal{D}(\mathbf{p}_1||\mathbf{p}_*) - \overline{\mathcal{D}}(\mathbf{p}_1||\mathbf{p}_*)\right]^2 P\left(\mathbf{p}_1|\mathbb{Y}_N^1\right)\, d\mathbf{p}_1. \quad \text{(5b)}$$

As before, these quantities can also be evaluated analytically for the KL divergence.

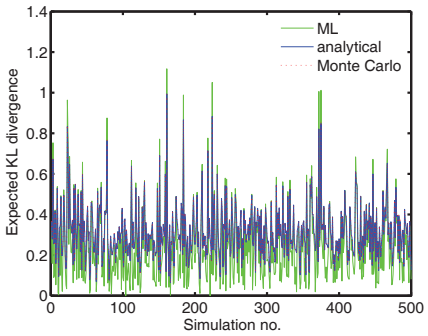
III. Simulation Results

We now present simulation results comparing the analytical calculations of Section II with Monte Carlo integration, and show example applications of the expected KL divergence measure for the assessment of convergence of a discrete single-run Markov chain and for testing uniformity.

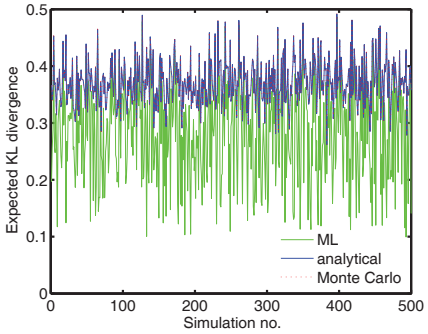
A. Validation of Derived Results

We compare estimates of the expected KL divergence measure using three methods: ML approach, analytical expected distance, and Monte Carlo integration. Figure 1 shows the results for estimating the expected KL divergence between two sets of synthetically generated data samples. 500 simulations were performed, with the following combinations of number of states and (average) number of data samples: (a) $M = 5$ and $N = 35$, (b) $M = 100$ and $N = 700$, (c) $M = 5$ and $N = 150$, and (d) $M = 100$ and $N = 3000$. In each simulation, about 500 to 1000 independent and identically distributed (i.i.d.) Dirichlet random samples were used for the Monte Carlo integration.

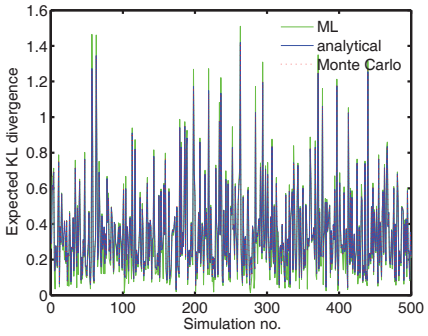
The plots show that the KL divergence estimate from the ML approach is close to the expected distance only in case (c) where the number of states is small and the number of data samples per state is large. The difference is particularly large in the case (b) where the number of states is large and the number of data samples per state is small. In all cases, the analytical expected KL divergence agrees very well with



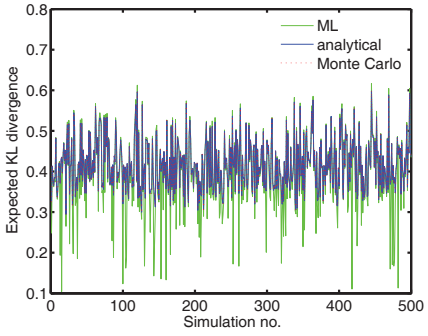
(a) $M = 5$ and $N = 35$



(b) $M = 100$ and $N = 700$



(c) $M = 5$ and $N = 150$



(d) $M = 100$ and $N = 3000$

Fig. 1. Estimation of expected KL divergence between two sets of synthetically generated data samples for various combinations of number of states M and (average) number of data samples N.

the Monte Carlo integration results, while being much more efficient to calculate than the Monte Carlo integration. The analytical and Monte Carlo integration results were observed to agree well for the calculation of variance also (not shown here).

B. Markov Chain Convergence Diagnosis

Consider a simple first-order time-homogeneous $M = 4$-state Markov chain with transition matrix A and stationary distribution π. Let A_1 and A_2 be two transition matrices, given by

$$A_1 = \begin{bmatrix} 0.9 & 0.1 & 0.0 & 0.0 \\ 0.0 & 0.9 & 0.0 & 0.1 \\ 0.1 & 0.0 & 0.7 & 0.2 \\ 0.1 & 0.0 & 0.0 & 0.9 \end{bmatrix},$$

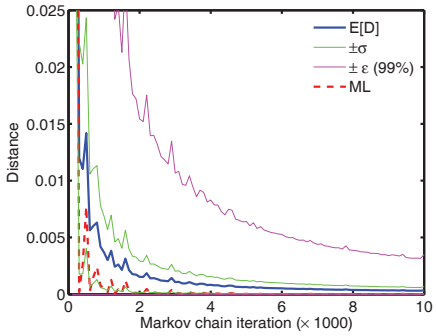
$$A_2 = \begin{bmatrix} 0.0 & 0.5 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.5 & 0.0 & 0.0 & 0.5 \\ 0.5 & 0.5 & 0.0 & 0.0 \end{bmatrix}.$$

The matrix A_1 has second largest eigenvalue magnitude 0.8544 and the matrix A_2 has second largest eigenvalue magnitude 0.7071. Therefore, the Markov chain with A_1 as transition matrix converges slower than that with A_2 as transition matrix. For each of the two Markov chains, a realization of length 10,000 samples was generated. Samples were collected into overlapping batches of growing size, in steps of 100 samples. That is, the first batch contained the first 100 samples, the second batch contained the first 200 samples (including the 100 samples from the first batch), and so on. The statistical KL divergence between consecutive batches was then computed and monitored to assess convergence.

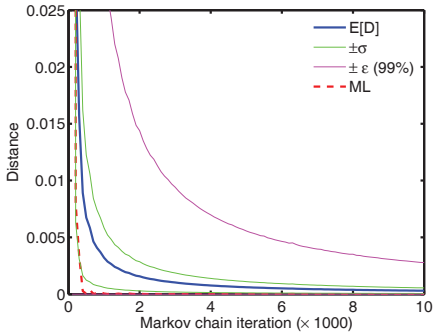
Figure 2 shows plots of the expected KL divergence between consecutive batches, 99% confidence intervals, $\pm 1\sigma$ deviation from the expected distance, and the corresponding ML estimates of distance for the two Markov chains as a function of the number of samples. Table I shows the expected KL divergence along with 99% confidence intervals and corresponding ML estimates, computed for the two chains at sample index 10,000 (100th batch). It can be seen that, for the same number of iterations, the faster converging Markov chain with transition matrix A_2 achieves lower expected KL divergence between consecutive batches than the chain with transition matrix A_1. Note that the comparison is less meaningful during the inital phase of the chains because of the larger variance, but as the chains progress the variance decreases (due to increasing batch size). In fact, the variance decreases at a rate proportional to $1/N$, where N is the number of samples. As the expected KL divergence between consecutive batches approach zero, it can be inferred that the batch samples are being drawn from similar distributions, indicating Markov chain convergence.

C. Test of Uniformity

The KL divergence can be utilized in statistical tests. For example, one can use the KL divergence to estimate how uniformly distributed a sample set is. This is an example of



(a) Transition matrix A_1 (second largest eigenvalue magnitude = 0.8544)



(b) Transition matrix A_2 (second largest eigenvalue magnitude = 0.7071)

Fig. 2. Expected KL divergence between consecutive Markov chain batches, 99% confidence intervals, and corresponding ML estimates for the two Markov chains as a function of the number of samples.

	Tr. matrix A_1	Tr. matrix A_2
$\overline{\mathcal{D}}_{\mathrm{KL}}$	3.0590×10^{-4}	3.0165×10^{-4}
$\mathbf{Var}\left[\mathcal{D}_{\mathrm{KL}}\right]$	7.7938×10^{-8}	6.0663×10^{-8}
$\mathcal{D}_{\mathrm{KL}}$ lower bound	0	0
$\mathcal{D}_{\mathrm{KL}}$ upper bound	3.0911×10^{-3}	2.7646×10^{-3}
$\mathcal{D}_{\mathrm{KL}}$ ML estimate	5.0200×10^{-6}	2.5827×10^{-7}

TABLE I
EXPECTED KL DIVERGENCE, 99% CONFIDENCE INTERVALS, AND CORRESPONDING ML ESTIMATES, COMPUTED FOR THE TWO MARKOV CHAINS AT SAMPLE INDEX 10,000 (100TH BATCH).

a problem where one of the pmfs, say $\mathbf{p}_2$, is known and is uniform. Then, using (5), one can calculate the expected KL divergence and the variance. Alternatively, the framework (1)-(4) may be used to provide an approximation, by setting $\alpha_2[i] = N$, $i = 1, \ldots, M$ where N is large. This is equivalent to imposing that $\mathbf{p}_2$ is a random pmf with Dirichlet distribution sharply concentrated around the uniform pmf.

We consider two $M = 4$-state Markov chains: one with uniform target distribution and state transition matrix A_2, and the other with non-uniform target distribution and state transition matrix given by

$$A_3 = \begin{bmatrix} 0.3 & 0.2 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.3 & 0.2 \end{bmatrix}.$$

For each of the two Markov chains, a realization of length 100,000 samples was generated. As before, the samples were collected into overlapping batches of growing size, in steps of 100 samples. The statistical KL divergence of these batches was then computed with respect to a uniform pmf.

Figure 3(a) shows the statistical distance for the Markov chain using transition matrix A_2. It can be seen that the expected distance approaches zero, confirming that the target distribution for this Markov chain is indeed uniform. An empirical estimate of the stationary distribution for the Markov chain using transition matrix A_2 is shown in Figure 3(b). Figure 3(c) shows the statistical distance for the Markov chain using transition matrix A_3. In this case, the distance converges to a non-zero value, which is the expected KL divergence between the non-uniform target distribution of this Markov chain and the uniform pmf. An empirical estimate of the stationary distribution for the Markov chain using transition matrix A_3 is shown in Figure 3(d).

IV. DISCUSSION

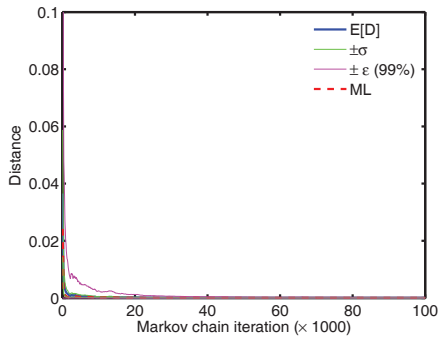
In this paper, a new measure has been described for the statistical difference between finite sized sample sets of discrete observations. The statistical distance measure is defined as the expected distance between underlying pmfs, which are Dirichlet distributed in light of the observed data samples. This approach by-design accounts for the uncertainty due to the finite size of the observation data sets, and yields the expected statistical distance along with probabilistic bounds. When the number of available data samples is small, the new statistical distance measure is more representative and reliable as compared to conventional ML estimates.

The expected distance measure and its bounds can be computed for general distance measures by using Monte Carlo integration. In particular, for the KL divergence measure, closed-form analytical expressions were derived for the expected distance and its bounds, which can be evaluated with $O(M^2)$ computational complexity (where M is the number of discrete states).

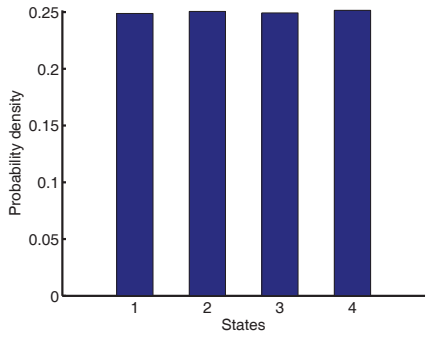
The utility of the expected statistical distance measure is demonstrated through simulation results for diagnosing the convergence of a discrete single-run Markov chain and for testing uniformity. Other potential applications include statistical hypothesis testing, detection and classification tasks, sequential Monte Carlo methods, etc. The approach can also be used to generalize and compute an expected discrepancy measure [14] with an associated confidence interval, which would benefit applications such as [1].

REFERENCES

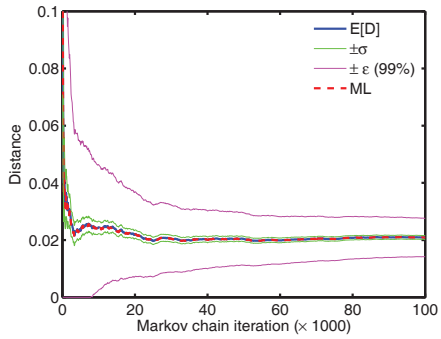
[1] D. Chakraborty, N. Kovvali, A. Papandreou-Suppappola, and A. Chattopadhyay, "Active learning data selection for



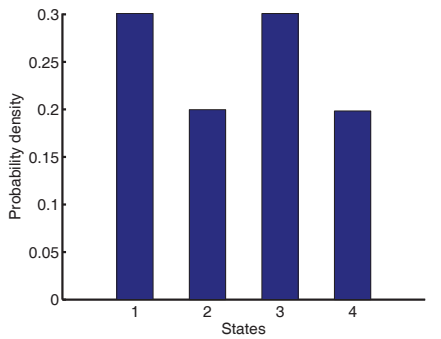
(a) Distance from uniform pmf for Markov chain with transition matrix A_2



(b) Empirically estimated target distribution



(c) Distance from uniform pmf for Markov chain with transition matrix A_3



(d) Empirically estimated target distribution

Fig. 3. Expected KL distance from a uniform pmf for Markov chains with uniform and non-uniform target distributions.

adaptive online structural damage estimation," in *Proc. of SPIE*, vol. 7649, 2010, p. 764915.

[2] A. Bhattacharya, P. Kar, and M. Pal, "On low distortion embeddings of statistical distance measures into low dimensional spaces." in *DEXA'09*, 2009, pp. 164–172.

[3] S. Boltz, E. Debreuve, and M. Barlaud, "High-dimensional statistical distance for region-of-interest tracking: Application to combining a soft geometric constraint with radiometry," in *in IEEE International Conference on Computer Vision and Pattern Recognition*, 2007, p. 07.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley Interscience, 2001.

[5] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[6] M. K. Cowles and B. P. Carlin, "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 883–904, Jun 1996.

[7] A. Gelman and D. B. Rubin, "A single series from the Gibbs sampler provides a false sense of security," *Bayesian Statistics*, vol. 4, pp. 625 – 631, 1992.

[8] ——, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, no. 4, pp. 457 – 511, 1992.

[9] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1996.

[10] H. L. V. Trees, *Detection, Estimation, and Modulation Theory, Part I*. Wiley Interscience, 2001.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[12] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[13] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967.

[14] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, L. Bers, P. Hilton, and H. Hochstadt, Eds. John Wiley & Sons Inc., 1974.