Supplementary Material

Guillaume Basse, Natesh Pillai and Aaron Smith

May, 2016

1 Proofs

*Proof of Theorem **5.2**.* Let $\Omega = \sqcup_{i=1}^{n} \Omega_i$ be the (random) partition used by Algorithm **1** in stage ℓ. Since the densities of $Q(x, \cdot)$ and $\pi(\cdot)$ are uniformly bounded away from 0 by a constant $c > 0$, so are the densities $\{\pi_i(\cdot)\}_{i=1}^{n}$. This immediately implies that the kernels $\{K_i\}_{i=1}^{n}$ used in stage ℓ are uniformly ergodic. The result then follows from the strong law of large numbers for uniformly ergodic Markov chains, combined with the consistency of bridge sampling. □

*Proof of Theorem **5.3**.* Fix $k \in \mathbb{N}$. We note that if the second line of the submethod **DoSpectralClustering** were replaced with

Sample N i.i.d. points from the mixture distribution $\frac{1}{n}\sum_{i=1}^{n} \pi_i$,

then Equation *(23)* would follow immediately from Theorem 16 of Ben-David et al. (2006). Thus, to prove Equation *(23)* in our setting, it is sufficient to prove that the distribution of the subsample obtained in **DoSpectralClustering**$(X, n, N(k), Q, \pi)$ converges in total variation distance to an i.i.d. sample as k goes to infinity. To do this, we need some additional notation.

We fix an ordering of the elements of $X = \{X_t^{(i)}\}_{0 \le t \le T(k),\, 1 \le i \le n}$ as follows: $X = \{X[1], \ldots, X[nT(k)]\} = \{X_1^{(1)}, X_2^{(1)}, \ldots, X_{T(k)}^{(1)}, X_1^{(2)}, \ldots, X_{T(k)}^{(n)}\}$. Let $1 \le s_1 < \ldots < s_{N(k)} \le nT(k)$ be the indices of the subsample selected in the second line of **DoSpectralClustering** when **DoSpectralClustering**$(X, n, N(k), Q, \pi)$ is called.

Next, let $X'_1, \ldots, X'_{N(k)}$ be an i.i.d. sequences drawn from the mixture distribution $\frac{1}{n}\sum_{i=1}^{n} \pi_i$. By the remark in the first paragraph of the proof, to prove Equation *(23)* it is sufficient to prove that:

$$\lim_{k \to \infty} \|\mathcal{L}(\{X[s_i]\}_{i=1}^{N(k)}) - \mathcal{L}(\{X'_i\}_{i=1}^{N(k)})\|_{\mathrm{TV}} = 0 \tag{1}$$

Next, define

$$\Delta(k) = \min\Big(\min_{1 \le i < N(k)} (s_{i+1} - s_i), \min_{1 \le i \le N(k),\, 0 \le j \le n} (|s_i - jT(k)|)\Big).$$

By Equation *(22)*, we have

$$\lim_{k\to\infty} \mathbb{P}[\Delta(k) < N(k)^\gamma] \leq \lim_{k\to\infty} \left(\left(1 - \frac{3N(k)^{1+\gamma}}{T(k)}\right)^{N(k)} + \left(\frac{3nN(k)}{T(k)}\right)\right) \tag{2}$$
$$= \lim_{k\to\infty} O\left(\frac{N(k)^{2+\gamma}}{T(k)} + \frac{N(k)}{T(k)}\right) = 0.$$

Next, define $D = \min_{1\leq j\leq n} \lambda(\Omega_j)$, where λ is the Lebesgue measure on $\mathbb{R}^d$. Conditioning on the indices $\{s_i\}_{i=1}^{N(k)}$, we have by Equation *(21)* and Equation (2) that

$$\lim_{k\to\infty} \|\mathcal{L}(\{X_{s_i}(k)\}_{i=1}^{N(k)}) - \mathcal{L}(\{X_i'\}_{i=1}^{N(k)})\|_{\text{TV}} \tag{3}$$
$$\leq \lim_{k\to\infty} (N(k)e^{-\lfloor \frac{N(k)^\gamma}{Dc} \rfloor} + \mathbb{P}[\Delta < N(k)^\gamma])$$
$$= 0.$$

We conclude that condition (1) holds, and thus that Equation *(23)* holds. This completes the proof. □

2 Simulations

2.1 Details of Example 1

The target $\pi(x)$ is a mixture of bivariate normals with the weights $w = (0.02, 0.2, 0.2, 0.58)$, means $\mu_1 = (3, 3), \mu_2 = (7, -3), \mu_3 = (2, 7), \mu_4 = (-5, 0)$, and covariance matrices $\Sigma_1 = [1, 0.2; 0.2, 1], \Sigma_2 = [2, -0.5; -0.5, 0.5], \Sigma_3 = [1.3, 0.3; 0.3, 0.4], \Sigma_4 = [1, 1; 1, 2.5]$. The proposal distribution we use for our method has the following density: $Q(x, \cdot) = MVN(x, 0.8I_2)$ where MVN denotes the multivariate normal distribution, and I_2 is the two by two identity matrix. For parallel tempering, on top of the original target $\pi(X)$ and the corresponding proposal just mentioned, we use the 3 tempered targets $\pi^{(2)}(x) = \pi(x)^{0.7}$, $\pi^{(3)}(x) = \pi(x)^{0.4}$, $\pi^{(4)}(x) = \pi(x)^{0.25}$ and their associated proposals $Q^{(2)}(x, \cdot) = MVN(x, I_2)$, $Q^{(3)}(x, \cdot) = MVN(x, 2I_2)$, $Q^{(4)}(x, \cdot) = MVN(x, 3I_2)$.

2.2 Details of Example 2

The proposal we use for this example is $Q(x, \cdot) = MVN(x, 0.015I_2)$.

3 Extensions

3.1 Estimating ratios of normalizing constants

In our article, we use a vanilla version of bridge sampling (Meng and Wong (1996)) to estimate the partition weights. Meng and Schilling (2002) proposed an extension of bridge sampling that essentially transforms the densities before applying the bridge, thus increasing the "overlap" between the two distributions. Gelman and Meng (1998) extends bridge sampling in a different direction by considering multiple bridges (eventually, an infinity of bridges).

An alternative family of strategies for estimating ratios of normalizing constants is Importance Sampling, as well as extensions such as Annealed Importance Sampling (Neal (2001)) and Linked Importance Sampling (Neal (2005)).

3.2 Scaling Spectral Clustering in High Dimensions

Like most clustering algorithms, spectral clustering experiences some difficulties in high dimension. Niu et al. (2011) identifies the main issue as being a lack of robustness to "noisy dimensions" and proposes a method to reduce the dimension of the data before clustering. This method can be plugged into our approach. More generally, any method that improves on spectral clustering can be integrated to our approach.

3.3 Extending clustering to a partition

In the paper, we use the centers $\{C_1, \ldots, C_n\}$ found by the **kmeans** algorithm in the eigenspace to define the partition. An alternative is to use the observed points, as follow:

$$\Omega_i = \{x \in \Omega \,:\, \sigma(\operatorname*{argmin}_{Z \in \{Z_1, \ldots, Z_N\}} \|Z - Z(x)\|) = i\}. \tag{4}$$

with the same notation as in the paper.

3.4 More on the heuristics

One may wonder how to choose the number of iterations ℓ. In practice, we want to run as many iterations as possible, so the real question of interest concerns the relationship between $T(k)$ and $N(k)$ for which we provide a heuristic in Theorem 5.3.

References

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

Xiao-Li Meng and Stephen Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002.

Xiao-li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Radford M Neal. Estimating ratios of normalizing constants using linked importance sampling. *arXiv preprint math/0511216*, 2005.

Donglin Niu, Jennifer G Dy, and Michael I Jordan. Dimensionality reduction for spectral clustering. In *International Conference on Artificial Intelligence and Statistics*, pages 552–560, 2011.