Proceedings of Machine Learning Research vol 99:1–43, 2019 32nd Annual Conference on Learning Theory

The Relative Complexity of Maximum Likelihood Estimation, MAP Estimation, and Sampling

Christopher Tosh C.TOSH@COLUMBIA.EDU
Columbia University

Sanjoy Dasgupta DASGUPTA@CS.UCSD.EDU
University of California, San Diego

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We prove that, for a broad range of problems, maximum-a-posteriori (MAP) estimation and approximate sampling of the posterior are at least as computationally difficult as maximum-likelihood (ML) estimation. By way of illustration, we show how hardness results for ML estimation of mixtures of Gaussians and topic models carry over to MAP estimation and approximate sampling under commonly used priors.

Keywords: Hardness, reductions, sampling

1. Introduction

When learning a probabilistic model, there are three computational tasks that commonly arise: maximum likelihood (ML) estimation of the model given data, maximum a posteriori (MAP) estimation when a prior distribution over models is specified, and (approximate) sampling of the posterior distribution over models. We are interested in the relative *computational* complexity of these three tasks: what does the hardness—or tractability—of one imply about the others?

At a high level, MAP estimation is rather like ML, with the added complication of the prior; and the two are known to converge to the same limit with infinite data, under certain conditions. Thus one would intuitively expect the MAP problem to be at least as hard as the ML problem.

The situation of approximately sampling is not as immediately clear. Sampling is known to be as hard as optimization for various statistical physics models with a "temperature" parameter: this temperature can be adjusted so that a sampler is essentially forced to return optimal or near-optimal solutions. In the usual setting of probabilistic learning, however, there is no such temperature knob. Nonetheless, there are other ways to produce a similar effect, and thus one would again expect, intuitively, that approximate sampling is no easier than ML estimation.

In this work, we make these intuitions precise. Considering probabilistic models in broad generality, we give simple conditions under which approximate MAP estimation and approximate posterior sampling can be shown to be at least as hard as approximate ML estimation. A key challenge here is formalizing issues of numerical precision.

We then illustrate these general reductions in two cases of interest. Starting from recent hardness results for ML estimation of Gaussian mixture models (Tosh and Dasgupta, 2018) and topic models (Arora et al., 2012), we show how in both settings, the hardness extends also to MAP estimation and approximate sampling.

1.1. Numerical precision

When discussing standard combinatorial optimization problems such as set cover or maximum cut, the first step is to consider the *exact* version of the problem and, when that proves intractable, to consider *approximate* solutions. For such problems, the exact solution lies in a discrete space and is of polynomial size, but is difficult to find.

By contrast, many of the problems that we have in mind—such as estimation of Gaussian mixture models or of topic distributions—take solutions in continuous spaces. The exact optimal solutions may therefore not be compactly representable. The following lemma, whose proof appears in the appendix, illustrates how this can happen even for extremely simple models.

Lemma 1 *When fitting a mixture model* $\frac{1}{2}\mathcal{N}(-\mu, 1) + \frac{1}{2}\mathcal{N}(\mu, 1)$ *to the data set of three points* $\{-2, 0, 2\}$*, the maximum-likelihood choice of* μ *is irrational.*

Thus, exact solutions are ruled out from the very beginning, and we are forced to restrict ourselves to approximate versions of ML and MAP estimation. In particular, we need to characterize the quality of polynomial-sized solutions.

The case of sampling is even more challenging, since we cannot hope to sample exactly from continuous domains. Much of the work on Markov chain mixing has focused on the discrete setting, and uses *total variation distance* to measure the difference between the target distribution μ, over a space Θ, and the distribution ν from which samples are actually drawn:

$$d_{TV}(\mu, \nu) = \sup_{\text{measurable } A \subset \Theta} |\mu(A) - \nu(A)|.$$

In cases where μ has continuous support, ν must still be discrete because samples must have bounded bit-length, and thus this distance will be identically 1.

To overcome this, we consider two notions of distribution approximation. The first is the Wasserstein distance, a metric over distributions that has become an increasingly popular measure of convergence in the optimization and sampling literature (Raginsky et al., 2017; Cheng et al., 2018). In the appendix, we also introduce a generalization of total variation distance that takes the supremum over a subfamily of measurable sets that captures Θ *at a certain granularity*. We show how to construct suitable families from ϵ-covers of Θ; our construction may be useful in other contexts.

1.2. Overview of results

In Section 3, we show that under conditions on the prior, there is a generic polynomial-time reduction from ML estimation to MAP estimation.

In Section 4, we define a notion of approximate posterior sampling that makes sense in both discrete and continuous domains and we then give a generic reduction from ML estimation to this problem, again under conditions on the prior.

Sections 3 and 4 extend the hardness of ML estimation problems to their Bayesian counterparts provided that the prior meets certain mild conditions. In these cases, we cannot hope for efficient MAP estimation algorithms or rapidly mixing Markov chains unless the data is specially constrained or $NP = RP$.

Throughout our exposition, topic modeling serves as a running example. In particular, we extend a hardness result of Arora et al. (2012) for ML estimation of topic models to the corresponding

Bayesian estimation problems. In Section 5, we do this also for the problem of estimating mixtures of Gaussians.

1.3. Methodology

Our goal is to reduce arbitrary maximum-likelihood (ML) estimation problems to their Bayesian counterparts. In the absence of special knowledge about the specific model under consideration, we opt for the simplest type of reduction: duplication.

Let $\mathcal{P} = \{p(\cdot \mid \theta) : \theta \in \Theta\}$ be a family of parameterized probability densities and let a q_0 be a prior density over Θ. Consider a data sequence $X = (x_1, \ldots, x_n)$ and suppose that we replicate this sequence k times, i.e., we make k copies $X^{(i)} = (x_1, \ldots, x_n)$. The posterior distribution, given $X^{(1)}, \ldots, X^{(k)}$, is of the form

$$\frac{1}{Z} q(\theta) p(X^{(1)}, \ldots, X^{(k)} \mid \theta) \;=\; \frac{1}{Z} q(\theta) \left(p(x_1, \ldots, x_n \mid \theta)\right)^k$$

where Z is the normalizing constant to make the density integrate to one. By simply replicating the data, we get an exponential increase on the weight of the likelihood function over the prior distribution. Thus, our general strategy will be to replicate the data until the posterior distribution is suitably concentrated around the maximum likelihood estimate.

The success of this approach ultimately hinges on the relationship of the prior distribution and the maximum likelihood estimate. If the prior distribution has very low density, say double-exponentially small, on parameters that are close to the maximum likelihood estimate, then to get large enough posterior weight on these parameters we need to duplicate the data a very large number of times, possibly more than polynomial. On the other hand, many prior distributions, especially those over unbounded parameter spaces, do put very small weight on *some* parameters. Thus the data duplication technique can fail on instances whose maximum likelihood estimate lies in some region of small prior density. How do we get around this?

The key observation is that many hardness reductions to ML estimation problems do not produce arbitrary instances. Indeed, these reductions often create instances with a special structure. And in many of these cases, the maximum likelihood solutions to these instances are themselves well-structured and, as a consequence, often lie close to parameter regions with non-negligible weight under many commonly-considered prior distributions. The upshot is that if we exploit the structure of hard instances, then we can often avoid the only obstacle to our duplication technique.

1.4. Related work, including connection with statistical literature

Of the computational tasks discussed in this work, ML estimation has seen the lion's share of hardness results (Chor and Tuller, 2006; Guruswami and Vardy, 2005). This is possibly because ML estimation does not have the additional complication of a prior distribution and can be easier to work with than MAP estimation and sampling.

In the computational literature, several algorithmic connections have been made between sampling and optimization. Kalai and Vempala (2006) showed that simulated annealing, a technique that involves approximately sampling from a sequence of distributions, can be used for certain convex optimization problems. Bubeck et al. (2015) used Langevin dynamics, a technique in which Gaussian noise is added to each step of gradient descent, to sample efficiently from log-concave distributions. What these, and other, works demonstrate is that certain optimization algorithms can

be turned into sampling algorithms, and vice versa. What is lacking, however, is a generic reduction between these two tasks.

In the statistical literature, there are many results to the effect that, under suitable conditions, for data sampled from some model in Θ, the maximum likelihood estimate in Θ asymptotically converges to this same model, and the posterior distribution asymptotically concentrates around it. Examples include the classical work of Le Cam (1953). Our hardness results are in a different setting—the data are arbitrary—but interestingly, require similar conditions on the prior. This is because our duplication technique gives the problem a statistical aspect: the replicated data look rather like multiple draws from an underlying distribution supported on the initial data points.

It is worth pointing out that the existence of efficient reductions from ML estimation to MAP estimation and posterior sampling, although intuitive, is by no means a foregone conclusion. There are various conditions that need to hold, and this work provides a blueprint that boils these reductions down to a few checks. Moreover, we illustrate how to apply this blueprint on two canonical examples, demonstrating hardness of MAP estimation and approximate sampling for topic modeling and mixtures of Gaussians for the first time.

Finally, we mention a closely related task to our estimation setting, but one which is not considered in the present work: using a fitted model to infer latent variables in new data. Notably, Sontag and Roy (2011) demonstrated that MAP estimation and posterior sampling for inference in LDA topic models are NP-hard. This result is particularly interesting as the corresponding maximum likelihood estimation problem is trivial.

2. Preliminaries and definitions

Let $\mathcal{X}$ be any data space. A *parameterized probability model* on $\mathcal{X}$ is a pair (p, Θ), where $p(\cdot \,|\, \theta)$ is a probability density over $\mathcal{X}$ for any $\theta \in \Theta$. We will be working with i.i.d. probability models, where for any sequence $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and any $\theta \in \Theta$,

$$p(X \,|\, \theta) = p(x_1, \ldots, x_n \,|\, \theta) = p(x_1 \,|\, \theta) \cdots p(x_n \,|\, \theta).$$

If we couple this probability model with a prior probability measure ν_0 over Θ, then the resulting triple (ν_0, p, Θ) is a *Bayesian parameterized probability model*. Let q_0 be a probability density corresponding to measure ν_0. The posterior density after observing X is then written as $q_X(\theta) \propto q_0(\theta) p(X \,|\, \theta)$ and we denote the corresponding measure as ν_X.

This notation conceals problem size: in reality, each input instance has some dimension m (the vocabulary size for documents, for instance) and requires parameters of corresponding dimensionality, in some $\Theta_m \subset \Theta$. We will suppress this dependence except where needed.

2.1. Maximum likelihood estimation

We formally define the maximum likelihood estimation problem as follows.

> ML ESTIMATION: MLE-(p, Θ)
> *Input*: A sequence of points $X \in \mathcal{X}^n$ and an accuracy parameter b in unary.
> *Output*: A parameter $\theta \in \Theta$ satisfying $\log p(X \,|\, \theta) \geq \sup_{\theta' \in \Theta} \log p(X \,|\, \theta') - 1/b$.

It might also be reasonable to ask for precision $1/2^b$. We adopt this particular formulation because it yields stronger hardness results.

2.2. Topic modeling

We will use the problem of topic modeling as a running example. We follow the model of Arora et al. (2012) where there is an unknown $V \times K$ topic matrix Ψ such that each column $\Psi^{(i)}$ is a distribution over a dictionary $[V]$, and there is a collection of D unknown, stochastically-generated distributions $\theta^{(1)}, \ldots, \theta^{(D)}$ over the topics $[K]$. The standard choice of prior on $\theta^{(1)}, \ldots, \theta^{(D)}$ is a symmetric Dirichlet(α) distribution. The generative process for a document d with words $w_1^{(d)}, w_2^{(d)}, \ldots$ is

$$\begin{aligned} \theta^{(d)} &\sim \text{Dirichlet}(\alpha) \\ z_i \mid \theta^{(d)} &\sim \text{Categorical}(\theta^{(d)}) \\ w_i^{(d)} \mid z_i, \Psi^{(z_i)} &\sim \text{Categorical}(\Psi^{(z_i)}) \end{aligned}$$

We observe the bags of words $X = [X^{(1)}| \cdots |X^{(D)}]$ where $X_i^{(d)} = |\{j : w_j^{(d)} = i\}|$ is the number of occurrences of word i in document d. Since each document is generated independently and $\theta^{(1)}, \ldots, \theta^{(D)}$ are assumed to be i.i.d., the likelihood of the corpus under a topic matrix Ψ is

$$p(X \mid \Psi) = \prod_{d=1}^{D} \mathbb{E}_{\theta^{(d)}} \left[p\left(X \mid \Psi, \theta^{(d)}\right) \right] = \prod_{d=1}^{D} \mathbb{E}_{\theta^{(d)}} \left[\prod_{i=1}^{V} \left(\sum_{k=1}^{K} \Psi_i^{(k)} \theta_k^{(d)} \right)^{X_i^{(d)}} \right]$$

How many bits does it take to approximate the maximum-likelihood Ψ? In the appendix, we show that for any discrete distribution $p = (p_1, \ldots, p_\ell)$ and any $\epsilon > 0$, there is a rounded distribution $\widehat{p}$ that uses $\lceil \log_2(\ell/\epsilon) \rceil$ bits per entry and has $\widehat{p}_i \geq p_i(1-\epsilon)$. By applying this construction to each individual topic distribution, we get the following.

Lemma 2 *Consider any $V \times K$ topic distribution matrix Ψ. For any $\epsilon > 0$ and any integer m, there is a topic matrix $\widehat{\Psi}$ that uses $\lceil \log_2(mV/\epsilon) \rceil$ bits per entry, such that $\log p(x|\Psi) - \log p(x|\widehat{\Psi}) \leq \epsilon$ for all documents x of $\leq m$ words.*

Thus, there exists a polynomially-sized solution to the problem of approximating the ML estimate of the topic modeling problem with a Dirichlet(α) prior on Θ, which we will refer to as TM-MLE(α). Arora et al. (2012) demonstrated that TM-MLE(α) is NP-hard for $\alpha = 1$. Their proof method works for any $\alpha > 0$; for completeness we present the following generalization of their result in the appendix.

Theorem 1 *[Implicit in Arora et al. (2012)] We say a topic matrix Ψ is* c-smooth *for $c > 0$ if $\min_i \max_j \Psi_i^{(j)} \geq c$. Given $\alpha > 0$,* TM-MLE(α) *is NP-hard when restricted to instances in which $K = 2$, all the documents are restricted to have 2 words, and any topic matrix within $\frac{1}{3(1+\alpha)}$ of optimal is guaranteed to be $1/V$-smooth.*

The result in the appendix applies to all symmetric priors and not just the Dirichlet. However, to keep our examples concrete we will only refer to the NP-hardness of TM-MLE(α). Given this result, what can we say about the complexity of MAP estimation and sampling for topic modeling?

3. MAP estimation is as hard as ML estimation

In this section we give a generic reduction from ML estimation to MAP estimation. For a fixed data space $\mathcal{X}$, let (p, Θ) be a parameterized probability model and let ν_0 be a prior probability measure with an associated density q_0. Recall that we use the notation q_X to denote the posterior density given data X. We define the MAP estimation problem as follows.

MAP ESTIMATION: MAP-(p, q_0, Θ)
Input: A sequence of points $X \in \mathcal{X}^n$ and accuracy parameter b in unary.
Output: A parameter $\theta \in \Theta$ satisfying $\log q_X(\theta) \geq \sup_{\theta' \in \Theta} \log q_X(\theta') - 1/b$.

For any instance $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let $Z = (X^{(1)}, \ldots, X^{(k)})$ be a sequence consisting of k copies of X. The lemma below relates the MAP estimate for Z to the ML estimate for X.

Lemma 3 *Pick any $\delta > 0$ and any $\theta \in \Theta$ within δ of the optimal MAP solution for Z, that is,*

$$\log q_Z(\theta) \geq \sup_{\theta' \in \Theta} \log q_Z(\theta') - \delta.$$

Then the log-likelihood of any $\theta' \in \Theta$ can exceed that of θ by at most

$$\log p(X|\theta') - \log p(X|\theta) \leq \frac{1}{k} \left(\delta + \log q_0(\theta) - \log q_0(\theta')\right).$$

Lemma 3 is a promising start to a reduction from ML estimation to MAP estimation, but it requires the prior density $q(\cdot)$ to be bounded above and below, which is often not the case. Recall our example of topic modeling, where we are given a matrix bag of words $X \in \mathbb{Z}^{V \times D}$ and the ML goal is to find the topic matrix $\Psi \in \mathbb{R}^{V \times K}$ which maximizes the objective

$$\log p(X \mid \Psi) = \sum_{d=1}^{D} \log \mathbb{E}_{\theta^{(d)}} \left[\prod_{i=1}^{V} \left(\sum_{k=1}^{K} \Psi_i^{(k)} \theta_k^{(d)} \right)^{X_i^{(d)}} \right]$$

where $\theta^{(d)} \sim$Dirichlet(α). A common choice of prior for Ψ is to assume that the columns of Ψ are drawn i.i.d. from a symmetric Dirichlet(β) distribution. If we let q denote the density of the Dirichlet(β) distribution, this prior density on Ψ can be written as

$$q_0(\Psi) = q(\Psi^{(1)})\, q(\Psi^{(2)}) \cdots q(\Psi^{(K)}).$$

We call the problem of maximizing the resulting posterior TM-MAP(α,β). For $\beta < 1$, the density q is not bounded from above, and TM-MAP(α,β) is consequently not well-defined: infinite a-posteriori scores can be achieved. Hence we will focus on the case $\beta \geq 1$. Here, however, q approaches 0 on the boundary of the simplex, which is problematic for the reduction because the ML solutions Ψ_{ml} of Theorem 1 contain topic distributions that are arbitrarily close to this boundary. Thus, Lemma 3 cannot be straightforwardly applied using $\theta' = \Psi_{ml}$. Instead, we need to ensure that, for any data set, there is some intermediate Ψ that is far enough from the boundary to have non-negligible probability mass under q_0 but has high enough likelihood to be considered a good estimate of Ψ_{ml}.

In summary, we want to guarantee that there are good ML estimates with non-negligible weight under the prior density. The following two sections will help us formalize these notions.

3.1. Admissible distances

Given a likelihood function p, we define the *log-likelihood distance* between $\theta_1, \theta_2 \in \Theta$ as

$$d_p(\theta_1, \theta_2) \;=\; \sup_{x \in \mathcal{X}} |\log p(x \,|\, \theta_1) - \log p(x \,|\, \theta_2)|.$$

Note that for any data sequence $X \in \mathcal{X}^n$, we have

$$d_{p,X}(\theta_1, \theta_2) \;:=\; |\log p(X \,|\, \theta_1) - \log p(X \,|\, \theta_2)| \;\leq\; n\, d_p(\theta_1, \theta_2).$$

In our setting, where we are concerned with how close parameters are in terms of their log-likelihood, d_p is a natural distance to consider. However, it can be difficult to work with directly. Thus we would like to upper-bound it—that is, upper-bound differences in log-likelihood—in terms of more convenient distance functions on parameters, such as ℓ_p distance. Such upper bounds might hold only on well-behaved subsets of parameter space.

Definition 2 *Given $\lambda \geq 1$ and $S \subset \Theta$, we say a distance $d(\cdot, \cdot)$ is* (λ, S)-admissible *if for all $\theta_1, \theta_2 \in S$,*

(i) $d(\theta_1, \theta_1) = 0$ and $d(\theta_1, \theta_2) = d(\theta_2, \theta_1)$, and

(ii) if $d(\theta_1, \theta_2) < 1/\lambda$, then $d_p(\theta_1, \theta_2) \leq \lambda d(\theta_1, \theta_2)$.

We call a λ satisfying (ii) an admissibility constant.

Returning to topic modeling, we define the max-norm distance between two topic matrices as

$$\|\Psi - \Phi\|_{max} = \max_{i,j} |\Psi_i^{(j)} - \Phi_i^{(j)}|.$$

The following lemma, combined with the fact that max-norm distance is a metric, demonstrates that max-norm distance is admissible over the set of smooth topic matrices, where smoothness was defined in Theorem 1.

Lemma 4 *Pick $c, m > 0$ and define $\alpha_0 = \sum \alpha_i$. Let $\mathcal{X}$ be the space of documents with length bounded by m and S be the space of c-smooth matrices. If*

$$\lambda = \left(\frac{2m}{c} + \max\left(1, \left(\frac{\alpha_0 + m}{K}\right)^K\right) \frac{K^{\alpha_0 + 2m - 1/2}}{c^m} \right),$$

then max-norm distance is (λ, S)-admissible.

Although our discussion of topic modeling has assumed a symmetric Dirichlet distribution, the above lemma holds for non-symmetric Dirichlet distributions as well. Additionally, for the instances produced in Theorem 1, $K = m = 2$ and $c = 1/V$; thus the parenthesized term is polynomial in V.

Finally, for $\theta \in \Theta$ and $\epsilon > 0$, define the ball around θ of radius ϵ with respect to distance d as the set $B_d(\theta, \epsilon) = \{\theta' \in \Theta \,:\, d(\theta, \theta') < \epsilon\}$.

3.2. Promise problems

When constructing polynomial time reductions from a language L to a language L', the typical approach is to demonstrate the existence of a polynomial-time computable function $f : \Sigma^* \to \Sigma^*$ such that $x \in L$ if and only if $f(x) \in L'$. However, it is often the case that reductions only demonstrate the hardness of certain well-behaved subsets of languages. Such subsets are captured in the notion of *promise problems*. Given a function $\Pi : \Sigma^* \to \{0, 1\}$, known as the promise, and a language $L \subset \Sigma^*$, the promise problem Π-L is the problem of determining if $x \in L$ given input instances with $\Pi(x) = 1$.

3.3. The reduction

To turn Lemma 3 into a generic reduction, we need to assert that for any valid data sequence X and any $\epsilon > 0$, there is some θ_ϵ whose log-likelihood is within ϵ of θ_{ml} such that $q_0(\theta_\epsilon)$ is bounded below from zero. Such a condition on θ_{ml} is implicitly a restriction on valid inputs X and therefore can be phrased as a promise.

Theorem 3 *Let m be some measure of the size of an input instance, and $\lambda(m)$ any function of this size. Let $S \subset \Theta$ be some subset of parameters, and d be a $(\lambda(m), S)$-admissible distance function. Suppose q_0 satisfies two properties:*

(i) it is bounded above by $2^{\mathrm{poly}(\lambda(m))}$ and

(ii) given $\epsilon > 0$ and $\theta \in S$, there exists $\theta_\epsilon \in B_d(\theta, \epsilon) \cap S$ such that $\log q_0(\theta_\epsilon) \geq -\mathrm{poly}(\lambda(m), 1/\epsilon)$.

If Π is the promise that $\theta_{ml} \in S$, then there is a reduction that is polynomial in the input length and $\lambda(m)$ from Π-MLE-(p, Θ) to MAP-(p, q_0, Θ).

Proof Let $X = (x_1, \ldots, x_n)$ and $b = 1^b$ be the input to MLE-(p, Θ) and let Z denote the sequence consisting of k copies of X. Our input to MAP-(p, q_0, Θ) will be the data sequence Z and the accuracy parameter b. Suppose that the output of this call is θ.

Let $\epsilon = 1/(2bn\lambda(m))$ and take $\theta_\epsilon \in S \cap B_d(\theta_{ml}, \epsilon)$ to be the point with $q_0(\theta_\epsilon) \geq 2^{-\mathrm{poly}(\lambda(m), 1/\epsilon)}$ whose existence is given by assumption (ii). By Lemma 3,

$$\begin{aligned}
\left|\log \frac{p(X \mid \theta_{ml})}{p(X \mid \theta)}\right| &\leq \left|\log \frac{p(X \mid \theta_{ml})}{p(X \mid \theta_\epsilon)}\right| + \left|\log \frac{p(X \mid \theta_\epsilon)}{p(X \mid \theta)}\right| \\
&\leq n \cdot d_p(\theta_{ml}, \theta_\epsilon) + \left|\log \frac{p(X \mid \theta_\epsilon)}{p(X \mid \theta)}\right| \\
&\leq n \cdot \lambda(m) d(\theta_{ml}, \theta_\epsilon) + \frac{1}{k}\left(\frac{1}{b} + \log \frac{q_0(\theta)}{q_0(\theta_\epsilon)}\right) \\
&\leq \lambda(m) n \epsilon + \frac{1}{k}\left(\frac{1}{b} + \mathrm{poly}(\lambda(m), 1/\epsilon)\right).
\end{aligned}$$

By taking k to be a large enough polynomial in b, $\lambda(m)$, and $1/\epsilon$, we can guarantee that θ is within $1/b$ of the ML solution. ■

Returning to topic modeling, let Π denote the promise that the data sequence has only 2 words per document and the ML solution is $1/V$-smooth. From Theorem 1, Π-TM-MLE(α) is NP-hard for any fixed $\alpha > 0$.

Now take λ to be the admissibility constant from Lemma 4 with $c = 1/V$, $m = K = 2$. In the appendix, we show that when the prior q_o is Dirichlet(β) with $\beta \geq 1$, then for any $\epsilon > 0$ and any input instance, there exists a $1/V$-smooth Ψ_ϵ that is within ϵ of Ψ_{ml} in max-norm distance and satisfies $\log q_0(\Psi_\epsilon) \geq -\operatorname{poly}(\lambda, 1/\epsilon)$. Letting S be the set of c-smooth matrices, Theorem 3 immediately gives us the following.

Theorem 4 *For any fixed $\alpha > 0$ and $\beta \geq 1$,* TM-MAP(α, β) *is NP-hard.*

4. Approximate sampling is as hard as ML estimation

We now turn to a reduction from ML estimation to posterior sampling. As pointed out in Section 1, total variation distance is not a suitable metric for approximate sampling in continuous domains. Thus we begin by describing a commonly-used alternative: Wasserstein distances. Afterwards, we provide reductions from ML estimation to approximate sampling and demonstrate how it applies to topic modeling.

4.1. Wasserstein approximate sampling

Given two probability distributions μ and ν and a metric d over Θ, the t-th Wasserstein distance is

$$\mathcal{W}_t(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma} d(X, Y)^t \right)^{1/t}.$$

Here, $\Gamma(\mu, \nu)$ is the space of all couplings of μ and ν, i.e. probability distributions over $\Theta \times \Theta$ whose marginals are μ and ν.

When (Θ, d) is the trivial metric space, i.e. $d(\theta, \theta') = 1$ whenever $\theta \neq \theta'$, then the t-th Wasserstein distance is just the t-th root of the total variation distance. In general, however, Wasserstein distances differ from total variation distance. In particular, when (Θ, d) admits an ϵ-covering $\widehat{\Theta}$, any distribution ν can be approximated within Wasserstein distance ϵ by a discrete distribution.

Theorem 5 *Let $\epsilon > 0$ and ν be a distribution over a metric space (Θ, d). If $\widehat{\Theta}$ is a countable ϵ-cover of Θ then there exists a measure $\widehat{\nu}$ over $\widehat{\Theta}$ such that $W_t(\nu, \widehat{\nu}) \leq \epsilon$.*

In particular, Theorem 5 demonstrates that when Θ is a bounded subset of $\mathbb{R}^k$ and d is an ℓ_p norm,[1] then any distribution ν over Θ has a discrete distribution within ϵ of ν in Wasserstein distance that is supported on points which can be written using a polynomial number of bits. Given this observation, we define the Wasserstein approximate posterior sampling problem as follows.

WASSERSTEIN APPROXIMATE POSTERIOR SAMPLING: $\mathcal{W}_t$-APPROX-SAMPLING-(p, ν_0, Θ)
Input: A sequence of points $X \in \mathcal{X}^n$, accuracy parameter b in unary.
Output: A random draw $\theta \sim \nu$, where $\mathcal{W}_t(\nu, \nu_X) \leq 1/b$.

1. Alternatively Θ could be unbounded and $d(x, y) = \min\{\|x - y\|_p, B\}$ for some $B > 0$.

4.2. The reduction

Recall the definition of $d_{p,X}$ as

$$d_{p,X}(\theta_1, \theta_2) \;=\; |\log p(X|\theta_1) - \log p(X|\theta_2)|$$

for a data sequence $X \in \mathcal{X}^n$ and $\theta_1, \theta_2 \in \Theta$. The following lemma tells us the rate at which the posterior of a duplicated data sequence concentrates around the maximum likelihood solution.

Lemma 5 *Take any $\epsilon, \delta > 0$ and $X \in \mathcal{X}^n$. If Z is the sequence created by duplicating X*

$$k \;\geq\; \frac{2}{\epsilon}\left(\log\left(\frac{1}{\delta} - 1\right) + \log\left(\frac{1 - \nu_0(B_{d_{p,X}}(\theta_{ml}, \epsilon))}{\nu_0(B_{d_{p,X}}(\theta_{ml}, \epsilon/2))}\right)\right)$$

times then $\nu_Z(B_{d_{p,X}}(\theta_{ml}, \epsilon)) \geq 1 - \delta$. (Recall θ_{ml} is the maximum-likelihood solution for X.)

With this lemma in hand, the reduction from ML estimation to Wasserstein approximate sampling can be given. The proof is deferred to the appendix.

Theorem 6 *Let m be some measure of the size of an input instance X, and let $\lambda(m)$ be any function of this size. Let d be a distance function and $S \subset S' \subset \Theta$ be subsets satisfying*

(i) if $\theta \in S$ then $B_d(\theta, 1/\lambda(m)) \subset S'$ and

(ii) d is $(\lambda(m), S')$-admissible.

If Π is the promise that $B_{d_{p,X}}(\theta_{ml}, 1/\lambda(m)) \subset S$ and $\nu_0(B_d(\theta_{ml}, \epsilon)) \geq 2^{-\text{poly}(\lambda(m), 1/\epsilon)}$ for all $\epsilon > 0$, then Π-MLE-$(p, \Theta) \leq_P \mathcal{W}_t$-APPROX-SAMPLING-$(p, \nu_0, \Theta)$ under randomized reductions which are polynomial in the input size and $\lambda(m)$.

To see how this reduction applies to our topic modeling scenario, recall that our posterior was formed by considering the likelihood in TM-MLE(α) and placing a Dirichlet(β) prior on each of the columns of Ψ. We call the problem of sampling from this distribution TM-APPROX-SAMPLING(α, β).

Notice that Theorem 6 only requires a lower bound on the probability of neighborhoods of ML solutions and not any type of upper bound as in Theorem 3. Thus, for approximate posterior sampling in topic modeling, we do not need to place the same lower bound on β as in MAP estimation. In particular, we prove the following in the appendix.

Theorem 7 *There is no poly-time algorithm for* TM-APPROX-SAMPLING(α, β) *for any $\alpha, \beta > 0$ unless NP=RP.*

4.3. Discretized total variation distance

Wasserstein distance is not the only way to compare continuous and discrete distributions. In the appendix, we introduce a generalization of total variation distance that captures the disagreement between distributions at a specified granularity. Under the same conditions as Theorem 6, we show ML estimation can be reduced to approximate posterior sampling with respect to this alternate distance.

5. Application: Mixtures of Gaussians

Consider the following maximum likelihood estimation problem for mixtures of k spherical Gaussians with the same variance.

> SAME VARIANCE MLE: MLE-MOG-SV(k)
> *Input*: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; unary parameter b.
> *Output*: Parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ such that $LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \pi_j N(x_i; \mu_j, \sigma^2)\right)$ is within an additive factor $1/b$ of optimal.

In the above, we have used boldface symbols to pack parameters into vectors, i.e. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$. Tosh and Dasgupta (2018) showed MLE-MOG-SV(k) is NP-hard for $k \geq 2$.

In this section, we examine the complexity of Bayesian estimation for mixtures of Gaussians. We consider common conjugate priors on the mixing weights, means, and variance. In particular, we place a symmetric Dirichlet(γ) prior on the mixing weights and a Normal-Inverse-Gamma($\alpha, \beta, \mu_0, n_0$) prior on the means and variance, wherein the variance σ^2 is first drawn from an inverse gamma distribution, IG(α, β), and the means are drawn i.i.d. from a normal distribution, $\mathcal{N}(\mu_0, \sigma^2/n_0 I_d)$. The full generative process is spelled out below.

$$(\pi_1, \ldots, \pi_k) \sim \text{Dirichlet}(\gamma) \qquad \mu_j \,|\, \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/n_0 I_d)$$

$$\sigma^2 \sim \text{IG}(\alpha, \beta) \qquad x_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \sigma^2 \sim \sum_{i=1}^{k} \pi_i \mathcal{N}(\mu_i, \sigma^2 I_d)$$

Let $\boldsymbol{\omega} = (\alpha, \beta, \gamma, \mu_0, n_0)$ denote a fixed set of hyper-parameters. We call the corresponding MAP estimation problem MAP-MOGS($k, \boldsymbol{\omega}$) and the corresponding approximate sampling problem APPROX-SAMPLING-MOGS($k, \boldsymbol{\omega}$). We will show these problems are hard when $k = 2$.

As in the topic modeling setting, we cannot simply start with a reduction from MLE-MOGS-SV(k). We will need a well-behaved promise version of this problem.

Theorem 8 *Let Π be the promise that there exists a low-order polynomial $\rho(\cdot, \cdot, \cdot)$ such that all input data points satisfy $\|x\| \leq \rho(n, d, k)$ and if $\theta_{ml} = (\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*)$ is a maximum likelihood solution and $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ is within 1 of optimality, then*

(i) $\|\mu_j\| \leq \rho(n, d, k)$ *for all* j,

(ii) $\sigma^2 \geq 1/\rho(n, d, k)$,

(iii) $\pi_j > 0$ *for all* j, *and*

(iv) $\pi_j^* \geq 1/\rho(n, d, k)$ *for all* j.

Then Π-MLE-MOGS-SV(k) *is NP-hard for* $k \geq 2$.

The proof of Theorem 8 is deferred to the appendix. Theorem 8 implies, among other things, that if we are reducing from Π-MLE-MOGS-SV(k), we may restrict our data space $\mathcal{X}$ to consist of points x satisfying $\|x\| \leq \rho(n, d, k)$.

As before, we also need a suitable distance in parameter space. We will consider the following distance between two parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ and $\hat{\theta} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}, \hat{\sigma})$:

$$d(\theta, \hat{\theta}) = \max\left\{\|\mu_i - \hat{\mu}_i\|^2, |\log \pi_i - \log \hat{\pi}_i|, |\sigma^2 - \hat{\sigma}^2|\right\}.$$

The following lemma, whose proof appears in the appendix, shows that this distance is admissible for well-behaved parameters.

Lemma 6 *Let $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ and $\hat{\theta} = (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\sigma})$ be two parameter vectors satisfying $\pi_j, \hat{\pi}_j > 0$ for all j. If $\mathcal{X} = \{x : \|x\| \leq B\}$ then $d_p(\theta, \hat{\theta}) \leq d(\theta, \hat{\theta}) \operatorname{poly}(1/\sigma_i^2, 1/\hat{\sigma}_i^2, \|\mu_i\|, \|\hat{\mu}_i\|, B)$.*

The next lemma, whose proof appears in the appendix, provides bounds on the prior density.

Lemma 7 *Let q and ν be the prior density and measure, respectively, for the Bayesian mixture of two spherical Gaussians generative model with fixed parameters $\alpha, \beta, \gamma, \mu_0, n_0$. For any $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ and any $\epsilon > 0$, we have*

(i) $\log q(\theta) \geq -\operatorname{poly}(1/\pi_i, 1/\sigma, \|\mu_i\|, d)$,

(ii) $\log \nu(B_d(\theta, \epsilon)) \geq -\operatorname{poly}(1/\pi_i, 1/\sigma, \|\mu_i\|, d, 1/\epsilon)$, *and*

(iii) if $\gamma \geq 1$, then $\log q(\theta) \leq \operatorname{poly}(d)$.

Given the above results, we can now demonstrate the hardness of MAP estimation and approximate posterior sampling.

Theorem 9 *Let $\boldsymbol{\omega} = (\alpha, \beta, \gamma, \mu_0, n_0)$ for $\alpha, \beta, \gamma, n_0 > 0$ and $\mu_0 \in \mathbb{R}^d$. Then*

(a) MAP-MOGS$(k = 2, \boldsymbol{\omega})$ *is NP-hard if $\gamma \geq 1$.*

(b) APPROX-SAMPLING-MOGS$(k = 2, \boldsymbol{\omega})$ *is NP-hard for all $\gamma > 0$.*

Proof We show (a) by reducing from Π-MLE-MOGS-SV(k) for $k = 2$ under the condition $\gamma \geq 1$. Since k is a constant, we may take the polynomial ρ from Theorem 8 to only have two free arguments. Let q denote the prior density and let

$$S = \left\{ (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) : \frac{1}{\sigma^2}, \frac{1}{\pi_i}, \|\mu_i\|^2 \leq \rho(n, d) \text{ for all } i \right\}.$$

Then we have the following.

(i) $\log q(\theta) \leq \operatorname{poly}(d, n_0, \alpha, \beta, \gamma, \|\mu_0\|)$ for any parameter $\theta \in \Theta$ (Lemma 7).

(ii) $\log q(\theta) \geq -\operatorname{poly}(n, d, n_0, \alpha, \beta, \gamma, \|\mu_0\|)$ for any parameter $\theta \in S$ (Lemma 7).

(iii) d is $(\operatorname{poly}(n, d), S)$-admissible (Lemma 6).

(iv) Π guarantees that $\theta_{ml} \in S$.

Given the above, Theorem 3 implies (a). The proof for (b) is provided in the appendix. ■

Acknowledgements

The authors are grateful to the anonymous reviewers for their feedback and to the National Science Foundation for support through awards CCF-1740833 and CCF-1813160.

References

D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond SVD. In *Proceedings of the 53rd IEEE Annual Symposium on Foundations of Computer Science*, pages 1–10, 2012.

A. Baker. *Transcendental Number Theory*. Cambridge University Press, 1975.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *arXiv preprint arXiv:1507.02564*, 2015.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323, 2018.

B. Chor and T. Tuller. Finding a maximum likelihood tree is hard. *Journal of the ACM*, 53(5): 722–744, 2006.

V. Guruswami and A. Vardy. Maximum-likelihood decoding of Reed-Solomon codes is NP-hard. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 470–478, 2005.

A. T. Kalai and S. Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

L. M. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, 1:277–330, 1953.

K. W. Ng, G.-L. Tian, and M.-L. Tang. *Dirichlet and Related Distributions: Theory, Methods and Applications*. Wiley, 2011.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

D. Sontag and D. Roy. Complexity of inference in latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1008–1016, 2011.

F. Topsøe. Some bounds for the logarithmic function. *Inequal. Theory Appl.*, pages 137–151, 2004.

C. Tosh and S. Dasgupta. Maximum likelihood estimation for mixtures of spherical Gaussians is NP-hard. *JMLR*, 18(1):1–11, 2018.

Appendix A. Proofs from Section 1

Lemma 8 *When fitting a mixture model* $\frac{1}{2}\mathcal{N}(-\mu, 1) + \frac{1}{2}\mathcal{N}(\mu, 1)$ *to the data set of three points* $\{-2, 0, 2\}$*, the maximum-likelihood choice of* μ *is irrational.*

Proof

Writing out the log-likelihood function,

$$\begin{aligned}\ln p(-2, 0, 2 \mid \mu) &= \ln\left(\frac{1}{2\sqrt{2\pi}}e^{-\mu^2/2} + \frac{1}{2\sqrt{2\pi}}e^{-\mu^2/2}\right) \\ &\quad + 2\ln\left(\frac{1}{2\sqrt{2\pi}}e^{-(2-\mu)^2/2} + \frac{1}{2\sqrt{2\pi}}e^{-(2+\mu)^2/2}\right) \\ &= -2\ln(2\sqrt{2\pi}) - 4 - \frac{3\mu^2}{2} + 2\ln\left(e^{2\mu} + e^{-2\mu}\right).\end{aligned}$$

Taking derivatives of the log-likelihood equation with respect to μ,

$$\frac{d}{d\mu}\ln p(-2, 0, 2 \mid \mu) = -3\mu + 4\tanh(2\mu).$$

This has two non-negative roots, one of which is zero. Evaluating the second derivative at zero, we have

$$\frac{d^2}{d^2\mu}\ln p(-2, 0, 2 \mid \mu)|_{\mu=0} = 8\,\mathrm{sech}^2(0) - 3 = 5.$$

Thus zero is a local minimum. Because $-3\mu + 4\tanh(2\mu)$ tends to $-\infty$ as μ goes to ∞, we can conclude that the other nonnegative root is the maximum likelihood estimate. Expanding the $\tanh$, we see that this root satisfies $(3\mu - 4)e^{2\mu} + (3\mu + 4)e^{-2\mu} = 0$; an application of the Lindemann-Weierstrass theorem (Baker, 1975, Chapter 1) then tells us that it must be transcendental, implying it is also irrational. ■

Appendix B. Proofs from Section 2

We start with a general result about the polynomial approximability of discrete distributions, and then consider an application to topic models.

Lemma 9 *Consider any distribution with finite support, say* $p = (p_1, \ldots, p_\ell)$*. Pick any positive integer* M*. Then there is a distribution* $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_\ell)$ *such that:*

- *Each* $\widehat{p}_i$ *is a non-zero multiple of* $1/M$*.*
- *For each* i*, we have* $\widehat{p}_i \geq (1 - \ell/M)p_i$*.*

Proof First define an intermediate distribution $\overline{p}$ as follows:

$$\overline{p}_i = (1 - \ell/M)p_i, \text{ rounded up to the nearest multiple of } 1/M.$$

Therefore, $(1-\ell/M)p_i \leq \overline{p}_i \leq (1-\ell/M)p_i + (1/M)$, and $\sum_i \overline{p}_i$ is some multiple of $1/M$ that is ≤ 1. To get $\widehat{p}$, take $\overline{p}$ and add multiples of $1/M$ to any coordinate(s) until the sum of the coordinates equals 1. ■

This construction can be used to show that the maximum-likelihood topic model admits a concise approximation.

Lemma 10 *Consider any $V \times K$ topic distribution matrix Ψ. For any $\epsilon > 0$ and any integer m, there is a topic matrix $\widehat{\Psi}$ that uses $\lceil \log_2(mV/\epsilon) \rceil$ bits per entry, such that $\log p(x|\Psi) - \log p(x|\widehat{\Psi}) \leq \epsilon$ for all documents x of $\leq m$ words.*

Proof Obtain $\widehat{\Psi}$ by applying the previous lemma to each individual topic distribution, with $M = \lceil 2mV/\epsilon \rceil$. Pick any document x of length m. Letting q denote the prior on topic weights (that is, a prior on the K-simplex), and letting $z \in \{1,\ldots,K\}^m$ denote the topic assignments to the words x_i, we have

$$\begin{aligned}\Pr(x|\Psi) &= \int q(\theta) \sum_z \Pr(z|\theta)\Pr(x|z,\Psi)\, d\theta \\ &= \sum_z \left(\prod_{i=1}^m \Psi_{x_i}^{(z_i)} \right) \int q(\theta)\Pr(z|\theta)\, d\theta\end{aligned}$$

By construction, for any z,

$$\prod_{i=1}^m \widehat{\Psi}_{x_i}^{(z_i)} \;\geq\; \prod_{i=1}^m \left((1-\epsilon/2m)\Psi_{x_i}^{(z_i)} \right) \;=\; (1-\epsilon/2m)^m \prod_{i=1}^m \Psi_{x_i}^{(z_i)} \;\geq\; e^{-\epsilon} \prod_{i=1}^m \Psi_{x_i}^{(z_i)},$$

and thus $\Pr(x|\Psi) \leq e^{\epsilon}\Pr(x|\widehat{\Psi})$, as claimed. ■

B.1. Proof of Theorem 1: hardness of finding the maximum-likelihood topic model

Our goal is to prove the following theorem.

Theorem 1 *[Implicit in Arora et al. (2012)] We say a topic matrix Ψ is c-*smooth *for $c > 0$ if $\min_i \max_j \Psi_i^{(j)} \geq c$. Given $\alpha > 0$, TM-MLE(α) is NP-hard when restricted to instances in which $K = 2$, all the documents are restricted to have 2 words, and any topic matrix within $\frac{1}{3(1+\alpha)}$ of optimal is guaranteed to be $1/V$-smooth.*

In fact, we will prove a more general result. Let Δ^N be the N-simplex, i.e.

$$\Delta^N \;=\; \left\{ \theta \in \mathbb{R}^N \;:\; \sum_{i=1}^N \theta_i = 1 \text{ and } \theta_i \geq 0 \right\}.$$

Theorem 10 *Let $\lambda_S, \lambda_X \geq 0$ and ν_0 be a distribution over Δ^K such that for $\theta \sim \nu_0$*

- $\mathbb{E}[\theta_1^2] = \cdots = \mathbb{E}[\theta_k^2] = \lambda_S$ *and*

- $\mathbb{E}[\theta_i\theta_j] = \lambda_X$ *for all* $i \neq j$.

Then TM-MLE-ν_0*, the problem of maximizing the same objective as* TM-MLE(α) *with the Dir(α) prior replaced with* ν_0*, is NP-hard when* $\lambda_S > \lambda_X \geq 0$ *when* $K = 2$*, there are exactly two words in each document, and any topic matrix within* $\frac{\lambda_S - \lambda_X}{3\lambda_S}$ *of optimal must be* $1/|V|$*-smooth.*

To see how this implies Theorem 1, note that for $\theta \sim \text{Dir}(\alpha)$ and $i \neq j$,

$$\lambda_S = \mathbb{E}[\theta_i^2] = \frac{\Gamma(K\alpha)\Gamma(\alpha+2)}{\Gamma(K\alpha+2)\Gamma(\alpha)} = \frac{\alpha+1}{K(\alpha K+1)} > \frac{\alpha}{K(\alpha K+1)}$$
$$= \frac{\Gamma(K\alpha)(\Gamma(\alpha+1))^2}{\Gamma(K\alpha+2)(\Gamma(\alpha))^2} = \mathbb{E}[\theta_i\theta_j] = \lambda_X.$$

Further, we have

$$\frac{\lambda_S - \lambda_X}{3\lambda_S} = \left(\frac{\alpha+1}{K(\alpha K+1)} - \frac{\alpha}{K(\alpha K+1)}\right)\frac{K(\alpha K+1)}{3(\alpha+1)} = \frac{1}{3(\alpha+1)}.$$

The proof of Theorem 10 follows the reduction from Arora et al. (2012) very closely. We start with an instance of MINIMUM-BISECTION. Here the input is a graph $G = (V, E)$ with $|V| = n$ even and $|E| = m$, and the goal is to find a cut (S, T) such that $|S| = n/2 = |T|$ and $|E(S, T)|$ is minimized.

Beginning with G, we construct our instance of TM-MLE-ν_0 as follows. The vocabulary is the set of vertices V. Our corpus will consist of the following documents:

- for each word $i \in V$, create N documents with only the word i repeated twice, and
- for each edge $(i, j) \in E$, create one document with only the word i and the word j.

Here N is a polynomial of n, m, λ_S, and λ_X to be determined later. Given a document with words i and j (possibly equal) and a topic matrix $\Psi = [\Psi^{(1)}|\Psi^{(2)}]$, what is the likelihood of the document under Ψ? This is simply

$$p(i, j \mid \Psi) = \mathbb{E}[\theta_1^2]\Psi_i^{(1)}\Psi_j^{(1)} + \mathbb{E}[\theta_1\theta_2]\left(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)}\right) + \mathbb{E}[\theta_2^2]\Psi_i^{(2)}\Psi_j^{(2)}$$
$$= \lambda_S\langle\Psi_i, \Psi_j\rangle + \lambda_X\left(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)}\right)$$

where $\Psi_i = (\Psi_i^{(1)}, \Psi_i^{(2)})$. Then the objective is to maximize the following function:

$$F(\Psi) = \sum_{\text{document}=(i,j)} \ln\left(\lambda_S\langle\Psi_i, \Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)})\right)$$
$$= \sum_{i\in V} N\ln\left(\lambda_S\|\Psi_i\|_2^2 + 2\lambda_X\Psi_i^{(1)}\Psi_i^{(2)}\right)$$
$$+ \sum_{(i,j)\in E} \ln\left(\lambda_S\langle\Psi_i, \Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)})\right).$$

For any bisection (S, T), define the *canonical solution* $\Psi = \Psi(S, T)$ to be the topic matrix which satisfies $\Psi_i^{(1)} = 2/n$ and $\Psi_i^{(2)} = 0$ for all $i \in S$; and $\Psi_i^{(1)} = 0$ and $\Psi_i^{(2)} = 2/n$ for all $i \in T$. We'll

see that the maximum-likelihood solution (or an approximation thereof) is approximately canonical, and therefore uniquely specifies a cut.

Write $F(\Psi) = G(\Psi) + H(\Psi)$, where

$$G(\Psi) = \sum_{i \in V} N \ln \left(\lambda_S \|\Psi_i\|_2^2 + 2\lambda_X \Psi_i^{(1)} \Psi_i^{(2)} \right)$$
$$H(\Psi) = \sum_{(i,j) \in E} \ln \left(\lambda_S \langle \Psi_i, \Psi_j \rangle + \lambda_X (\Psi_i^{(1)} \Psi_j^{(2)} + \Psi_j^{(1)} \Psi_i^{(2)}) \right).$$

When N is made large enough, G dominates H.

Each of the n rows of Ψ is a pair $(\Psi_i^{(1)}, \Psi_i^{(2)})$. We start by characterizing approximately optimal solutions subject to specific row-sum constraints.

Lemma 11 *Suppose the row-sums are constrained to be $\Psi_i^{(1)} + \Psi_i^{(2)} = r_i$, for some $r_1, \ldots, r_n$ summing to 2. Then:*

(a) G is bounded as follows:

$$G(\Psi) \;\leq\; \sum_{i=1}^{n} N \ln(\lambda_S r_i^2) - N \sum_{i=1}^{n} \frac{\lambda_S - \lambda_X}{\lambda_S} \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}$$

with equality if each row has $\min(\Psi_i^{(1)}, \Psi_i^{(2)}) = 0$.

(b) H lies in a smaller range:

$$m \ln \lambda_X \;\leq\; H(\Psi) - \sum_{(i,j) \in E} \ln(r_i r_j) \;\leq\; m \ln \lambda_S.$$

Proof To see (a), first note that

$$\frac{2\Psi_i^{(1)} \Psi_i^{(2)}}{r_i^2} = \frac{2\Psi_i^{(1)} \Psi_i^{(2)}}{(\Psi_i^{(1)} + \Psi_i^{(2)})^2} \geq \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{\Psi_i^{(1)} + \Psi_i^{(2)}} = \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}.$$

Therefore, we can write

$$\begin{aligned}
G(\Psi) &= \sum_{i=1}^{n} N \ln \left(\lambda_S \|\Psi_i\|_2^2 + 2\lambda_X \Psi_i^{(1)} \Psi_i^{(2)} \right) \\
&= \sum_{i=1}^{n} N \ln \left(\lambda_S r_i^2 \right) + N \sum_{i=1}^{n} \ln \left(\frac{r_i^2 - 2\Psi_i^{(1)} \Psi_i^{(2)}}{r_i^2} + \frac{2\lambda_X \Psi_i^{(1)} \Psi_i^{(2)}}{\lambda_S r_i^2} \right) \\
&= \sum_{i=1}^{n} N \ln \left(\lambda_S r_i^2 \right) + N \sum_{i=1}^{n} \ln \left(1 - \frac{2\Psi_i^{(1)} \Psi_i^{(2)}}{r_i^2} \cdot \frac{\lambda_S - \lambda_X}{\lambda_S} \right) \\
&\leq \sum_{i=1}^{n} N \ln \left(\lambda_S r_i^2 \right) - N \sum_{i=1}^{n} \frac{2\Psi_i^{(1)} \Psi_i^{(2)}}{r_i^2} \cdot \frac{\lambda_S - \lambda_X}{\lambda_S}
\end{aligned}$$

$$\leq \sum_{i=1}^{n} N \ln(\lambda_S r_i^2) - N \sum_{i=1}^{n} \frac{\lambda_S - \lambda_X}{\lambda_S} \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}.$$

(b) follows directly from algebra and applying the inequality $\lambda_S > \lambda_X$. ■

This immediately gives us the following corollary.

Corollary 11 *Fix any row-sums $r_1, \ldots, r_n$, and any $\Delta > 0$. Let Ψ be any solution whose $F(\cdot)$ value is within Δ of optimal, subject to these constraints. Then for each i,*

$$\frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i} \leq \frac{1}{N} \left(\frac{m\lambda_S}{\lambda_X} + \frac{\Delta \lambda_S}{\lambda_S - \lambda_X} \right).$$

Thus each row has one entry that is approximately zero, whereupon, returning to Lemma 11, we see that $G(\cdot)$ is roughly $2N \sum_i \ln r_i$, ignoring constants. The following technical lemma then implies that in an approximately optimal solution, all row-sums must be roughly equal.

Lemma 12 *Subject to the constraint that $r_1, \ldots, r_n$ are nonnegative and sum to 2:*

(a) The quantity $\sum_i \ln r_i$ is maximized when the r_i are equal, in which case

$$\sum_{i=1}^{n} \ln r_i = n \ln \frac{2}{n}.$$

(b) Pick any $\epsilon > 0$. If there is some $r_i \notin [\frac{2}{n}(1-\epsilon), \frac{2}{n}(1+\epsilon)]$, then no matter how the other r_j are set,

$$\sum_{i=1}^{n} \ln r_i \leq n \ln \frac{2}{n} - \frac{1}{4}\epsilon^2.$$

Proof (a) follows directly from Jensen's inequality. To see (b), we make use of the following logarithmic inequalities, which can be found in (Topsøe, 2004). For $0 \leq x < 1$,

$$\ln(1+x) \leq \frac{x}{2} \cdot \frac{2+x}{1+x} \qquad \text{and} \qquad \ln(1-x) \leq \frac{-2x}{2-x}.$$

Now let $\delta > 0$ and suppose that there is some i such that $r_i = 2(1+\delta)/n$. Then by (a),

$$\begin{aligned}
\sum_{j=1}^{n} \ln r_j &= \ln\left(\frac{2}{n}(1+\delta)\right) + \sum_{j \neq i} \ln r_j \\
&\leq \ln\left(\frac{2}{n}(1+\delta)\right) + (n-1)\ln\left(\frac{1}{n-1}\left(2 - \frac{2}{n}(1+\delta)\right)\right) \\
&= n \ln \frac{2}{n} + \ln(1+\delta) + (n-1)\ln\left(1 - \frac{\delta}{n-1}\right) \\
&\leq n \ln \frac{2}{n} + \frac{\delta}{2} \cdot \frac{2+\delta}{1+\delta} - \frac{2\delta(n-1)}{2(n-1)-\delta}
\end{aligned}$$

$$\leq n \ln \frac{2}{n} - \frac{1}{4}\delta^2$$

They proof for the case where $r_i = 2(1-\delta)/n$ is similar. Thus, we have (b). ■

The $G(\cdot)$ function dominates $H(\cdot)$ and forces (approximately) canonical solutions.

Lemma 13 *Pick any $0 < \epsilon < 1$ and any $\Delta > 0$. Define*

$$N_0 = \frac{2}{\epsilon^2}\left(\Delta + m \ln\left(\frac{\lambda_S}{\lambda_X}\frac{n^2}{4}\right)\right)$$
$$N_1 = \frac{1}{\epsilon}\left(\frac{m\lambda_S}{\lambda_X} + \frac{\Delta\lambda_S}{\lambda_S - \lambda_X}\right).$$

Let Ψ^* be a maximizer of $F(\cdot)$. Then, if $N \geq \max(N_0, N_1)$, any solution Ψ with $F(\Psi) \geq F(\Psi^*) - \Delta$ must satisfy the following conditions for each i:

(a) $\Psi_i^{(1)} + \Psi_i^{(2)} \in [\frac{2}{n}(1-\epsilon), \frac{2}{n}(1+\epsilon)]$.

(b) $\min(\Psi_i^{(1)}, \Psi_i^{(2)}) \leq \epsilon \cdot \frac{2}{n}$.

Proof Let Φ be any canonical solution (which trivially implies $F(\Psi^*) \geq F(\Phi)$), let Ψ be a solution satisfying $F(\Psi) \geq F(\Psi^*) - \Delta$, and let $r_1, \ldots, r_n$ be the row sums of Ψ. Then because Φ is canonical, we know from the above lemmas

$$\begin{aligned}
\Delta &\geq F(\Phi) - F(\Psi) \\
&= G(\Phi) - G(\Psi) + H(\Phi) - H(\Psi) \\
&\geq G(\Phi) - G(\Psi) + \sum_{(i,j)\in E} \ln\left(\frac{4}{n^2}\right) + m \ln \lambda_X - \sum_{(i,j)\in E} \ln(r_i r_j) - m \ln \lambda_S \\
&= G(\Phi) - G(\Psi) - m \ln\left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X}\right) - \sum_{(i,j)\in E} \ln(r_i r_j) \\
&\geq N\left(n \ln\left(\lambda_S \frac{4}{n^2}\right) - \sum_{i=1}^{n} \ln(\lambda_S r_i^2) + \sum_{i=1}^{n} \frac{\lambda_S - \lambda_X}{\lambda_S} \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}\right) \\
&\qquad - m \ln\left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X}\right) - \sum_{(i,j)\in E} \ln(r_i r_j).
\end{aligned}$$

Now suppose by contradiction that Ψ does not satisfy condition (a). We know that because the columns of Ψ sum to 1, it must be the case that $r_i r_j \leq 1$. Applying Lemma 12(b),

$$\begin{aligned}
\Delta &\geq N\left(n \ln\left(\lambda_S \frac{4}{n^2}\right) - \sum_{i=1}^{n} \ln(\lambda_S r_i^2) + \sum_{i=1}^{n} \frac{\lambda_S - \lambda_X}{\lambda_S} \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}\right) - m \ln\left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X}\right) \\
&> N\left(n \ln\left(\lambda_S \frac{4}{n^2}\right) - n \ln\left(\lambda_S \frac{4}{n^2}\right) + \frac{\epsilon^2}{2}\right) - m \ln\left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X}\right) \\
&= \frac{N\epsilon^2}{2} - m \ln\left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X}\right).
\end{aligned}$$

But this implies that

$$N < \frac{2}{\epsilon^2}\left(\Delta + m \ln\left(\frac{\lambda_S}{\lambda_X}\frac{n^2}{4}\right)\right) = N_0$$

which is a contradiction.

To see that Ψ must satisfy condition (b), note that by Corollary 11, if Ψ did not satisfy condition (b), then $F(\Psi)$ could not be within Δ of $F(\Psi^*)$. ■

Once we are within the realm of approximately canonical solutions, which uniquely designate a bisection cut, the lower-order term $H(\cdot)$ serves to choose a cut of small size.

Lemma 14 *Pick any $0 < \epsilon < 1$. We will describe any Ψ that satisfies conditions (a) and (b) of Lemma 13 as being ϵ-approximately canonical.*

(a) For any canonical solution Ψ,

$$H(\Psi) \;=\; m \ln \frac{4\lambda_S}{n^2} - |\text{cut}(\Psi)| \cdot \ln \frac{\lambda_S}{\lambda_X}.$$

(b) For any ϵ-approximately canonical solution Ψ,

$$H(\Psi) \;\leq\; m \ln \frac{4\lambda_S}{n^2} - |\text{cut}(\Psi)| \cdot \ln \frac{\lambda_S}{\lambda_X} + 2m\epsilon\frac{\lambda_S}{\lambda_X}.$$

Proof Recall that

$$H(\Psi) = \sum_{(i,j)\in E} \ln\left(\lambda_S\langle\Psi_i,\Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)})\right).$$

Therefore, if Ψ is a canonical solution corresponding to the bisection (S,T), then if $E(S,T)$ denotes the subset of edges with one endpoint in S and the other in T we have

$$\begin{aligned} H(\Psi) &= \sum_{(i,j)\in E(S,T)} \ln\frac{4\lambda_X}{n^2} + \sum_{(i,j)\in E\setminus E(S,T)} \ln\frac{4\lambda_S}{n^2} \\ &= m\ln\frac{4\lambda_S}{n^2} - |\text{cut}(\Psi)|\cdot\ln\frac{\lambda_S}{\lambda_X}. \end{aligned}$$

Now let Ψ be an ϵ-approximately canonical solution. Use it to define a cut (S,T) in the natural way:

$$S = \{i : \Psi_i^{(2)} \leq 2\epsilon/n\},\;\; T = [n]\setminus S.$$

Given an edge $(i,j)\in E$, how do we bound $Q_{i,j}(\Psi) = \lambda_S\langle\Psi_i,\Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)})$? We consider two cases.

Case 1: $(i,j)\in E\setminus E(S,T)$. Assume w.l.o.g. that $i,j\in S$. Then because Ψ is ϵ-approximately canonical, we know $\|\Psi_i\|_1, \|\Psi_j\|_1 \in [\frac{2}{n}(1-\epsilon), \frac{2}{n}(1+\epsilon)]$ and $\Psi_i^{(2)}, \Psi_j^{(2)} \leq \frac{2}{n}\epsilon$. Letting $\Psi_i^{(2)} = \frac{2}{n}\delta_i$ and $\Psi_j^{(2)} = \frac{2}{n}\delta_j$, we have

$$\begin{aligned} Q_{i,j}(\Psi) &= \lambda_S\langle\Psi_i,\Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)}) \\ &\leq \frac{4}{n^2}\left(\lambda_S((1+\epsilon-\delta_i)(1+\epsilon-\delta_j)+\delta_i\delta_j) + \lambda_X((1+\epsilon-\delta_j)\delta_i + (1+\epsilon-\delta_i)\delta_j)\right) \\ &= \frac{4}{n^2}\left(\lambda_S(1+\epsilon)^2 + (\lambda_S-\lambda_X)(2\delta_i\delta_j - (1+\epsilon)(\delta_i+\delta_j)\right) \end{aligned}$$

Since $\delta_i, \delta_j \leq \epsilon < 1$ and $\lambda_S > \lambda_X$, the above is maximized whenever $\delta_i = \delta_j = 0$. Thus,

$$Q_{i,j}(\Psi) \leq \frac{4\lambda_S(1+\epsilon)^2}{n^2}.$$

Case 2: $(i,j) \in E(S,T)$. Assume w.l.o.g. that $i \in S$ and $j \in T$. Then because Ψ is ϵ-approximately canonical, we know $\|\Psi_i\|_1, \|\Psi_j\|_1 \in [\frac{2}{n}(1-\epsilon), \frac{2}{n}(1+\epsilon)]$ and $\Psi_i^{(2)}, \Psi_j^{(1)} \leq \frac{2}{n}\epsilon$. Letting $\Psi_i^{(2)} = \frac{2}{n}\delta_i$ and $\Psi_j^{(1)} = \frac{2}{n}\delta_j$, we have

$$\begin{aligned}
Q_{i,j}(\Psi) &= \lambda_S\langle\Psi_i, \Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)}) \\
&\leq \frac{4}{n^2}\left(\lambda_S((1+\epsilon)\delta_i + (1+\epsilon)\delta_j) + \lambda_X((1+\epsilon-\delta_i)(1+\epsilon-\delta_j) + \delta_i\delta_j\right) \\
&= \frac{4}{n^2}\left(\lambda_X(1+\epsilon)^2 + (\lambda_S - \lambda_X)((1+\epsilon)(\delta_i+\delta_j) - \delta_i\delta_j\right) \\
&\leq \frac{4}{n^2}\left(\lambda_X(1+\epsilon)^2 + 2(\lambda_S - \lambda_X)(1+\epsilon)\epsilon\right) \\
&= \frac{4\lambda_X}{n^2}(1+\epsilon)\left(1 + \frac{\epsilon(2\lambda_S - \lambda_X)}{\lambda_X}\right)
\end{aligned}$$

Combining the above two cases, we can bound on $H(\Psi)$ above by

$$\begin{aligned}
&\sum_{(i,j)\in E\setminus E(S,T)} \ln\left(\frac{4\lambda_S(1+\epsilon)^2}{n^2}\right) + \sum_{(i,j)\in E(S,T)} \ln\left(\frac{4\lambda_X}{n^2}(1+\epsilon)\left(1 + \frac{\epsilon(2\lambda_S-\lambda_X)}{\lambda_X}\right)\right) \\
&= m\ln\frac{4\lambda_S}{n^2} - |\text{cut}(\Psi)|\cdot\ln\frac{\lambda_S}{\lambda_X} + |\text{cut}(\Psi)|\ln\left((1+\epsilon)\left(1 + \frac{\epsilon(2\lambda_S-\lambda_X)}{\lambda_X}\right)\right) \\
&\quad + (m - |\text{cut}(\Psi)|)\ln((1+\epsilon)^2) \\
&\leq m\ln\frac{4\lambda_S}{n^2} - |\text{cut}(\Psi)|\cdot\ln\frac{\lambda_S}{\lambda_X} + m\epsilon\max\left(2, 1 + \frac{2\lambda_S - \lambda_X}{\lambda_X}\right).
\end{aligned}$$

Using the fact that $\lambda_S > \lambda_X$ gives us the lemma. ■

Let $\Delta, \epsilon, N_0, N_1, N > 0$ satisfy the relationship specified in Lemma 13. We will argue that for an appropriate, but polynomial setting, of these variables, any Δ-optimal solution must correspond to the minimum bisection.

Let Ψ be a Δ-optimal solution. By Lemma 13, Ψ must be ϵ-approximately canonical. As in the proof of Lemma 14, we can use Ψ to define a cut (S,T). For $\epsilon < 1/(2n)$, this cut is a bisection. Now let (S^*, T^*) be an optimal bisection and let Ψ^* be the solution corresponding to this. Then we can say

$$\begin{aligned}
\Delta &\geq \max_{\Psi'} F(\Psi') - F(\Psi) \geq F(\Psi^*) - F(\Psi) \\
&= G(\Psi^*) - G(\Psi) + H(\Psi^*) - H(\Psi) \geq H(\Psi^*) - H(\Psi).
\end{aligned}$$

Now by Lemma 14, we have

$$\begin{aligned}\Delta &\geq H(\Psi^*) - H(\Psi) \\ &\geq m \ln \frac{4\lambda_S}{n^2} - |\text{cut}(\Psi^*)| \cdot \ln \frac{\lambda_S}{\lambda_X} - m \ln \frac{4\lambda_S}{n^2} + |\text{cut}(\Psi)| \cdot \ln \frac{\lambda_S}{\lambda_X} - 2m\epsilon \frac{\lambda_S}{\lambda_X} \\ &= (|\text{cut}(\Psi)| - |\text{cut}(\Psi^*)|) \ln \left(\frac{\lambda_S}{\lambda_X}\right) - 2m\epsilon \frac{\lambda_S}{\lambda_X} \\ &\geq (|\text{cut}(\Psi)| - |\text{cut}(\Psi^*)|) \left(\frac{\lambda_S - \lambda_X}{\lambda_S}\right) - 2m\epsilon \frac{\lambda_S}{\lambda_X}.\end{aligned}$$

Thus, if $\Delta \leq \frac{1}{3}\left(\frac{\lambda_S - \lambda_X}{\lambda_S}\right)$ and $\epsilon \leq \frac{1}{6m}\left(\frac{\lambda_X}{\lambda_S}\right)\left(\frac{\lambda_S - \lambda_X}{\lambda_S}\right)$, then we must conclude that

$$|\text{cut}(\Psi)| = |\text{cut}(\Psi^*)|.$$

These settings of ϵ and Δ, give us

$$\begin{aligned}N_0 &= \frac{2}{\epsilon^2}\left(\Delta + m \ln \left(\frac{\lambda_S}{\lambda_X}\frac{n^2}{4}\right)\right) \\ &= 72m^2 \left(\frac{\lambda_S}{\lambda_X}\right)^2 \left(\frac{\lambda_S}{\lambda_S - \lambda_X}\right)^2 \left((m + 1/2) \ln \left(\frac{\lambda_S}{\lambda_X}\right) + m \ln \left(\frac{n^2}{4}\right)\right)\end{aligned}$$

and

$$\begin{aligned}N_1 &= \frac{1}{\epsilon}\left(\frac{m\lambda_S}{\lambda_X} + \frac{\Delta\lambda_S}{\lambda_S - \lambda_X}\right) \\ &= 6m \left(\frac{\lambda_S}{\lambda_X}\right)\left(\frac{\lambda_S}{\lambda_S - \lambda_X}\right)\left(\frac{m\lambda_S}{\lambda_X} + \frac{\lambda_S}{2(\lambda_S - \lambda_X)} \ln \left(\frac{\lambda_S}{\lambda_X}\right)\right).\end{aligned}$$

Thus, we have that TM-MLE(α) is NP-hard when $K = 2$ and there are exactly two words in each document.

Now suppose Ψ is a topic matrix within $\frac{\lambda_S - \lambda_X}{3\lambda_S}$ of optimal. Lemma 13 guarantees

$$\min_i \max\left\{\Psi_i^{(1)}, \Psi_i^{(2)}\right\} \geq \frac{2}{|V|}(1 - 2\epsilon) \geq \frac{2}{|V|}\left(1 - \frac{2}{6|E|}\left(\frac{\lambda_X}{\lambda_S}\right)\left(\frac{\lambda_S - \lambda_X}{\lambda_S}\right)\right) \geq \frac{1}{|V|}.$$

Thus, Ψ must be $1/|V|$-smooth.

Appendix C. Proofs from Section 3

C.1. Proof of Lemma 3

Lemma 15 *Pick any $\delta > 0$ and any $\theta \in \Theta$ within δ of the optimal MAP solution for Z, that is,*

$$\log q_Z(\theta) \geq \sup_{\theta' \in \Theta} \log q_Z(\theta') - \delta.$$

Then the log-likelihood of any $\theta' \in \Theta$ can exceed that of θ by at most

$$\log p(X|\theta') - \log p(X|\theta) \leq \frac{1}{k}\left(\delta + \log q_0(\theta) - \log q_0(\theta')\right).$$

Proof Note that since θ is within δ of the supremum of $\ln q_Z$, we have

$$-\delta \leq \ln q_Z(\theta) - \ln q_Z(\theta') = \ln \frac{q_0(\theta)p(X \mid \theta)^k}{q_0(\theta')p(X \mid \theta')^k} = \ln \frac{q_0(\theta)}{q_0(\theta')} - k \ln \frac{p(X \mid \theta')}{p(X \mid \theta)}.$$

Rearranging the above gives us

$$\ln p(X \mid \theta') - p(X \mid \theta) = \ln \frac{p(X \mid \theta')}{p(X \mid \theta)} \leq \frac{1}{k}\left(\delta + \ln \frac{q_0(\theta)}{q_0(\theta')}\right).$$

■

C.2. Proof of Lemma 4

The goal of this section is to prove the following lemma.

Lemma 16 *Pick $c, m > 0$ and define $\alpha_0 = \sum \alpha_i$. Let $\mathcal{X}$ be the space of documents with length bounded by m and S be the space of c-smooth matrices. If*

$$\lambda = \left(\frac{2m}{c} + \max\left(1, \left(\frac{\alpha_0 + m}{K}\right)^K\right) \frac{K^{\alpha_0 + 2m - 1/2}}{c^m}\right),$$

then max-norm distance is (λ, S)-admissible.

To do so, we need to introduce some notation. Suppose $x = (i_1, \ldots, i_m)$ is some length m document. Then $z \in [K]^m$ is a *labeling* of x, that is an assignment of each word in x to some topic. For some fixed labeling z, let $n_i(z) = |\{j : z_j = i\}|$ denote the number of times that topic i appears in z. Define the likelihood of z under Ψ by

$$q(\Psi, z) = \left(\prod_{i=1}^{K} \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}\right) \prod_{j=1}^{m} \Psi_{i_j}^{(z_j)}.$$

Then we see that summing over all labelings gives us the likelihood of document x.

Lemma 17 *For any length m document x and any topic matrix Ψ,*

$$p(x|\Psi) = \sum_{z \in [K]^m} q(\Psi, z).$$

Proof To generate document $x = (i_1, \ldots, i_m)$ given Ψ, we can first sample $\theta \sim Dir(\alpha_1, \ldots, \alpha_K)$. Given θ, we can sample $z_1, \ldots, z_m$ independently from the distribution θ and then independently sample each word j from the distribution Ψ^{z_j}. Marginalizing over θ and z and recognizing that x is

independent of θ given the z's,

$$\begin{aligned} p(x \mid \Psi) &= \mathbb{E}_\theta[p(x \mid \Psi, \theta)] \\ &= \sum_{z \in [K]^m} \mathbb{E}_\theta \left[p(z|\theta) p(x|\Psi, \theta, z) \right] \\ &= \sum_{z \in [K]^m} \mathbb{E}_\theta \left[\prod_{j=1}^m \theta_{z_j} \right] \prod_{j=1}^m \Psi_{i_j}^{(z_j)} \\ &= \sum_{z \in [K]^m} \mathbb{E}_\theta \left[\prod_{i=1}^K \theta_i^{n_i(z)} \right] \prod_{j=1}^m \Psi_{i_j}^{(z_j)} \end{aligned}$$

The expectation in the last line deals with the moments of the Dirichlet distribution. Ng et al. (2011) provides the following identity for the moments of the Dirichlet distribution

$$\mathbb{E}_\theta \left[\prod_{i=1}^k \theta_i^{n_i} \right] = \frac{\Gamma(\sum \alpha_i)}{\Gamma(\sum \alpha_i + n_i)} \cdot \prod_{i=1}^k \frac{\Gamma(\alpha_i + n_i)}{\Gamma(\alpha_i)}$$

for positive integers $n_1, \ldots, n_K$. Plugging this into the above gives us the lemma. ■

Therefore, proving Lemma 4 amounts to getting a handle on the ratio

$$\frac{p(x \mid \Psi)}{p(x \mid \Phi)} = \frac{\sum_{z \in [K]^n} q(\Psi, z)}{\sum_{z \in [K]^n} q(\Phi, z)},$$

for topic matrices Φ, Ψ that are close in max-norm distance. The next few technical lemmas deal with bounding ratios of sums.

C.2.1. RATIOS OF SUMS

Lemma 18 *Let $a_1, \ldots, a_n, b_1, \ldots, b_n, c > 0$ such that $a_i / b_i \leq c$, then $\frac{\sum a_i}{\sum b_i} \leq c$.*

Proof We have that $a_i \leq c b_i$ for all i. Thus, $\frac{\sum a_i}{\sum b_i} \leq \frac{\sum c b_i}{\sum b_i} \leq c$. ■

Lemma 19 *Suppose $a, b, c, d, \epsilon > 0$, $x, y \in [0, 1]$, and $|x - y| \leq \epsilon$ then*

$$\frac{a + cx}{b + dy} \leq \max \left(\frac{a + c\epsilon}{b}, \frac{a + c}{b + d(1 - \epsilon)} \right).$$

Proof There are two cases.

Case 1: $y \geq 1 - \epsilon$. In this case we have

$$\frac{a + cx}{b + dy} \leq \frac{a + c}{b + d(1 - \epsilon)}.$$

Case 2: $y \leq 1 - \epsilon$. In this case we have

$$\frac{a + cx}{b + dy} \leq \frac{a + c(y + \epsilon)}{b + dy} =: f(y)$$

Then it can be shown that the sign of f' is independent of y (since $y \geq 0$). Therefore f is monotonic in y and reaches the maximum at the boundary $\{0, 1-\epsilon\}$. ■

Lemma 20 *Let $a, b, c, \epsilon > 0$ and $x_1, \ldots, x_n, y_1, \ldots, y_n \in [0,1]$ such that $|x_i - y_i| < \epsilon$. Then*

$$\frac{a + c\prod_{i=1}^n x_i}{b + c\prod_{i=1}^n y_i} \leq \max\left(\frac{a+\epsilon c}{b}, \frac{a+c}{b+(1-\epsilon)^n c}\right).$$

Proof The proof is by induction on n. The base case is simply an appeal to Lemma 19. Now assume it holds for $n-1$. There are three cases we need to consider.

Case 1: $y_n = 0$. In this case we know $x_n \leq \epsilon$, therefore

$$\frac{a + c\prod_{i=1}^n x_i}{b + c\prod_{i=1}^n y_i} \leq \frac{a + \epsilon c\prod_{i=1}^{n-1} x_i}{b} \leq \frac{a+\epsilon c}{b}.$$

Case 2: $x_n = 0$. In this case,

$$\frac{a + c\prod_{i=1}^n x_i}{b + c\prod_{i=1}^n y_i} \leq \frac{a}{b} \leq \frac{a+\epsilon c}{b}.$$

Case 3: $x_n, y_n > 0$. In this case we can use our inductive assumption to see the following

$$\begin{aligned}
\frac{a + c\prod_{i=1}^n x_i}{b + c\prod_{i=1}^n y_i} &= \frac{x_n}{y_n} \cdot \frac{a/x_n + c\prod_{i=1}^{n-1} x_i}{b/y_n + c\prod_{i=1}^{n-1} y_i} \\
&\leq \frac{x_n}{y_n} \max\left(\frac{a/x_n + \epsilon c}{b/y_n}, \frac{a/x_n + c}{b/y_n + (1-\epsilon)^{n-1}c}\right) \\
&= \max\left(\frac{a + \epsilon x_n c}{b}, \frac{a + x_n c}{b + y_n(1-\epsilon)^{n-1}c}\right) \\
&\leq \max\left(\frac{a+\epsilon c}{b}, \frac{a + x_n c}{b + y_n(1-\epsilon)^{n-1}c}\right).
\end{aligned}$$

By appealing again to Lemma 19, we have

$$\frac{a + x_n c}{b + y_n(1-\epsilon)^{n-1}c} \leq \max\left(\frac{a+\epsilon c}{b}, \frac{a+c}{b+(1-\epsilon)^n c}\right).$$

Combining all of the above gives us the lemma. ■

Lemma 21 *Let $a, b, c_i, \epsilon > 0$ and $x_{i,j}, y_{i,j} \in [0,1]$ such that $|x_{i,j} - y_{i,j}| \leq \epsilon$ for $i \in [m], j \in [n]$. Then there exists a partition Ω_1, Ω_2 of $[m]$ such that*

$$\frac{a + \sum_{i=1}^m c_i \prod_{j=1}^n x_{i,j}}{b + \sum_{i=1}^m c_i \prod_{j=1}^n y_{i,j}} \leq \frac{a + \sum_{i\in\Omega_1} \epsilon c_i + \sum_{i\in\Omega_2} c_i}{b + \sum_{i\in\Omega_2}(1-\epsilon)^n c_i}.$$

Proof We prove by induction on m. The base case of $m = 1$ follows directly from Lemma 20. We can assume that the lemma holds for $m - 1$, then

$$\frac{a + \sum_{i=1}^{m} c_i \prod_{j=1}^{n} x_{i,j}}{b + \sum_{i=1}^{m} c_i \prod_{j=1}^{n} y_{i,j}} = \frac{\overbrace{a + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} x_{i,j}}^{a'} + c_m \prod_{j=1}^{n} x_{m,j}}{\underbrace{b + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} y_{i,j}}_{b'} + c_m \prod_{j=1}^{n} y_{m,j}}.$$

By applying Lemma 20, we have that this is bounded by

$$\max\left(\frac{a' + \epsilon c_m}{b'}, \frac{a' + c_m}{b' + (1-\epsilon)^n c_m}\right).$$

We will bound each of these quantities separately. Denoting $a_1 = a + \epsilon c_m$, then by induction we have that there exists a partition Ω_1', Ω_2' of $[m-1]$ such that

$$\begin{aligned}
\frac{a' + \epsilon c_m}{b'} &= \frac{a_1 + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} x_{i,j}}{b + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} y_{i,j}} \\
&\leq \frac{a_1 + \sum_{i \in \Omega_1'} \epsilon c_i + \sum_{i \in \Omega_2'} c_i}{b + \sum_{i \in \Omega_2'} (1-\epsilon)^n c_i} \\
&= \frac{a + \sum_{i \in \Omega_1' \cup \{m\}} \epsilon c_i + \sum_{i \in \Omega_2'} c_i}{b + \sum_{i \in \Omega_2'} (1-\epsilon)^n c_i}.
\end{aligned}$$

On the other hand, if we let $a_2 = a + c_m$ and $b_2 = b + (1-\epsilon)^n c_m$, then by induction there exists a partition Ω_1'', Ω_2'' of $[m-1]$ such that

$$\begin{aligned}
\frac{a' + c_m}{b' + (1-\epsilon)^n c_m} &= \frac{a_2 + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} x_{i,j}}{b_2 + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} y_{i,j}} \\
&\leq \frac{a_2 + \sum_{i \in \Omega_1''} \epsilon c_i + \sum_{i \in \Omega_2''} c_i}{b_2 + \sum_{i \in \Omega_2''} (1-\epsilon)^n c_i} \\
&= \frac{a + \sum_{i \in \Omega_1''} \epsilon c_i + \sum_{i \in \Omega_2'' \cup \{m\}} c_i}{b + \sum_{i \in \Omega_2'' \cup \{m\}} (1-\epsilon)^n c_i}.
\end{aligned}$$

By taking Ω_1, Ω_2 to be the partitions corresponding to the larger of these two scenarios (either $\Omega_1' \cup \{m\}, \Omega_2'$ or $\Omega_1'', \Omega_2'' \cup \{m\}$), we have the lemma statement. ■

C.2.2. ACTUAL PROOF OF LEMMA 4

We are now ready to prove the main lemma of this section.

Lemma 22 *Pick $c, m > 0$ and define $\alpha_0 = \sum \alpha_i$. Let $\mathcal{X}$ be the space of documents with length bounded by m and S be the space of c-smooth matrices. If*

$$\lambda = \left(\frac{2m}{c} + \max\left(1, \left(\frac{\alpha_0 + m}{K}\right)^K\right) \frac{K^{\alpha_0 + 2m - 1/2}}{c^m} \right),$$

then max-norm distance is (λ, S)-admissible.

Proof Let $x = (i_1, \ldots, i_m)$ be a document of length m, and let $\Omega = [K]^m$ denote the space of all labelings. Pick $\Phi, \Psi \in S$. From the smoothness condition, we see that there is a labeling $z^* \in \Omega$ such that $\Phi_{i_j}^{(z_j^*)} \geq c$ for $j = 1, \ldots, m$. Recalling the definition of $q(\cdot, z)$ from Lemma 17 and applying Lemma 21, we know that we can partition $\Omega \setminus \{z^*\}$ into Ω_1, Ω_2 such that

$$\begin{aligned}
\frac{p(x \mid \Psi)}{p(x \mid \Phi)} &= \frac{q(\Psi, z^*) + \sum_{z \in \Omega \setminus \{z^*\}} q(\Psi, z)}{q(\Phi, z^*) + \sum_{z \in \Omega \setminus \{z^*\}} q(\Phi, z)} \\
&= \frac{q(\Psi, z^*) + \sum_{z \in \Omega \setminus \{z^*\}} \left(\prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}\right) \prod_{j=1}^m \Psi_{i_j}^{(z_j)}}{q(\Phi, z^*) + \sum_{z \in \Omega \setminus \{z^*\}} \left(\prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}\right) \prod_{j=1}^m \Phi_{i_j}^{(z_j)}} \\
&\leq \frac{q(\Psi, z^*) + \sum_{z \in \Omega_1} \epsilon \prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)} + \sum_{S \in \Omega_2} \prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}}{q(\Phi, z^*) + \sum_{z \in \Omega_2} (1-\epsilon)^m \prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}}.
\end{aligned}$$

From Lemma 18 we know that we can separately bound

$$\frac{q(\Psi, z^*) + \sum_{z \in \Omega_1} \epsilon \prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}}{q(\Phi, z^*)} \quad \text{and} \quad \frac{\sum_{z \in \Omega_2} \prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}}{\sum_{z \in \Omega_2} (1-\epsilon)^m \prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}}.$$

The second quantity is simply bounded above by $(1-\epsilon)^{-m} \leq \exp\left(\frac{\epsilon m}{1-\epsilon}\right) \leq \exp(2\epsilon m)$.

By properties of the gamma function, $\prod_{i=1}^K \Gamma(\alpha_i + r_i) \leq \Gamma(\alpha_0 + m)$ for any $r_1, \ldots, r_K \geq 0$ satisfying $r_1 + \cdots + r_K = m$. Since $\|\Phi - \Psi\|_{max} \leq \epsilon$ and $\Phi_{i_j}^{(z_j^*)} \geq c$ for all j, we have

$$\begin{aligned}
&\frac{q(\Psi, z^*) + \sum_{z \in \Omega_1} \epsilon \prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}}{q(\Phi, z^*)} \\
&= \frac{\left(\prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z^*))}{\Gamma(\alpha_i)}\right) \prod_{j=1}^m \Psi_{i_j}^{(z_j^*)} + \sum_{z \in \Omega_1} \epsilon \prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)}}{\left(\prod_{i=1}^K \frac{\Gamma(\alpha_i + n_i(z^*))}{\Gamma(\alpha_i)}\right) \prod_{j=1}^m \Phi_{i_j}^{(z_j^*)}} \\
&\leq \frac{\prod_{j=1}^m \Psi_{i_j}^{(z_j^*)}}{\prod_{j=1}^m \Phi_{i_j}^{(z_j^*)}} + \frac{\epsilon |\Omega_1| \Gamma(m + \alpha_0)}{\left(\prod_{j=1}^K \Gamma(\alpha_j + n_j(z^*))\right) \left(\prod_{j=1}^m \Phi_{i_j}^{(z_j^*)}\right)} \\
&\leq (1 + \epsilon/c)^m + \frac{\epsilon |\Omega_1| \Gamma(m + \alpha_0)}{c^m \prod_{j=1}^K \Gamma(\alpha_j + n_j(z^*))} \\
&\leq e^{\epsilon m / c} + \frac{\epsilon K^m \Gamma(m + \alpha_0)}{c^m \prod_{j=1}^K \Gamma(\alpha_j + n_j(z^*))}.
\end{aligned}$$

Where the last line follows by observing that $|\Omega_1| \leq |\Omega| = K^m$.

Additionally, by the log-convexity of Γ on the positive reals, we know that for positive $x_1, \ldots, x_K$, $\Gamma(x_1) \cdots \Gamma(x_K) \geq (\Gamma(x_1/K + \cdots x_K/K))^K$. Thus

$$\frac{p(x \mid \Psi)}{p(x \mid \Phi)} \;\leq\; e^{\epsilon m/c} + \frac{\epsilon K^m \Gamma(m + \alpha_0)}{c^m (\Gamma(\alpha_0/K + m/K))^K}.$$

Taking logs and making use of $\epsilon < c/m$, we have

$$\begin{aligned}
\ln \frac{p(x \mid \Psi)}{p(x \mid \Phi)} &\leq \ln \left(e^{\epsilon m/c} + \frac{\epsilon \Gamma(\alpha_0 + m) K^m}{c^m (\Gamma(\alpha_0/K + m/K))^K} \right) \\
&\leq \ln \left(1 + \frac{2m\epsilon}{c} + \frac{\epsilon \Gamma(\alpha_0 + m) K^m}{c^m (\Gamma(\alpha_0/K + m/K))^K} \right) \\
&\leq \epsilon \left(\frac{2m}{c} + \left(\frac{K}{c} \right)^m \frac{\Gamma(\alpha_0 + m)}{(\Gamma(\alpha_0/K + m/K))^K} \right)
\end{aligned}$$

By Gauss' multiplicative theorem and the log-convexity of Γ, we know for any positive integer k and any $a > 0$,

$$\frac{\Gamma(ka)}{\Gamma(a)^k} \;\leq\; \max(1, a^k) k^{ak - 1/2}.$$

Applying this to the above gives us the lemma statement. ■

C.3. Proof of Theorem 4

Define the TM-MLE(α, K, m) problem to be the TM-MLE(α) problem where the number of topics is K and the number of words per document is bounded from above by m. TM-MAP(α, β, K, m) and TM-APPROX-SAMPLING(α, β, K, m) are defined analogously.

Let Δ^V denote the simplex of all probability distributions over V outcomes. For every $c > 0$, define

$$S_c = \left\{ \Psi \in \Delta^{V \times K} \;:\; \Psi \text{ is } c\text{-smooth} \right\}.$$

If m is the length of the longest document, then we have by Lemma 4 that the max-norm is $(g(K, m, c, \alpha), S_c)$-admissible for

$$g(K, m, c, \alpha) = \frac{2m}{c} + \max \left(1, \left(\frac{\alpha_0 + m}{K} \right)^K \right) \frac{K^{\alpha_0 + 2m - 1/2}}{c^m}.$$

The next thing we need to establish to apply our results from Sections 3 and 4 is that the prior distribution is well-behaved on neighborhoods of the maximum likelihood estimate. The following lemma gives us a handle on the Dirichlet distribution.

Lemma 23 *Suppose that ν is the measure and q is the density associated with the symmetric Dirichlet distribution over Δ^N with parameter α. Then for any $\epsilon > 0$ and any point $x \in \Delta^N$ s.t. $\min_i x_i \geq \epsilon$ we have*

$$\log q(x) \;\geq\; -\operatorname{poly}(N, \alpha, 1/\alpha, 1/\epsilon)$$

which implies for any $x \in \Delta^N$

$$\log \nu(B_{\ell_2}(x, \epsilon)) \;\geq\; -\operatorname{poly}(N, \alpha, 1/\alpha, 1/\epsilon).$$

Further, if $\alpha \geq 1$*, we have*

$$\log q(x) \;\leq\; \operatorname{poly}(N, \alpha).$$

Proof Recall that

$$q(x) \;=\; \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N}\, x_1^{\alpha-1} \cdots x_N^{\alpha-1}.$$

We will first show that if $\alpha \geq 1$, then $\log q(x) \;\leq\; \operatorname{poly}(N, \alpha)$. Note that when $\alpha \geq 1$, q is a concave density with whose maximum is achieved at $(1/N, \ldots, 1/N)$. Thus,

$$q(x) \leq \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N} \cdot N^{N(1-\alpha)} \leq (N\alpha)^{N\alpha} \cdot 2^N \cdot N^{N(1-\alpha)} \leq 2^{\operatorname{poly}(N,\alpha)}$$

where the last inequality follows from the bounds $\Gamma(x) \leq x^x$ and $\Gamma(x) \geq 1/2$ for $x \geq 1$.

Now we turn to showing the first two inequalities. We consider two cases.

Case 1: $\alpha < 1$**.** In this case q is a convex probability density with minimum at $(1/N, \ldots, 1/N)$. Thus,

$$q(x) \geq \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N} \cdot \left(\frac{1}{N}\right)^{N(\alpha-1)} \geq \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N}.$$

Notice by Γ's recurrence relation

$$\Gamma(\alpha) = \frac{\Gamma(1+\alpha)}{\alpha} \leq \frac{1}{\alpha}$$

for $\alpha \in (0, 1)$. Moreover, $\Gamma(t) \geq 3/4$ for any real $t > 0$. Thus, we have

$$q(x) \geq \frac{3}{4}\left(\frac{\alpha}{N}\right)^N \geq 2^{-\operatorname{poly}(N, 1/\alpha)}.$$

Case 2: $\alpha \geq 1$**.** When $x_i \geq \epsilon$ for $i = 1, \ldots, n$, we have

$$q(x) \geq \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N} \epsilon^{N(\alpha-1)} \geq 2^{-\operatorname{poly}(N,\alpha,1/\epsilon)}.$$

Then the inequalities dealing with the density q in the lemma statement can be gleaned from the above two cases.

Now we turn to lower bounding $\nu(B_{\ell_2}(x, \epsilon) \cap \Delta^N)$. First, note that $\operatorname{vol}(B_{\ell_2}(x, \epsilon) \cap \Delta^N)$ is minimized for $x \in \Delta^N$ when x is a corner of Δ^n. Thus we can consider $x \in \Delta^n$ such that w.l.o.g. $x_1 = 1$ and $x_i = 0$ for $i = 2, \ldots, N$. We claim $B_{\ell_2}(x, \epsilon) \cap \Delta^N$ contains a regular simplex S with edge length $\epsilon/2N$ satisfying that $\min_i x_i \geq \epsilon/2N$ for all $x \in S$. To see this, let S be the simplex created by the convex hull of $x^{(1)}, \ldots, x^{(N)} \in \Delta^N$ where

$$x_i^{(1)} = \begin{cases} 1 - \frac{(N-1)\epsilon}{2N} & \text{if } i = 1 \\ \frac{\epsilon}{2N} & \text{o/w} \end{cases}$$

and

$$x_i^{(k)} = \begin{cases} 1 - \frac{\epsilon}{2N}\left(N - 2 + \frac{1+\sqrt{2}}{\sqrt{2}}\right) & \text{if } i = 1 \\ \frac{(1+\sqrt{2})\epsilon}{2\sqrt{2}N} & \text{if } i = k \\ \frac{\epsilon}{2N} & \text{o/w} \end{cases}$$

for $k = 2, \ldots, N$. Then one can see that

- $x^{(k)} \in \Delta^N$ for $k = 1, \ldots, N$,
- $\|x^{(k)} - x^{(k')}\| = \frac{\epsilon}{2N}$ for all $k \neq k'$,
- $x^{(k)} \in B_{\ell_2}(x, \epsilon)$ for $k = 1, \ldots, N$, and
- $x_i^{(k)} \geq \frac{\epsilon}{2N}$ for $i, k = 1, \ldots, N$.

Then the simplex S lying in the convex hull of $x^{(1)}, \ldots, x^{(N)}$ is a regular simplex with edge length $\epsilon/2N$ satisfying that $\min_i x_i \geq \epsilon/2N$ for all $x \in S$. Therefore for any $x \in \Delta^N$,

$$\nu(B_{\ell_2}(x, \epsilon) \cap \Delta^N) \geq \text{vol}(S) \cdot \inf_{x \in S} q(x) = \frac{\sqrt{N+1}}{N! 2^{N/2}} \cdot \left(\frac{\epsilon}{2N}\right)^N \cdot \inf_{x \in S} q(x) \geq 2^{-\text{poly}(N, \alpha, 1/\alpha, 1/\epsilon)}.$$

■

We are now ready to apply Theorem 3.

Theorem 12 *Let $\alpha > 0$, $c = 1/V$, $\beta \geq 1$, $K, m \in \mathbb{N}$, and let Π_c denote the promise that $\Psi_{ml} \in S_c$, then Π_c-TM-MLE$(\alpha, K, m) \leq_P$ TM-MAP(α, β, K, m) where the reduction is polynomial in the input size and $(1/c)^m$, K^m, and $\max\{\beta, 1/\beta\}$.*

Proof Suppose that q is the density associated with the symmetric Dirichlet distribution over Δ^V with parameter β. The prior density q_0 we are interested in is the product distribution, i.e. for $\Psi \in \mathbb{R}^{V \times K}$,

$$q_0(\Psi) = q(\Psi^{(1)}) \cdots q(\Psi^{(K)}).$$

From Lemma 23, we know that q is bounded above by $2^{\text{poly}(V, \beta)}$. The density q_0 of the product distribution is thus bounded above by $2^{\text{poly}(V, K, \beta)}$.

From Lemma 4, we know that max-norm is $(g(K, m, c, \alpha), S_c)$-admissible. Thus, in order to apply Theorem 3, we need to show the existence of a topic matrix $\widehat{\Psi}$ satisfying the following three conditions. For small enough $\epsilon > 0$,

(a) $\widehat{\Psi} \in S_c$,

(b) $\|\widehat{\Psi} - \Psi_{ml}\|_{max} \leq \epsilon$, and

(c) $q_0(\widehat{\Psi}) \geq 2^{-\text{poly}(V, K, \beta, 1/\epsilon)}$

To construct such a $\widehat{\Psi}$, let us first denote $\Psi = \Psi_{ml}$ and let $s = \min(\epsilon, 1/V^2)$. Consider a particular column j. If it is the case that $\Psi_i^{(j)} \geq s$ for all rows i, then we take $\widehat{\Psi}^{(j)} = \Psi^{(j)}$. Otherwise, because $\Psi^{(j)}$ is a distribution over V words and sums to one, this implies that there exists a row i^* such that $\Psi_{i^*}^{(j)} \geq \frac{1}{V} + \frac{s}{V}$. Then we take

$$\widehat{\Psi}_i^{(j)} = \begin{cases} \Psi_i^{(j)} - \frac{s}{V} & \text{if } i = i^* \\ \Psi_i^{(j)} + \frac{s}{V(V-1)} & \text{otherwise} \end{cases}$$

Then $\widehat{\Psi}$ is a valid topic matrix. It is easy to check that it satisfies (a) and (b). To see (c), notice that $\widehat{\Psi}_i^{(j)} \geq \frac{s}{V(V-1)}$ for all i, j. By Lemma 23, this implies that every column j satisfies $q(\widehat{\Psi}^{(j)}) \geq 2^{-\operatorname{poly}(V,\beta,1/\epsilon)}$, which implies $q_0(\Psi) \geq 2^{-\operatorname{poly}(V,K,\beta,1/\epsilon)}$. ■

The ML estimate in the construction in Theorem 1 lies in S_c for $c = 1/V$. The construction also satisfies that $K = 2$ and $m = 2$ (and that α is a constant), which means that the dominating factor $g(K, m, c, \alpha)$ is bounded above by $\operatorname{poly}(V)$. Theorem 4 follows as an immediate corollary.

Theorem 4 *For any fixed $\alpha > 0$ and $\beta \geq 1$,* TM-MAP(α, β) *is NP-hard.*

Appendix D. Proofs from Section 4

Theorem 5 *Let $\epsilon > 0$ and ν be a distribution over a metric space (Θ, d). If $\widehat{\Theta}$ is a countable ϵ-cover of Θ then there exists a measure $\widehat{\nu}$ over $\widehat{\Theta}$ such that $W_t(\nu, \widehat{\nu}) \leq \epsilon$.*

Proof For every $\widehat{\theta} \in \widehat{\Theta}$, define the inner Voronoi cell of $\widehat{\theta}$ to be

$$C^i(\widehat{\theta}) := \{\theta \,:\, d(\theta, \widehat{\theta}) < d(\theta, \bar{\theta}) \;\; \forall \bar{\theta} \in \widehat{\Theta} \setminus \{\widehat{\theta}\}\}.$$

The Voronoi cell $C(\widehat{\theta})$ consists of $C^i(\widehat{\theta})$ as well as part of its boundary. To ensure that these cells are disjoint and cover all of Θ, we can order $\widehat{\Theta}$ and adopt the convention that the boundary occurring among any Voronoi cells belongs to the cell whose center comes earliest in the ordering.

Now take $\widehat{\nu}$ to be the distribution over $\widehat{\Theta}$ such that $\widehat{\nu}(\widehat{\theta}) = \nu(C(\widehat{\theta}))$. We will show that $\mathcal{W}_t(\nu, \widehat{\nu}) \leq \epsilon$. To see this, consider the following coupling (X, Y) of ν and $\widehat{\nu}$:

- Draw $X \sim \nu$.
- There exists a $\widehat{\theta} \in \widehat{\Theta}$ such that $X \in C(\widehat{\theta})$.
- Take $Y = \widehat{\theta}$.

It is not hard to see that the marginal distributions of X and Y are ν and $\widehat{\nu}$, respectively, making this a valid coupling. Moreover, since $\widehat{\Theta}$ is an ϵ-cover of Θ, we have $d(X, Y) \leq \epsilon$ with probability 1. Thus,

$$\mathcal{W}_t(\nu, \widehat{\nu}) \leq \mathbb{E}[d(X, Y)^t]^{1/t} \leq \epsilon.$$

■

Lemma 24 *Take any $\epsilon, \delta > 0$ and $X \in \mathcal{X}^n$. If Z is the sequence created by duplicating X*

$$k \;\geq\; \frac{2}{\epsilon}\left(\log\left(\frac{1}{\delta}-1\right)+\log\left(\frac{1-\nu_0(B_{d_{p,X}}(\theta_{ml},\epsilon))}{\nu_0(B_{d_{p,X}}(\theta_{ml},\epsilon/2))}\right)\right)$$

times then $\nu_Z(B_{d_{p,X}}(\theta_{ml},\epsilon)) \geq 1-\delta$. (Recall θ_{ml} is the maximum-likelihood solution for X.)

Proof For any measurable set B, we may write

$$\nu_Z(B) \;=\; \frac{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}[\theta\in B]p(X|\theta)^k]}{\mathbb{E}_{\theta\sim\nu_0}[p(X|\theta)^k]}.$$

Thus,

$$\begin{aligned}
\frac{\nu_Z(B_{d_{p,X}}(\theta_{ml},\epsilon))}{\nu_Z(\Theta\setminus B_{d_{p,X}}(\theta_{ml},\epsilon))} &= \frac{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta\in B_{d_{p,X}}(\theta_{ml},\epsilon))p(X|\theta)^k]}{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta\notin B_{d_{p,X}}(\theta_{ml},\epsilon))p(X|\theta)^k]}\\
&\geq \frac{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta\in B_{d_{p,X}}(\theta_{ml},\epsilon/2))(e^{-\epsilon/2}p(X|\theta_{ml}))^k]}{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta\notin B_{d_{p,X}}(\theta_{ml},\epsilon))(e^{-\epsilon}p(X|\theta_{ml}))^k]}\\
&\geq e^{k\epsilon/2}\frac{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta\in B_{d_{p,X}}(\theta_{ml},\epsilon/2))]}{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta\notin B_{d_{p,X}}(\theta_{ml},\epsilon))}\\
&= e^{k\epsilon/2}\frac{\nu_0(B_{d_{p,X}}(\theta_{ml},\epsilon/2))}{\nu_0(\Theta\setminus B_{d_{p,X}}(\theta_{ml},\epsilon))}
\end{aligned}$$

Note that if the above is greater than $1/\delta-1$, we have

$$\nu_Z(B_{d_{p,X}}(\theta_{ml},\epsilon)) \;=\; \frac{\nu_Z(B_{d_{p,X}}(\theta_{ml},\epsilon))}{\nu_Z(B_{d_{p,X}}(\theta_{ml},\epsilon))+\nu_Z(\Theta\setminus B_{d_{p,X}}(\theta_{ml},\epsilon))} \;\geq\; 1-\delta.$$

However, this condition is satisfied when

$$k \;\geq\; \frac{2}{\epsilon}\left(\log\left(\frac{1}{\delta}-1\right)+\log\left(\frac{\nu_0(\Theta\setminus B_{d_{p,X}}(\theta_{ml},\epsilon))}{\nu_0(B_{d_{p,X}}(\theta_{ml},\epsilon/2))}\right)\right).$$

■

D.1. Proof of Theorem 7

Theorem 6 *Let m be some measure of the size of an input instance X, and let $\lambda(m)$ be any function of this size. Let d be a distance function and $S \subset S' \subset \Theta$ be subsets satisfying*

(i) if $\theta \in S$ then $B_d(\theta, 1/\lambda(m)) \subset S'$ and

(ii) d is $(\lambda(m), S')$-admissible.

If Π is the promise that $B_{d_{p,X}}(\theta_{ml}, 1/\lambda(m)) \subset S$ and $\nu_0(B_d(\theta_{ml},\epsilon)) \geq 2^{-\mathrm{poly}(\lambda(m),1/\epsilon)}$ for all $\epsilon > 0$, then Π-MLE-$(p,\Theta) \leq_P \mathcal{W}_t$-APPROX-SAMPLING-$(p,\nu_0,\Theta)$ under randomized reductions which are polynomial in the input size and $\lambda(m)$.

Proof Let $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and b be input to Π-MLE-(p, Θ). We may assume that $b \geq \lambda(m)$. If not, we can replace b with $\lambda(m)$. Further, we may assume Π is true of X, since we can return anything and terminate if it is not.

Pick any $\delta > 0$. By Lemma 5, duplicating the data

$$\begin{aligned} k(b, \delta) &= 4b \log\left(\frac{1}{\delta} - 1\right) - \log \nu_0 \left(B_d\left(\theta_{ml}, \frac{1}{4bn\lambda(m)}\right)\right) \\ &\geq 4b \log\left(\frac{1}{\delta} - 1\right) - \log \nu_0 \left(B_{d_{p,X}}\left(\theta_{ml}, \frac{1}{2b}\right)\right) \end{aligned}$$

times will ensure that $\nu_Z(B_{d_{p,X}}(\theta_{ml}, 1/(2b))) \geq 1 - \delta$ for $Z = (X^{(1)}, \ldots, X^{(k(b,\delta))})$.

Set $b' = 4b\lambda(m)n/\delta^{1/t}$ to be the accuracy parameter for $\mathcal{W}_t$-APPROX-SAMPLING-(p, ν_0, Θ). Let $\widehat{\nu}$ be the distribution the sample is generated from. For any $\alpha > 0$, we have from the definition of Wasserstein distance that there is some coupling $\gamma \in \Gamma(\widehat{\nu}, \nu_Z)$ satisfying

$$\left(\mathbb{E}_{(\theta,\theta') \sim \gamma}[d(\theta, \theta')^t]\right)^{1/t} \;\leq\; 1/b' + \alpha.$$

Letting $(\theta, \theta') \sim \gamma$ and $\alpha = 1/b'$, we have

$$\begin{aligned} \Pr\left(d_{p,X}(\theta, \theta_{ml}) \geq \frac{1}{b}\right) &\leq \Pr\left(d_{p,X}(\theta, \theta_{ml}) \geq \frac{1}{b}, d_{p,X}(\theta', \theta_{ml}) \leq \frac{1}{2b}\right) + \Pr\left(d_{p,X}(\theta', \theta_{ml}) > \frac{1}{2b}\right) \\ &\leq \Pr\left(d_{p,X}(\theta, \theta') \geq \frac{1}{2b}, d_{p,X}(\theta', \theta_{ml}) \leq \frac{1}{2b}\right) + \delta \\ &\overset{(1)}{\leq} \Pr\left(d(\theta, \theta') \geq \frac{1}{2b\lambda(m)n}, d_{p,X}(\theta', \theta_{ml}) \leq \frac{1}{2b}\right) + \delta \\ &\leq \Pr\left(d(\theta, \theta') \geq \frac{1}{2b\lambda(m)n}\right) + \delta \\ &\overset{(2)}{\leq} (2b\lambda(m)n)^t \, \mathbb{E}\left[d(\theta, \theta')^t\right] + \delta \\ &\overset{(3)}{\leq} \left(4b\lambda(m)n/b'\right)^t + \delta \;\leq\; 2\delta \end{aligned}$$

where (1) follows from the fact that, if $\theta' \in B_{d_{p,X}}(\theta_{ml}, 1/(2b)) \subset S$ and $d(\theta, \theta') < 1/(2b\lambda(m)n)$, then $\theta \in S'$ and $d_{p,X}(\theta, \theta') < 1/(2b)$; (2) is Markov's inequality; and (3) follows from our choice of b'. ■

Much of the proof of Theorem 7 is similar to the proof of Theorem 4. One key difference is that we care about lower bounding the probability mass of balls with respect to the Dirichlet(β) distribution. Because the ℓ_∞ and ℓ_2 norms are related by a factor which is polynomial in the dimension, Lemma 23 also implies that

$$\log \nu(B_{\ell_\infty}(x, \epsilon)) \;\geq\; -\operatorname{poly}(N, \beta, 1/\beta, 1/\epsilon)$$

for any $x \in \Delta^N$.

Theorem 7 *There is no poly-time algorithm for* TM-APPROX-SAMPLING(α, β) *for any* $\alpha, \beta > 0$ *unless NP=RP.*

Proof We will reduce from an instance of TM-MLE(α) from Theorem 1. In order to apply Theorem 6, let S be the set of all $1/V$-smooth matrices, S' be the set of all $1/(2V)$-smooth matrices, and let d be the max-norm distance.Then

(ii) if $\Psi \in S$ then $B_d(\Psi, 1/(2V)) \subset S'$ (max-norm distance),

(ii) d is $(\text{poly}(V), S')$-admissible (Lemma 4, $K = m = 2$, and α is a constant)

(iii) $B_{d_{p,X}}(\Psi_{ml}, 1/(3(1+\alpha))) \subset S$ (a promise from Theorem 1), and

(iv) for all $\epsilon > 0$ and all $\Psi \in S$, $\nu_0(B_d(\Psi, \epsilon)) \geq 2^{-\text{poly}(V,K,\alpha,1/\alpha,1/\epsilon)}$ (Lemma 23).

Thus, Theorem 6 implies that TM-MLE(α) $\leq_P$ TM-APPROX-SAMPLING(α, β). From Theorem 1, we know that there are no poly-time algorithms for TM-APPROX-SAMPLING(α, β) unless $NP = RP$. ■

Appendix E. Proofs from Section 5

E.1. Proof of Theorem 8

To prove Theorem 8, we will reduce from the following problem.

> k-MEANS
> *Input*: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; positive integer k.
> *Output*: A collection of k "centers" $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ in $\mathbb{R}^d$ that minimize the cost function
> $$\Phi(\boldsymbol{\mu}) = \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|x_i - \mu_j\|^2.$$

Let Π' denote the promise that there are low-order polynomials $\alpha(\cdot)$ and $\beta(\cdot)$ such that

- For an instance containing n points, each point is unique and has dimension at most $\alpha(n)$, with individual coordinates taking values in $\{-1, 0, 1\}$.
- Any set of means with k-means cost within a multiplicative factor $1 + 1/\beta(n)$ of optimal induces an optimal k-means partition of the data.

The following was shown by Aloise et al. (2009).

Theorem 13 (Aloise et al. (2009)) Π'-k-MEANS *is NP-hard.*

Tosh and Dasgupta (2018) demonstrated that there is a simple reduction from Π'-k-MEANS to MOGS-SV.

Lemma 25 (Tosh and Dasgupta (2018)) *Let $x_1, \ldots, x_n \in \mathbb{R}^{d_o}$ be an instance of Π'-k-MEANS and suppose we pad the points with 0's until the dimension reaches*

$$d \geq \max\{16\beta(n) \ln k, 2n\alpha(n)\sqrt{1 + 2\ln k}\}$$

If $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ satisfies $LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \leq 1$ then

$$\Phi(\boldsymbol{\mu}) \leq \left(1 + \frac{1}{\beta(n)}\right) \Phi_{OPT}.$$

To prove our promise problem is NP-hard, we will utilize two results from Tosh and Dasgupta (2018). The first relates the log-likelihood of a mixture model to the costs of certain partitions.

Lemma 26 (Tosh and Dasgupta (2018)) *Pick any mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ and data set $\mathcal{X} = \{x_1, \ldots, x_n\}$.*

(a) For any partition $(\mathcal{X}'_1, \ldots, \mathcal{X}'_k)$ of $\mathcal{X}$, we have

$$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \;\geq\; \sum_{j=1}^{k} \sum_{x \in \mathcal{X}'_j} \ln(\pi_j N(x; \mu_j, \sigma^2)).$$

(b) Let $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ correspond to the partition

$$\mathcal{X}_j \;=\; \left\{ x \in \mathcal{X} \;:\; j = \operatorname*{argmax}_{\ell} \; \pi_\ell N(x; \mu_\ell, \sigma^2) \right\}$$

(breaking ties arbitrarily). Then

$$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \;\leq\; n \ln k + \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \ln(\pi_j N(x; \mu_j, \sigma^2)).$$

The second result shows that k-means cost of the means of a mixture can be related to its log-likelihood.

Lemma 27 (Tosh and Dasgupta (2018)) *Fix any data set $x_1, \ldots, x_n \in \mathbb{R}^d$ and any positive integer k. Let LL_{OPT} denote the log-likelihood of the optimal solution to* MOG-SV*, and Φ_{OPT} the lowest achievable k-means cost. For any parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$, we have*

$$\ln \frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}} \;\leq\; \frac{4 \ln k}{d} + \frac{2}{nd} \left(LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \right).$$

Given the above, we are now ready to prove the main result of this section.

Theorem 8 *Let Π be the promise that there exists a low-order polynomial $\rho(\cdot, \cdot, \cdot)$ such that all input data points satisfy $\|x\| \leq \rho(n, d, k)$ and if $\theta_{ml} = (\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*)$ is a maximum likelihood solution and $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ is within 1 of optimality, then*

(i) $\|\mu_j\| \leq \rho(n, d, k)$ *for all j,*

(ii) $\sigma^2 \geq 1/\rho(n, d, k)$*,*

(iii) $\pi_j > 0$ *for all j, and*

(iv) $\pi^*_j \geq 1/\rho(n, d, k)$ *for all j.*

Then Π-MLE-MOGS-SV(k) *is NP-hard for $k \geq 2$.*

Proof We will reduce from Π'-k-MEANS. Our reduction is to pad the points in the k-means instance with zeros until the dimension d satisfies

$$d \;\geq\; \max\{16\beta(n) \ln k, 2n\alpha(n)\sqrt{1 + 2 \ln k}\}$$

and solve the resulting MOGS-SV with $b = 1$. From Lemma 27 and the promise Π', this solves the original problem. Note that since we are reducing from Π'-k-MEANS and we are padding the data points with 0's, all the resulting points satisfy $x_i \in \{-1, 0, 1\}^d$. Thus, $\|x_i\| \leq \sqrt{d}$ for all i.

Now we need to demonstrate that conditions (i), (ii), (iii), and (iv) hold for any $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ satisfying $d_{p,X}(\theta, \theta_{ml}) \leq 1$.

Proof of (i) From the above, we know if $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ has log-likelihood on this data set within $b = 1$ of θ_{ml}, then the partition $(\mathcal{X}'_1, \ldots, \mathcal{X}'_k)$ induced by $\boldsymbol{\mu}$ is an optimal k-means partition of the data set. By a bias-variance decomposition, this implies

$$\begin{aligned}\Phi(\boldsymbol{\mu}) &= \sum_{j=1}^{k} \sum_{x \in \mathcal{X}'_j} \|x - \text{mean}(\mathcal{X}'_j)\|^2 + \sum_{j=1}^{k} |X'_j| \|\mu_j - \text{mean}(\mathcal{X}'_j)\|^2 \\ &= \Phi_{OPT} + \sum_{j=1}^{k} |X'_j| \|\mu_j - \text{mean}(\mathcal{X}'_j)\|^2\end{aligned}$$

where Φ_{OPT} is the optimal k-means cost of this data set. If $\|\mu_j\| \geq 2\sqrt{d}$ for some j, then we have

$$\Phi(\boldsymbol{\mu}) \;\geq\; \Phi_{OPT} + d,$$

since all points are in $\{-1, 0, 1\}^d$ and thus all means have length $\|\text{mean}(\mathcal{X}'_j)\| \leq \sqrt{d}$. But Lemma 27 implies

$$d_{p,X}(\theta_{ml}, \theta) \;=\; LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)) \;\geq\; \frac{nd}{2} \ln\left(\frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}}\right) - 2n \ln k.$$

Setting the left-hand side ≤ 1 and rearranging, we see

$$\Phi_{OPT} + d \;\leq\; \Phi(\boldsymbol{\mu}) \;\leq\; \Phi_{OPT}\left(1 + \frac{4}{d}(1 + 2\ln k)\right).$$

We know $\Phi_{OPT} \leq n\alpha(n)^2$, since this is the cost of taking the origin to be the only center. Thus, we have

$$d^2 \;\leq\; 4n\alpha(n)^2(1 + 2\ln k)$$

which is not possible by our choice of d. Therefore, we have $\|\mu_j\| < 2\sqrt{d}$ for all j.

Proof of (ii) Let $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ be the partition induced by the mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$. Taking $\sigma^2 = \gamma \frac{\Phi_{OPT}}{nd}$, Lemma 26 implies

$$\begin{aligned}LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \;&\leq\; n \ln k + \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\ln \pi_j + \frac{d}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\|x - \mu_j\|^2}{2\sigma^2}\right) \\ &\leq\; n \ln k + \frac{nd}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2 \\ &\leq\; n \ln k + \frac{nd}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\Phi_{OPT}}{2\sigma^2} \\ &=\; n \ln k + \frac{nd}{2} \ln\left(\frac{nd}{2\pi\gamma\Phi_{OPT})}\right) - \frac{nd\gamma}{2}\end{aligned}$$

Suppose the optimal k-means solution is given by centers $\boldsymbol{\mu}' = (\mu'_1, \ldots, \mu'_k)$. Let $\boldsymbol{\pi}' = (1/k, \ldots, 1/k)$ and $\sigma'^2 = \Phi_{OPT}/nd$. Note that the partition induced by the mixture $(\boldsymbol{\pi}', \boldsymbol{\mu}', \sigma')$ is precisely the

partition induced by the assigning each point to the closest center μ'_j, which in turn is an optimal k-means partition $(\mathcal{X}'_1, \ldots, \mathcal{X}'_k)$. Thus, Lemma 26 tells us

$$\begin{aligned} LL_{OPT} \;\geq\; LL(\boldsymbol{\pi}', \boldsymbol{\mu}', \sigma') \;&\geq\; \sum_{j=1}^{k} \sum_{x \in \mathcal{X}'_j} \left(\ln \pi'_j + \frac{d}{2} \ln \left(\frac{1}{2\pi \sigma'^2} \right) - \frac{\|x - \mu'_j\|^2}{2\sigma'^2} \right) \\ &= -n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi \sigma'^2} \right) - \frac{1}{2\sigma'^2} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}'_j} \|x - \mu'_j\|^2 \\ &= -n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi \sigma'^2} \right) - \frac{1}{2\sigma'^2} \Phi_{OPT} \\ &= -n \ln k + \frac{nd}{2} \ln \left(\frac{nd}{2\pi \Phi_{OPT}} \right) - \frac{nd}{2} \end{aligned}$$

Rearranging, we have

$$d_{p,X}(\theta_{ml}, \theta) \;=\; LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \;\geq\; \frac{nd}{2} \ln \gamma + \frac{nd}{2\gamma} - \frac{nd}{2} - 2n \ln k \;\geq\; \frac{nd}{2} \left(\frac{1}{2\gamma} - 1 \right) - 2n \ln k$$

where the last inequality follows from $2x \ln x > -1$ for all $x > 0$. Utilizing $d_{p,X}(\theta_{ml}, \theta) \leq 1$ and $\gamma = \frac{\sigma^2 nd}{\Phi_{OPT}}$, we have

$$\sigma^2 \geq \frac{\Phi_{OPT}}{2nd(1 + \frac{2}{d}(1 + 2 \ln k))}.$$

Since all the data points are unique and at distance at least 1 from each other, there must be at least one mean in any optimal solution that lies at least at distance 1/2 from one of the data points. Thus $\Phi_{OPT} \geq 1/4$ and

$$\sigma^2 \geq \frac{1}{16nd(1 + \ln k)}.$$

Proof of (iii) Again, take $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ to be the partition induced by the mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$. As pointed out above, $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ is an optimal k-means partition of the data, implying that no $\mathcal{X}_j$ is empty. However, from the definition of $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ as

$$\mathcal{X}_j \;=\; \left\{ x \in \mathcal{X} \;:\; j = \operatorname*{argmax}_{\ell} \; \pi_\ell N(x; \mu_\ell, \sigma^2) \right\}$$

we know that if $\mathcal{X}_j \neq \emptyset$, then $\pi_j > 0$.

Proof of (iv) Now let $\theta_{ml} = (\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*)$. From (iii), we know that the partition $(\mathcal{X}^*_1, \ldots \mathcal{X}^*_k)$ such that

$$X^*_j = \{x \;:\; j = \operatorname*{argmax}_i \pi^*_i \mathcal{N}(x; \mu^*_i, \sigma^{*2})\}$$

must satisfy that X^*_j is non-empty for all j. Further, by the convergence of the EM algorithm, we know that for any j,

$$\pi^*_j \;=\; \frac{1}{n} \sum_{x \in \mathcal{X}} \frac{\pi^*_j \mathcal{N}(x; \mu^*_j, \sigma^{*2})}{\sum_{i=1}^{k} \pi^*_i \mathcal{N}(x; \mu^*_i, \sigma^{*2})} \;\geq\; \frac{1}{n} \sum_{x \in \mathcal{X}_j} \frac{\pi^*_j \mathcal{N}(x; \mu^*_j, \sigma^{*2})}{\sum_{i=1}^{k} \pi^*_i \mathcal{N}(x; \mu^*_i, \sigma^{*2})} \;\geq\; \frac{1}{kn}$$

To complete the proof, take $\rho(n,d,k)$ to be some polynomial satisfying

$$\rho(n,d,k) \;\geq\; \max\{16nd(1+\ln k), kn\}\,.$$

■

E.2. Proof of Lemma 6

Recall that for two parameter vectors $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ and $\hat{\theta} = (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\sigma})$, their parameter distance is defined as

$$d(\theta, \hat{\theta}) \;=\; \max_i \left(\left| \ln \frac{\pi_i}{\hat{\pi}_i} \right|, |\sigma^2 - \hat{\sigma}^2|, \|\mu_i - \hat{\mu}_i\|^2 \right).$$

Lemma 28 *Suppose* $\pi, \hat{\pi}, \sigma^2, \hat{\sigma}^2 > 0$ *and* $|\ln(\pi/\hat{\pi})|, |\sigma^2 - \hat{\sigma}^2|, \|\mu - \hat{\mu}\| \leq \epsilon$. *Then for any* $x \in \mathbb{R}^d$,

$$\left| \ln \frac{\pi\, \mathcal{N}(x \,|\, \mu, \sigma^2)}{\hat{\pi}\, \mathcal{N}(x \,|\, \hat{\mu}, \hat{\sigma}^2)} \right| \leq \epsilon \cdot \max\left(1 + \frac{d}{2\sigma^2} + \frac{2\|x-\mu\| + \epsilon}{2\hat{\sigma}^2} + \frac{\|x-\mu\|^2}{2\sigma^2\hat{\sigma}^2}, \right.$$
$$\left. 1 + \frac{d}{2\hat{\sigma}^2} + \frac{2\|x-\hat{\mu}\| + \epsilon}{2\sigma^2} + \frac{\|x-\hat{\mu}\|^2}{2\sigma^2\hat{\sigma}^2} \right).$$

Proof

The proof consists of first demonstrating

$$\ln \frac{\mathcal{N}(x \,|\, \mu, \sigma^2)}{\mathcal{N}(x \,|\, \hat{\mu}, \hat{\sigma}^2)} \leq \epsilon \left(\frac{d}{2\sigma^2} + \frac{2\|x-\mu\| + \epsilon}{2\hat{\sigma}^2} + \frac{\|x-\mu\|^2}{2\sigma^2\hat{\sigma}^2} \right)$$

and then demonstrating

$$\ln \frac{\mathcal{N}(x \,|\, \hat{\mu}, \hat{\sigma}^2)}{\mathcal{N}(x \,|\, \mu, \sigma^2)} \leq \epsilon \cdot \left(\frac{d}{2\hat{\sigma}^2} + \frac{2\|x-\hat{\mu}\| + \epsilon}{2\sigma^2} + \frac{\|x-\hat{\mu}\|^2}{2\sigma^2\hat{\sigma}^2} \right).$$

Because the proofs are symmetric, we will only demonstrate the first inequality. To begin, note that we can write out the likelihood ratio as follows.

$$\begin{aligned}
\frac{\mathcal{N}(x \,|\, \mu, \sigma^2)}{\mathcal{N}(x \,|\, \hat{\mu}, \hat{\sigma}^2)} &= \left(\frac{\hat{\sigma}^2}{\sigma^2} \right)^{d/2} \exp\left[\frac{\|x-\hat{\mu}\|^2}{2\hat{\sigma}^2} - \frac{\|x-\mu\|^2}{2\sigma^2} \right] \\
&\leq \left(1 + \frac{\epsilon}{\sigma^2} \right)^{d/2} \exp\left[\frac{(\|x-\mu\| + \|\hat{\mu}-\mu\|)^2}{2\hat{\sigma}^2} - \frac{\|x-\mu\|^2}{2\sigma^2} \right] \\
&\leq \exp\left[\frac{d\epsilon}{2\sigma^2} + \frac{\|x-\mu\|^2 + 2\epsilon\|x-\mu\| + \epsilon^2}{2\hat{\sigma}^2} - \frac{\|x-\mu\|^2}{2\sigma^2} \right] \\
&= \exp\left[\frac{d\epsilon}{2\sigma^2} + \frac{2\epsilon\|x-\mu\| + \epsilon^2}{2\hat{\sigma}^2} + \frac{\|x-\mu\|^2}{2} \left(\frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma^2} \right) \right] \\
&\leq \exp\left[\frac{d\epsilon}{2\sigma^2} + \frac{2\epsilon\|x-\mu\| + \epsilon^2}{2\hat{\sigma}^2} + \frac{\epsilon\|x-\mu\|^2}{2\sigma^2\hat{\sigma}^2} \right]
\end{aligned}$$

Taking logs and factoring out ϵ gives us the inequality. ■

Given the above, Lemma 6 follows immediately.

Lemma 29 *Let* $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ *and* $\hat{\theta} = (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\sigma})$ *be two parameter vectors satisfying* $\pi_j, \hat{\pi}_j > 0$ *for all* j. *If* $\mathcal{X} = \{x : \|x\| \leq B\}$ *then* $d_p(\theta, \hat{\theta}) \leq d(\theta, \hat{\theta})\,\mathrm{poly}(1/\sigma_i^2, 1/\hat{\sigma}_i^2, \|\mu_i\|, \|\hat{\mu}_i\|, B)$.

E.3. Proof of Lemma 7

Before we prove Lemma 7, we need to bound quantities related to the Normal-Inverse-Gamma distribution and the Beta distribution.

Lemma 30 *Fix $\alpha, \beta, n_0 > 0$ and $\mu_0 \in \mathbb{R}^d$. Let q and ν be the measure associated with the Normal-Inverse-Gamma distribution with these parameters. Then for any $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$, we have*

$$-\operatorname{poly}(\alpha, 1/\alpha, \beta, 1/\beta, n_0, d, \|\mu\|, \|\mu_0\|, \sigma^2, 1/\sigma^2) \le \log q(\mu, \sigma^2) \le \operatorname{poly}(\alpha, 1/\alpha, \beta, 1/\beta, n_0, d).$$

Moreover, if $d((\mu, \sigma^2), (\hat{\mu}, \hat{\sigma}^2)) = \max\{\|\mu - \hat{\mu}\|, |\sigma^2 - \hat{\sigma}^2|\}$, then

$$\log \nu(B_d((\mu, \sigma^2), \epsilon)) \ge -\operatorname{poly}(\alpha, \beta, n_0, d, \|\mu\|, \|\mu_0\|, \sigma^2, 1/\epsilon).$$

Proof The density q can be written out as

$$q(\mu, \sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{\beta}{\sigma^2}\right) \left(\frac{n_0}{2\pi\sigma^2}\right)^{d/2} \exp\left(-\frac{\|\mu - \mu_0\|^2}{2\sigma^2/n_0}\right).$$

To see the upper bound on the density, note that the mode of this distribution occurs at $\mu = \mu_0$ and $\sigma^2 = \frac{\beta}{\alpha + d/2 + 1}$.

The lower bound on the density follows by noting that (i) $\|\mu - \mu_0\|^2 \le 2\|\mu\|^2 + 2\|\mu_0\|^2$ and (ii) $\Gamma(x) \ge 3/4$ for all x and

$$\Gamma(x) \le \begin{cases} x^x = 2^{x \log x} & \text{for } x > 1 \\ \frac{1}{x} & \text{for } x \le 1 \end{cases}.$$

The lower bound on the measure follows by combining the lower bound on the density for any point in $B_d((\mu, \sigma^2), \epsilon)$ along with the volume of $B_d((\mu, \sigma^2), \epsilon)$. ■

Lemma 31 *Let $\gamma > 0$ and take ν be the measure and q be the density associated with the symmetric $Beta(\gamma, \gamma)$ distribution. For $\theta = (w, 1-w)$, $\hat{\theta} = (\hat{w}, 1-\hat{w})$, let $d(\theta, \hat{\theta}) = \max(|\log(w/\hat{w})|, |\log((1-w)/(1-\hat{w}))|)$. If $w, 1 - w \ge \delta > 0$, we have*

$$q(\theta) \ge 2^{-\operatorname{poly}(1/\gamma, \gamma, 1/\delta)}$$

and for $\epsilon \in (0, \gamma)$,

$$\nu(B_d(\theta, \epsilon)) \ge 2^{-\operatorname{poly}(1/\gamma, \gamma, 1/\epsilon, 1/\delta)}.$$

Proof Writing out the density, we have

$$q(\theta) = \frac{\Gamma(2\gamma)}{\Gamma(\gamma)^2} \theta^{\gamma-1}(1-\theta)^{\gamma-1}.$$

The bound on $q(\theta)$ follows from Lemma 23. To see the lower bound on $\nu(B_d(\theta, \epsilon))$, assume w.l.o.g. that $w \le 1/2$. For any $\hat{w} > 0$, if $|\log(w/\hat{w})| \le \epsilon$, then $|\log((1-w)/(1-\hat{w}))| \le 2\epsilon$. This implies

$$\mathcal{I} := \{\hat{\theta} = (\hat{w}, 1-\hat{w}) : e^{-\epsilon/2} w \le \hat{w} \le e^{\epsilon/2} w\} \subset B_d(\theta, \epsilon).$$

Then we have

$$\nu(B_d(\theta,\epsilon)) \;\geq\; \nu(\mathcal{I}) \;\geq\; (e^{\epsilon/2}w - e^{-\epsilon/2}w)\min_{\hat{\theta}\in\mathcal{I}} q(\hat{\theta}) \;\geq\; \frac{\delta\epsilon}{2}\min_{\hat{\theta}\in\mathcal{I}} q(\hat{\theta}) \;\geq\; 2^{-\operatorname{poly}(\gamma,1/\gamma,1/\delta,1/\epsilon)}$$

■

Given the above two lemmas, Lemma 7 follows immediately.

Lemma 32 *Let q and ν be the prior density and measure, respectively, for the Bayesian mixture of two spherical Gaussians generative model with fixed parameters $\alpha, \beta, \gamma, \mu_0, n_0$. For any $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ and any $\epsilon > 0$, we have*

(i) $\log q(\theta) \;\geq\; -\operatorname{poly}(1/\pi_i, 1/\sigma, \|\mu_i\|, d)$,

(ii) $\log \nu(B_d(\theta,\epsilon)) \;\geq\; -\operatorname{poly}(1/\pi_i, 1/\sigma, \|\mu_i\|, d, 1/\epsilon)$, *and*

(iii) if $\gamma \geq 1$, *then* $\log q(\theta) \leq \operatorname{poly}(d)$.

E.4. Proof of Theorem 9

Theorem 9 *Let $\boldsymbol{\omega} = (\alpha, \beta, \gamma, \mu_0, n_0)$ for $\alpha, \beta, \gamma, n_0 > 0$ and $\mu_0 \in \mathbb{R}^d$. Then*

(a) MAP-MOGS$(k = 2, \boldsymbol{\omega})$ *is NP-hard if* $\gamma \geq 1$.

(b) APPROX-SAMPLING-MOGS$(k = 2, \boldsymbol{\omega})$ *is NP-hard for all* $\gamma > 0$.

Proof Here we prove (b). We again reduce from Π-MLE-MOGS-SV(k) for $k = 2$. Define

$$S = \left\{(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) : \frac{1}{\sigma^2}, \|\mu_i\|^2 \leq \rho(n,d), \pi_i > 0 \text{ for all } i\right\}$$
$$S' = \left\{(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \;:\; \frac{1}{\sigma^2}, \|\mu_i\|^2 \leq 2\rho(n,d), \pi_i > 0 \text{ for all } i\right\}$$
$$S^* = \left\{(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \;:\; \frac{1}{\sigma^2}, \|\mu_i\|^2, \frac{1}{\pi_i} \leq \rho(n,d) \text{ for all } i\right\}$$

Then we have the following.

(i) If $\theta \in S$ then $B_d(\theta, 1/2\rho(n,d)) \subset S'$ (definition of distance d).

(ii) d is $(\operatorname{poly}(n,d), S')$-admissible (Lemma 6).

(iii) Π guarantees $\theta_{ml} \in S^*$ and $B_{d_{p,X}}(\theta_{ml}, 1) \subset S$.

(iv) From (iii), $\log \nu_0(B_d(\theta_{ml}, \epsilon)) \geq -\operatorname{poly}(n, d, 1/\epsilon)$ for all $\epsilon > 0$ (Lemma 7).

Putting the above together, Theorem 6 implies (b). ■

Figure 1: Rounding our samples produces a gridding of Θ. The resulting distribution is indistinguishable from the original distribution with respect to any union of grid boxes.

Appendix F. Discretized total variation distance

Given measures μ and ν over a set Θ and a collection $\mathcal{B}$ of measurable subsets of Θ, define the *$\mathcal{B}$-variation distance* as

$$d_{\mathcal{B}}(\mu, \nu) = \sup_{B \in \mathcal{B}} |\mu(B) - \nu(B)|.$$

When $\mathcal{B}$ is the collection of all measurable subsets, this is total variation distance. For smaller collections, $d_{\mathcal{B}}$ may differ significantly from d_{TV} but is still a pseudometric.

What are minimal requirements on $\mathcal{B}$ to ensure that $d_{\mathcal{B}}$ is a meaningful probability distance? Suppose that Θ is equipped with a pseudometric $d(\cdot, \cdot)$; and, to avoid pathologies, assume (Θ, d) is separable (has a countable dense subset). Define $B_d(\theta, r) = \{\theta' \in \Theta : d(\theta, \theta') < r\}$. For $\epsilon > 0$ and $c \geq 1$, we say that a collection $\mathcal{B}$ is (d, c, ϵ)-*fine* if for every point $\theta \in \Theta$ there exists a $B \in \mathcal{B}$ such that $B_d(\theta, \epsilon) \subset B \subset B_d(\theta, c\epsilon)$. Intuitively, $\mathcal{B}$ captures the space Θ at a resolution of roughly ϵ.

For total variation distance, the supremum is taken over all measurable sets, which are closed under countable union and intersection. Likewise, we say $\mathcal{B}$ is a *standard* collection if it is closed under countable union. Note that if we have a (d, c, ϵ)-fine collection and consider its closure under countable union, the result remains (d, c, ϵ)-fine.

To understand the effect of choosing a family of sets $\mathcal{B}$, consider a simple example: suppose we sample from some distribution μ over Θ and then round the sample to r bits of precision. What is a suitable family $\mathcal{B}$? One option, illustrated in Figure 1, is to grid Θ with boxes of width $O(2^{-r})$, and let $\mathcal{B}$ be all unions of such boxes.

The following theorem generalizes this intuition and demonstrates the existence of standard (d, c, ϵ)-fine collections as well as the existence of perfect discretizations of arbitrary distributions.

Theorem 14 *Let ν be a distribution over a space Θ equipped with a pseudometric $d(\cdot, \cdot)$. For $\epsilon > 0$, suppose $\widehat{\Theta}$ is a countable ϵ-cover of Θ with respect to d. Then there exists a standard collection $\mathcal{B}$ of measurable subsets and a discrete measure $\widehat{\nu}$ over $\widehat{\Theta}$ such that*

(i) $\mathcal{B}$ is (d, c, ϵ)-fine for $c = 3$,

(ii) $d_{\mathcal{B}}(\widehat{\nu}, \nu) = 0$, and

(iii) for any discrete distribution $\widehat{\mu}$ over $\widehat{\Theta}$, $d_{\mathcal{B}}(\widehat{\mu}, \nu) = d_{TV}(\widehat{\mu}, \widehat{\nu})$.

Proof For every $\widehat{\theta} \in \widehat{\Theta}$, define the inner Voronoi cell of $\widehat{\theta}$ to be

$$C^i(\widehat{\theta}) := \{\theta \,:\, d(\theta, \widehat{\theta}) < d(\theta, \bar{\theta}) \;\; \forall \bar{\theta} \in \widehat{\Theta} \setminus \{\widehat{\theta}\}\}.$$

The Voronoi cell $C(\widehat{\theta})$ consists of $C^i(\widehat{\theta})$ as well as part of its boundary. To ensure that these cells are disjoint and cover all of Θ, we can order $\widehat{\Theta}$ and adopt the convention that the boundary occurring among any Voronoi cells belongs to the cell whose center comes earliest in the ordering.

Define $\mathcal{B}$ to be the union-closure of the set of Voronoi cells:

$$\mathcal{B} \;=\; \left\{ \cup_{\widehat{\theta} \in \mathcal{I}} C(\widehat{\theta}) \;:\; \mathcal{I} \subset \widehat{\Theta} \right\}.$$

By the countability of $\widehat{\Theta}$ we have that $\mathcal{B}$ is closed under countable union. To see that $\mathcal{B}$ is (d, c, ϵ)-fine we need to show that for every $\theta \in \Theta$, there exists a $B \in \mathcal{B}$ such that

$$B_d(\theta, \epsilon) \;\subset\; B \;\subset\; B_d(\theta, 3\epsilon).$$

Let B be the union of Voronoi cells that intersect $B_d(\theta, \epsilon)$. The first set inclusion follows immediately. To see the second set inclusion, note that because $\widehat{\Theta}$ is an ϵ-covering, $C(\widehat{\theta}) \subset B_d(\widehat{\theta}, \epsilon)$. If $C(\widehat{\theta}) \cap B_d(\theta, \epsilon) \neq \emptyset$, then we have $d(\widehat{\theta}, \theta) \leq 2\epsilon$. This implies that $C(\widehat{\theta}) \subset B_d(\theta, 3\epsilon)$. Thus, the union of such sets must also be contained in $B_d(\theta, 3\epsilon)$.

Now let $\widehat{\nu}$ denote the discrete distribution over $\widehat{\Theta}$ such that $\widehat{\nu}(\widehat{\theta}) = \nu(C(\widehat{\theta}))$. Then any $B \in \mathcal{B}$ is the countable union of such sets, so we have $\widehat{\nu}(B) = \nu(B)$, which implies $d_{\mathcal{B}}(\widehat{\nu}, \nu) = 0$.

Now consider $\widehat{\mu}$ to be any other discrete distribution over $\widehat{\Theta}$. For any $\delta > 0$, there is some $A_\delta \subset \widehat{\Theta}$ that achieves

$$|\widehat{\mu}(A_\delta) - \widehat{\nu}(A_\delta)| \;\geq\; d_{TV}(\widehat{\mu}, \widehat{\nu}) - \delta.$$

If $B = \cup_{\widehat{\theta} \in A_\delta} C(\widehat{\theta})$, then

$$d_{TV}(\widehat{\mu}, \widehat{\nu}) \;\leq\; |\widehat{\mu}(A) - \widehat{\nu}(A)| + \delta \;=\; |\widehat{\mu}(B) - \nu(B)| + \delta \;\leq\; d_{\mathcal{B}}(\widehat{\mu}, \nu) + \delta.$$

But because $d_{\mathcal{B}}$ is a pseudometric, we have

$$d_{\mathcal{B}}(\widehat{\mu}, \nu) \;\leq\; d_{\mathcal{B}}(\widehat{\mu}, \widehat{\nu}) + d_{\mathcal{B}}(\widehat{\nu}, \nu) \;=\; d_{\mathcal{B}}(\widehat{\mu}, \widehat{\nu}) \;=\; d_{TV}(\widehat{\mu}, \widehat{\nu}).$$

Since our choice of $\delta > 0$ was arbitrary, we can conclude $d_{TV}(\widehat{\mu}, \widehat{\nu}) = d_{\mathcal{B}}(\widehat{\mu}, \nu)$. ■

Since c takes a constant value in Theorem 14, we will say a collection is (d, ϵ)-fine if it is (d, c, ϵ)-fine for some constant c. With these notions in hand, we are ready to give the definition of the approximate sampling problem.

> DISCRETIZED VARIATION APPROXIMATE POSTERIOR SAMPLING: D-APPROX-SAMPLING-(p, ν_0, Θ)-d
> *Input*: A sequence of points $X \in \mathcal{X}^n$, accuracy parameter b in unary.
> *Output*: A random draw $\theta \sim \nu$ such that $d_{\mathcal{B}}(\nu, \nu_X) \leq 1/b$ where $\mathcal{B}$ is a standard $(d, 1/b)$-fine collection.

When can we guarantee that a θ from the above problem will be polynomially sized? If we take $\widehat{\Theta}$ to be a $1/b$-covering of Θ, then Theorem 14 guarantees the existence of a $(d, 1/b)$-fine collection $\mathcal{B}$ and discrete distribution ν over $\widehat{\Theta}$ such that $d_{\mathcal{B}}(\nu, \nu_X) = 0$. In the case where Θ is a bounded subset of $\mathbb{R}^m$ and d is an ℓ_p norm, for example, every element of $\widehat{\Theta}$ can be written using a polynomial number of bits. Thus, every draw from ν will be polynomially sized.

Given this, we can provide a reduction similar to the one given in Theorem 5.

Theorem 15 *Let m, $\lambda(m)$ and Π be as defined in Theorem 6, then Π-MLE-$(p,\Theta) \leq_P$ APPROX-SAMPLING-(p,Θ,ν_0)-d under randomized reductions which are polynomial in the input size and $\lambda(m)$.*

Proof Let $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and b be input to Π-MLE-(p,Θ) and let $\delta > 0$. If Π is not true, then we can return anything and terminate.

Otherwise, let $\epsilon > 0$. By Lemma 5, we can duplicate the data

$$k(\epsilon,\delta) = \frac{2}{\epsilon}\log\left(\frac{1}{\delta} - 1\right) - \log\nu_0\left(B_d\left(\theta_{ml}, \frac{\epsilon}{2n\lambda(m)}\right)\right)$$

times to ensure $\nu_Z(B_{d_{p,X}}(\theta_{ml},\epsilon)) \geq 1-\delta$ for $Z = (X^{(1)}, \ldots, X^{(k(\epsilon,\delta))})$.

If our accuracy parameter given to APPROX-SAMPLING-(p,Θ,ν_0)-d is b', the collection $\mathcal{B}$ our approximate distribution is measured against is a standard $(d,c,1/b')$-fine collection. Thus, for every $\theta \in \Theta$, there exists a $B_\theta \in \mathcal{B}$ such that $B_d(\theta,1/b') \subset B_\theta \subset B_d(\theta,c/b')$. Since $\mathcal{B}$ is standard, we also have the set

$$B = \bigcup_{\theta \in B_{d_{p,X}}(\theta_{ml},\epsilon)} B_\theta$$

is in $\mathcal{B}$. Therefore, if $\widehat{\nu}$ satisfies $d_{\mathcal{B}}(\widehat{\nu},\nu_Z) \leq \delta$, then

$$\widehat{\nu}(B) \geq \nu(B) - \delta \geq 1 - 2\delta.$$

From this we know if $\theta \sim \widehat{\nu}$, then $\theta \in B$ with probability $1-2\delta$. Let us condition on this occurring. Then there exists $\theta' \in B_{d_{p,X}}(\theta_{ml},\epsilon)$ such that $d(\theta,\theta') \leq c/b'$. For $\epsilon < 1/n\lambda(m)$ and $b' \geq c/\epsilon$, we have $\theta' \in S$ and $\theta \in S'$ and

$$\begin{aligned}
\left|\log\frac{p(X|\theta)}{p(X|\theta_{ml})}\right| &\leq \left|\log\frac{p(X|\theta)}{p(X|\theta')}\right| + \left|\log\frac{p(X|\theta')}{p(X|\theta_{ml})}\right| \\
&\leq nd_p(\theta,\theta') + \left|\log\frac{p(X|\theta')}{p(X|\theta_{ml})}\right| \\
&\leq n\lambda(m)d(\theta,\theta') + \left|\log\frac{p(X|\theta')}{p(X|\theta_{ml})}\right| \\
&\leq \epsilon(n\lambda(m)+1)
\end{aligned}$$

Let $\epsilon = 1/(n\lambda(m)+1)$, $k = k(\epsilon,\delta)$, and $b' = cb/\epsilon$. If our input to APPROX-SAMPLING-(p,Θ,ν_0)-d is a k-fold replication of X and the accuracy parameter b', then with probability at least $1-2\delta$ the output of APPROX-SAMPLING-(p,Θ,ν_0)-d is within $1/b$ of θ_{ml}. ■