Variational Inference for DPGMM with Coresets

Zalán Borsos, Olivier Bachem, Andreas Krause
Department of Computer Science
ETH Zurich
{zalan.borsos, olivier.bachem}@inf.ethz.ch, krausea@ethz.ch

Abstract

Performing estimation and inference on massive datasets under time and memory constraints is a critical task in machine learning. One approach to tackle these challenges is offered by *coresets*, succinct data summaries that come with strong theoretical guarantees, and can operate under computational resource restrictions. In this work, we explore how such data summaries can be used in posterior inference through *variational methods*. We develop a novel coreset construction for approximate posterior inference in the nonparametric Dirichlet process Gaussian mixture model. We empirically demonstrate how our method allows trading small approximation error for large gains in runtime and memory usage.

1 Introduction

Coresets are data summarization techniques that come with strong theoretical guarantees. They are domain specific and provide an upper bound on the performance gap of a specific algorithm trained on the coreset versus trained of the full data. Many problems admit coreset constructions that produce sublinear-sized summaries of the data set, examples being K-Means (Feldman et al., 2013; Lucic et al., 2016), GMM (Lucic et al., 2017), logistic regression (Huggins et al., 2016) and DP-Means (Bachem et al., 2015).

In the domain of posterior inference for mixture models, coresets have also received attention. These works all rely on reusing the coresets constructed for the negative log likelihood of the data for posterior inference, and use them to perform weighted Gibbs sampling (McGrory et al., 2014) and weighted coordinate ascent variational inference (CAVI) (Zhang et al., 2016) on Bayesian GMMs. Although the latter two works demonstrate the effectiveness of coresets in Bayesian inference on GMMs, they provide no theoretical approximation guarantees.

In this paper, we consider the problem of scaling up variational inference through coresets for Dirichlet process Gaussian mixture models (DPGMM). We develop a novel coreset construction for variational inference in the GMM model that possesses strong theoretical guarantees and relates to posterior regularization. We show how the same coreset generation method extends to the DPGMM and empirically validate our method with a weighted CAVI.

2 Background

2.1 Variational Inference

Variational inference solves the approximate posterior inference problem by finding a distribution q from a restricted density family $\mathcal{Q}$ close to the true posterior in terms of the KL divergence, This objective is equivalent to maximizing the evidence lower bound $\mathrm{ELBO}_{\mathbf{X}}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \, \mathrm{d}\mathbf{Z}$, where $\mathbf{X}$ is the observed data, and $\mathbf{Z}$ denotes all latent variables. By taking the mean field assumption that q factorizes into L disjoint groups, this objective can be maximized in the conjugate setting with coordinate ascent (CAVI). More advanced methods extended to the stochastic (Hoffman et al., 2013)

and the streaming (Broderick et al., 2013) settings, while black-box variational inference (Ranganath et al., 2014) and automatic differentiation variational inference Kucukelbir et al. (2015) work even in the non-conjugate setting, where closed-form updates are unknown.

2.2 Coresets

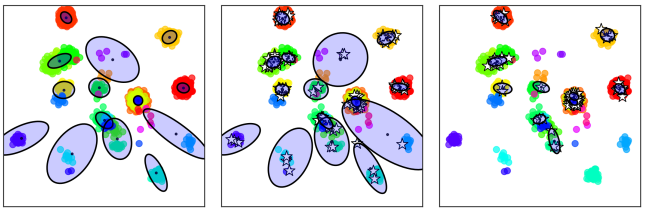
Let $\gamma_1, ..., \gamma_M$ be non-negative weights, $\mathbf{C} = \{(\gamma_1, \mathbf{x}_1), ..., (\gamma_M, \mathbf{x}_M)\}$ be a weighted set, $\mathscr{Q}$ a query space, $\mathbf{Q} \in \mathscr{Q}$ a query and a non-negative, additively decomposable cost function over the contributions f of data points, defined as $\text{cost}(\mathbf{X}, \mathbf{Q}) = \sum_{\mathbf{x} \in \mathbf{X}} f(\mathbf{x}, \mathbf{Q})$ and $\text{cost}(\mathbf{C}, \mathbf{Q}) = \sum_{(\gamma, \mathbf{x}) \in \mathbf{C}} \gamma f(\mathbf{x}, \mathbf{Q})$, respectively. Then $\mathbf{C}$ is a strong ε-coreset for $\mathbf{X}$ if for all queries $\mathbf{Q} \in \mathscr{Q}$ it holds that $|\text{cost}(\mathbf{X}, \mathbf{Q}) - \text{cost}(\mathbf{C}, \mathbf{Q})| \leq \varepsilon \text{cost}(\mathbf{X}, \mathbf{Q})$, that is, the cost function on the coreset approximates the cost on the full data up to a multiplicative factor $1 + \varepsilon$ uniformly over all queries.

From the definition above, we see that the notion of coreset is inherently problem-specific. Recently, the notion of *lightweight coreset* was coined (Bachem et al., 2017), which relaxes the traditional definition of coresets by allowing an additive approximation error in conjunction with the multiplicative error guarantee.

3 Variational Inference with Coresets

3.1 Why coresets?

The key intuition behind employing coresets for posterior inference is that these importance sampling schemes capture dense areas with a few high-weighted points, and represent small components with low-weighted points. On the other hand, uniform subsampling misses the small components with high probability. This phenomenon is illustrated by running our proposed method on a toy example, sampled from a Chinese restaurant process, where the coreset size is fixed to 2% of the data. The results of posterior inference is depicted in Figure 1, where the coresets visibly outperform the uniform subsample in representing the component posteriors more precisely. In this light, we hope to obtain a more accurate posterior approximation with coresets, while being able to work under both memory and time constraints.



(a) Full data (b) Coreset (c) Uniform

Figure 1: Posterior means and covariances of DPGMM on the full data, coreset and uniform sample. The color coding of the points denote the generated cluster belongings, coreset points are depicted by white stars, and the blue ellipses are the one standard deviation density contours of each component.

3.2 Coresets for Variational Inference in BGMM

In this section, we review the state-of-the-art method for constructing coresets for the data log likelihood of GMM and highlight its weakness in Bayesian inference. Then we propose a novel coreset construction method for variational inference in Bayesian GMM. We then show how the same algorithm is suitable for inference in Dirichlet process Gaussian mixture models.

Let $\mathbf{Z}$ denote the latent assignment variables, and let $\boldsymbol{\theta}$ refer to all latent variables except $\mathbf{Z}$. In a T-component GMM, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ includes the mixture weights, component means and precision matrices, so $\boldsymbol{\theta} = [(w_1, \boldsymbol{\mu_1}, \boldsymbol{\Lambda}_1), ..., (w_T, \boldsymbol{\mu_T}, \boldsymbol{\Lambda}_T)]$. The negative log likelihood of the data can be decomposed as $-\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) = -n \ln W(\boldsymbol{\theta}) + \phi(\mathbf{X}|\boldsymbol{\theta})$, where $W(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{w_t}{\sqrt{|2\pi \boldsymbol{\Lambda}_t^{-1}|}}$ is the normalization constant

assuring that function

$$\phi(\mathbf{X}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \phi(\boldsymbol{x}_n|\boldsymbol{\theta}) = -\sum_{n=1}^{N} \ln \left(\sum_{t=1}^{T} \frac{w_t}{W(\boldsymbol{\theta})\sqrt{|2\pi\boldsymbol{\Lambda}_t^{-1}|}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_t)^\intercal \boldsymbol{\Lambda}_t (\boldsymbol{x}_n - \boldsymbol{\mu}_t) \right) \right)$$

is positive for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Coresets for GMM were developed by Feldman et al. (2011) followed by an update from Lucic et al. (2017). Their main result claims that on the restricted set of model parameter space $\boldsymbol{\Theta}_\lambda$, where the eigenvalues of the covariance matrices are bounded by $[\lambda, 1/\lambda]$, $\lambda \in (0, 1]$, a coreset $\mathbf{C}$ of size $\mathcal{O}\left(\frac{D^4T^6 + T^2 \ln \frac{1}{\delta}}{\lambda^4 \varepsilon^2}\right)$ a multiplicative approximation guarantee for ϕ can be obtained with high probability. This in turn imply that $|\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) - \mathcal{L}(\mathbf{C}|\boldsymbol{\theta})| \to 0$ as $\varepsilon \to 0$.

In the Bayesian inference for GMM, one seeks to obtain a posterior over the covariance matrices, which entails integrating over the whole cone of positive semidefinite matrices. Since the previous guarantees only hold for the restricted parameter space, they are not easily extendable to the Bayesian framework. In this section, we show that the recently developed lightweight coresets (Bachem et al., 2017) sampling scheme also applies to the GMM. Moreover, it gives a handle on the issue of restraining the parameter space, which previous coreset guarantees suffered from.

Algorithm 1 Coreset for GMM

Input: $\mathbf{X}$ data set, M summary size
$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$
for $\boldsymbol{x} \in \mathbf{X}$ **do**
$q(\boldsymbol{x}) = \frac{1}{2N} + \frac{||\boldsymbol{x} - \boldsymbol{\mu}||^2}{2\sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2}$
end for
$\mathbf{C} \leftarrow$ sample M points with probability $q(\boldsymbol{x})$ from $\mathbf{X}$ and assign weights $\gamma_{\boldsymbol{x}} = \frac{1}{M \cdot q(\boldsymbol{x})}$
return Coreset $\mathbf{C}$

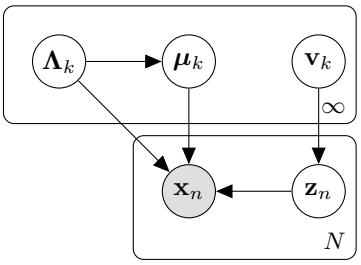


Figure 2: Graphical representation of the DPGMM

Theorem 3.1. *Let $\varepsilon > 0$, $\delta > 0$ and $k \in \mathbb{N}$. Let $\mathbf{X}$ be a set of N points from $\mathbb{R}^D$ and let $\mathbf{C}$ be the output of Algorithm 1 with $M \in \Omega\left(\frac{D^4T^4 + \log \frac{1}{\delta}}{\varepsilon^2}\right)$. Then, with probability at least $1 - \delta$:*

$$|\phi(\mathbf{X}|\boldsymbol{\theta}) - \phi(\mathbf{C}|\boldsymbol{\theta})| \leq \varepsilon\phi(\mathbf{X}|\boldsymbol{\theta}) + \varepsilon \sum_{t=1}^{T} \mathrm{Tr}(\boldsymbol{\Lambda}_t) \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2$$

uniformly for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$ is the data mean.

The resulting sampling algorithm is identical to the embarrassingly parallel lightweight coreset construction (Bachem et al., 2017). The guarantee of the theorem is scale-invariant, and the required coreset size is independent of the data set size. In contrast to the current methods, the coreset size is *independent* of underlying model parameter assumptions. Instead, the relaxing additive term is affected by "collapsing" mixture components. Moreover, this term resembles the Tikhonov regularizer in Gaussian maximum likelihood estimation (Honorio and Jaakkola, 2013).

Now, since the guarantee holds uniformly over the parameter space, we can apply the Bayesian posterior inference framework and connect the result with variational inference. We posit an approximate posterior distribution family $\mathcal{Q}_{\boldsymbol{\theta}}$ over $\boldsymbol{\theta}$ and we want to find $q^*_{\boldsymbol{\theta}} = \arg\max_{q_{\boldsymbol{\theta}} \in \mathcal{Q}_{\boldsymbol{\theta}}} f_{\mathbf{X}}(q_{\boldsymbol{\theta}})$, where $f_{\mathbf{X}}(q_{\boldsymbol{\theta}}) = \int q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{\theta}$ and $f_{\mathbf{X}}(q_{\boldsymbol{\theta}}) \geq \mathrm{ELBO}_{\mathbf{X}}(q_{\boldsymbol{\theta}})$. Taking the expectation with respect to $q_{\boldsymbol{\theta}}$ over the guarantee of Theorem 3.1, we get the following corollary:

Corollary 3.1.1. $|\mathbb{E}[\phi(\mathbf{X}|\boldsymbol{\theta})] - \mathbb{E}[\phi(\mathbf{C}|\boldsymbol{\theta})]| \leq \varepsilon\mathbb{E}[\phi(\mathbf{X}|\boldsymbol{\theta})] + \varepsilon \cdot \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 \cdot \sum_{t=1}^{T} \mathbb{E}[\mathrm{Tr}(\boldsymbol{\Lambda}_t)]$ *for all $q_{\boldsymbol{\theta}} \in \mathcal{Q}_{\boldsymbol{\theta}}$ w.p. at least $1 - \delta$. Moreover, $|f_{\mathbf{X}}(q_{\boldsymbol{\theta}}) - f_{\mathbf{C}}(q_{\boldsymbol{\theta}})| \to 0$ as $\varepsilon \to 0$.*

This lemma gives a handle on the arbitrarily bad bound of the rhs of the guarantee in Theorem 3.1, since now the additive term is defined in terms of the expectation under the approximate posterior distribution. Since the optimization objective is well approximated by employing the coreset instead of the full data, it is sensible to perform the optimization with the coreset.

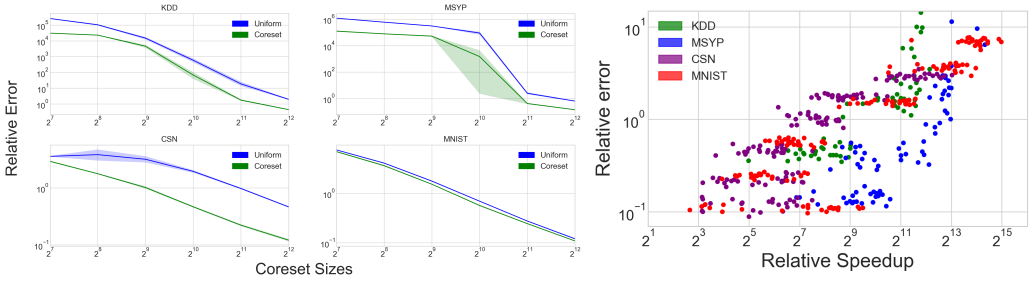
3.3 DPGMM

The CAVI for Dirichlet process mixture models (DPMM) (Antoniak, 1974; Teh, 2011) of exponential families was presented by Blei et al. (2006). In our application, we employ a member of the exponential family: a multivariate normal distribution, with means and the precisions following the Normal-Wishart distribution. We call this model the Dirichlet process Gaussian mixture model (DPGMM) and we show how our coreset construction method extends to this setting. The graphical model of the problem can be seen in Figure 2, where $\pi(\mathbf{v})$ are the mixing proportions given by the stick-breaking construction. In order to cope with the infinite mixture, the variational distribution is *truncated* at level T and is assumed to be fully decomposable over the latent variables of the model:

$$q(\mathbf{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \mathbf{Z}) = \prod_{t=1}^{T-1} \underbrace{q_{\rho_t}(v_t)}_{\text{Beta}} \prod_{t=1}^{T} \underbrace{q_{\mathbf{m}_t, \beta_t, \mathbf{W}_t, \nu_t}(\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)}_{\text{Normal-Wishart}} \prod_{n=1}^{N} \underbrace{q_{\phi_n}(\boldsymbol{z}_n)}_{\text{Multinomial}}$$

By the truncation, calculating expectations with respect to q nullifies component contributions with $t > T$ and thus the guarantees of Lemma 3.1.1 hold, i.e. the coreset construction for variational inference in DPGMM is reduced to the coreset construction of the GMM. In order to perform inference on coresets, the standard VI algorithms should accommodate weighted data. For simplicity, we illustrate it with CAVI for DPGMM in Algorithm 2 in the supplementary materials. However, our coreset retains the guarantees under more advanced methods that operate in the stochastic setting.

4 Experiments and Discussion



(a) Relative approximation error of by coresets and uniform subsampling.

(b) Speedup-accuracy tradeoff. The time measurements include also the coreset construction.

In our experiments, we compare our variational inference for DPGMM with coresets against inference on a uniform subsample. Uniform subsampling in some cases is a strong competitor, since if the data spreads out in balanced clusters, then clever importance sampling methods are not helping much. We perform the experiments on several datasets presented in the supplementary materials.

One of the most important tasks performed with Bayesian models is determining the likelihood of unseen data. With the predictive posterior (Blei et al., 2006) of our DPGMM model, one can approximate the likelihood of the $N+1$-th unseen point as $p(\mathbf{x}_{N+1}|\mathbf{X}, \boldsymbol{\theta}_0) \approx \sum_{t=1}^{T} \mathbb{E}[\pi_t(\mathbf{v})]\mathbb{E}[p(\mathbf{x}_{N+1}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)]$. We split the data into 80% training and 20% testing portion randomly. The CAVI optimization is performed with the following hyperparameters: mean precision prior $\beta_0 = 0.8$, degree of freedom of the Inverse-Wishart covariance prior $\nu_0 = D$, weight concentration prior $\alpha_0 = 5$, $\boldsymbol{\mu}_0$ set to $\mathbf{0}$, $\mathbf{W}_0$ set to $10 \cdot \mathbf{I}$ and truncation level $T = 100$. We treat every point from the test set as the $N+1$-th unseen point and report the relative error to the total log probability of the test data. The results can be seen in Figure 3a on logarithmic scale with 95 percent confidence intervals over 25 runs. The coreset outperforms uniform subsampling on some datasets, whereas on others, the performance is similar. The factor of improvement over uniform subsampling is dependent on the normalized Shannon entropy of the produced coreset sampling distribution: in case of MNIST, these values are over 0.98, which in turn means that the coreset construction samples almost uniformly.

Finally, we investigate the speedup-accuracy tradeoff of coresets. Assuming coreset size $M > D$, one iteration of CAVI on coresets offers a speedup of N/M compared to the full data. Figure 3b shows the relative speedup versus the relative error in the held-out log likelihood.

References

(2004). KDD Cup 2004. Protein Homology Dataset. `http://osmot.cs.cornell.edu/kddcup/`. Accessed: 10.11.2016.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Bachem, O., Lucic, M., and Krause, A. (2015). Coresets for nonparametric estimation-the case of dp-means. In *ICML*, pages 209–217.

Bachem, O., Lucic, M., and Krause, A. (2017). Scalable and distributed clustering via lightweight coresets. *arXiv preprint arXiv:1702.08248*.

Bertin-Mahieux, T. and Ellis, D. P. (2011). The Million Song Dataset. In *Proceedings of the 2th International Society for Music Information Retrieval Conference*.

Blei, D. M., Jordan, M. I., et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–144.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. (2013). Streaming variational bayes. In *Advances in Neural Information Processing Systems*, pages 1727–1735.

Faulkner, M., Olson, M., Chandy, R., Krause, J., Chandy, K. M., and Krause, A. (2011). The next big one: Detecting earthquakes and other rare events from community-based sensors. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, pages 13–24. IEEE.

Feldman, D., Faulkner, M., and Krause, A. (2011). Scalable training of mixture models via coresets. In *Advances in neural information processing systems*, pages 2142–2150.

Feldman, D., Schmidt, M., and Sohler, C. (2013). Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453. SIAM.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Honorio, J. and Jaakkola, T. S. (2013). Inverse covariance estimation for high-dimensional data in linear time and space: Spectral methods for riccati and sparse models. *arXiv preprint arXiv:1309.6838*.

Huggins, J., Campbell, T., and Broderick, T. (2016). Coresets for scalable bayesian logistic regression. In *Advances In Neural Information Processing Systems*, pages 4080–4088.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. (2015). Automatic variational inference in stan. In *Advances in neural information processing systems*, pages 568–576.

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

Li, Y., Long, P. M., and Srinivasan, A. (2001). Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527.

Lucic, M., Bachem, O., and Krause, A. (2016). Strong coresets for hard and soft bregman clustering with applications to exponential family mixtures. In *International Conference on Artificial Intelligence and Statistics*.

Lucic, M., Faulkner, M., Krause, A., and Feldman, D. (2017). Training Mixture Models at Scale via Coresets. *ArXiv e-prints*.

McGrory, C. A., Ahfock, D. C., Horsley, J. A., and Alston, C. L. (2014). Weighted gibbs sampling for mixture modelling of massive datasets via coresets. *Stat*, 3(1):291–299.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.

Teh, Y. W. (2011). Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer.

Zhang, M., Fu, Y., Bennett, K. M., and Wu, T. (2016). Computational efficient variational bayesian gaussian mixture models via coreset. In *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5.

5 Supplementary materials

5.1 Proof of Theorem 3.1

First we prove the following lemma, which is similar to Lemma 6 of Lucic et al. (2017). In the following, the norms refer to the $||.||_2$ norms of vectors and matrices.

Lemma 5.1. *Let* $\mathbf{X}$ *be a set of* N *points from* $\mathbb{R}^D$*. Define*

$$\tilde{f}(\boldsymbol{\theta}) = \frac{1}{N}\phi(\mathbf{X}|\boldsymbol{\theta}) + \frac{\sum_{t=1}^{T}\mathrm{Tr}(\boldsymbol{\Lambda}_t)}{N}\sum_{n=1}^{N}||\boldsymbol{x}_n - \boldsymbol{\mu}||^2$$

Then for all $n \in \{1, 2, ..., N\}$*:*

$$\sup_{\boldsymbol{\theta}}\frac{\phi(\boldsymbol{x}_i|\boldsymbol{\theta})}{\tilde{f}(\boldsymbol{\theta})} \leq \frac{6N \cdot ||\boldsymbol{x}_i - \boldsymbol{\mu}||^2}{\sum_{n=1}^{N}||\boldsymbol{x}_n - \boldsymbol{\mu}||^2} + 6$$

Proof. Rewrite $\phi(\boldsymbol{x}_i|\boldsymbol{\theta})$ as

$$\phi(\boldsymbol{x}_i|\boldsymbol{\theta}) = -\ln\left(\sum_{t=1}^{T} w_t' \exp\left(-\frac{1}{2}||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu}_t)||^2\right)\right)$$

where $w_t' = \frac{w_t}{W(\boldsymbol{\theta})\sqrt{|2\pi\boldsymbol{\Lambda}_t^{-1}|}}$. By the triangle inequality

$$||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu}_t)|| \leq ||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu})|| + ||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{\mu} - \boldsymbol{\mu}_t)||$$

Taking the squares and using $2ab \leq a^2 + b^2$ we have

$$||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu}_t)||^2 \leq 2||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu})||^2 + 2||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{\mu} - \boldsymbol{\mu}_t)||^2 \tag{1}$$

We focus on the first term on the rhs. Since the precision matrix is the inverse of the covariance matrix, it admits the eigenvalue decomposition $\boldsymbol{\Lambda}_t = \mathbf{Q}_t\mathbf{D}_t\mathbf{Q}_t^{-1}$, thus $\boldsymbol{\Lambda}_t^{\frac{1}{2}} = \mathbf{Q}_t\mathbf{D}_t^{\frac{1}{2}}\mathbf{Q}_t^{-1}$ and

$$\begin{aligned}
||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu})||^2 &= ||\mathbf{D}_t^{\frac{1}{2}}\mathbf{Q}_t^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})||^2 \\
&\leq ||\mathbf{D}_t^{\frac{1}{2}}||^2||\mathbf{Q}_t^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})||^2 \\
&= ||\mathbf{D}_t^{\frac{1}{2}}||^2||\boldsymbol{x}_i - \boldsymbol{\mu}||^2 \\
&= \lambda_t^*||\boldsymbol{x}_i - \boldsymbol{\mu}||^2 \\
&\leq \mathrm{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_i - \boldsymbol{\mu}||^2 \\
&\leq \sum_{t=1}^{T}\mathrm{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_i - \boldsymbol{\mu}||^2
\end{aligned}$$

where λ_t^* is the largest eigenvalue of $\boldsymbol{\Lambda}_t$. In the first and third line we used the 2-norm preserving property of orthogonal matrices, the inequality of second line follows from the definition of the norm, the fourth line follows from the fact that the 2-norm of the matrix equals its largest eigenvalue. Finally, in the fifth and sixth line, since $\boldsymbol{\Lambda}_t$ is positive semidefinite, we can upper bound the largest eigenvalue by the sum of all eigenvalues or equivalently, the by the trace and then we sum over all components. Plugging this result into Equation 1 and then to the definition of ϕ, we get

$$\begin{aligned}
\phi(\boldsymbol{x}_i|\boldsymbol{\theta}) &\leq -\ln\left(\sum_{t=1}^{T} w_t' \exp\left(-\sum_{t=1}^{T}\mathrm{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_i - \boldsymbol{\mu}||^2 - ||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{\mu} - \boldsymbol{\mu}_t)||^2\right)\right) \\
&= \sum_{t=1}^{T}\mathrm{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_i - \boldsymbol{\mu}||^2 - \ln\left(\sum_{t=1}^{T} w_t' \exp\left(-\frac{1}{2}||\boldsymbol{\Lambda}_t^{\frac{1}{2}}(\boldsymbol{\mu} - \boldsymbol{\mu}_t)||^2\right)^2\right) \\
&\leq \sum_{t=1}^{T}\mathrm{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_i - \boldsymbol{\mu}||^2 + 2\phi(\boldsymbol{\mu}|\boldsymbol{\theta})
\end{aligned}$$

where the last inequality follows from Jensen's inequality and the convexity of $h(x) = x^2$. We apply the inequality we just obtained for $\boldsymbol{\mu}$ and for all $\boldsymbol{x}_n \in \mathbf{X}$:

$$\phi(\boldsymbol{\mu}|\boldsymbol{\theta}) \leq \sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 + 2\phi(\boldsymbol{x}_n|\boldsymbol{\theta})$$

Now we sum over all $\boldsymbol{x}_n \in \mathbf{X}$ and divide by N:

$$\phi(\boldsymbol{\mu}|\boldsymbol{\theta}) \leq \frac{1}{N}\sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t) \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 + \frac{2}{N}\phi(\mathbf{X}|\boldsymbol{\theta})$$

Combining the last three results, we have

$$\phi(\boldsymbol{x}_i|\boldsymbol{\theta}) \leq \sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_i - \boldsymbol{\mu}||^2 + \frac{2}{N}\sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t) \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 + \frac{4}{N}\phi(\mathbf{X}|\boldsymbol{\theta})$$

We are ready to finish the proof by:

$$\begin{aligned}
\frac{\phi(\boldsymbol{x}_i|\boldsymbol{\theta})}{\tilde{f}(\boldsymbol{\theta})} &\leq \frac{\sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_i - \boldsymbol{\mu}||^2}{\tilde{f}(\boldsymbol{\theta})} + \frac{2\sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t) \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2}{N\tilde{f}(\boldsymbol{\theta})} + \frac{4\phi(\mathbf{X}|\boldsymbol{\theta})}{N\tilde{f}(\boldsymbol{\theta})} \\
&\leq \frac{\sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t)||\boldsymbol{x}_i - \boldsymbol{\mu}||^2}{\frac{1}{N}\sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t) \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2} + \frac{2\sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t) \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2}{\sum_{t=1}^{T} \operatorname{Tr}(\boldsymbol{\Lambda}_t) \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2} + \frac{4\phi(\mathbf{X}|\boldsymbol{\theta})}{\phi(\mathbf{X}|\boldsymbol{\theta})} \\
&= \frac{6N \cdot ||\boldsymbol{x}_i - \boldsymbol{\mu}||^2}{\sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2} + 6
\end{aligned}$$

where in the last inequality we applied the definition of $\tilde{f}(\boldsymbol{\theta})$ and ignored some positive terms on demand. Since the rhs does not depend on $\boldsymbol{\theta}$, this proves the claim.

□

For the rest of the proof, we denote

$$s(\boldsymbol{x}_i) = \frac{6N \cdot ||\boldsymbol{x}_i - \boldsymbol{\mu}||^2}{\sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2} + 6$$

so we have $\sup_{\boldsymbol{\theta}} \frac{\phi(\boldsymbol{x}_i|\boldsymbol{\theta})}{\tilde{f}(\boldsymbol{\theta})} \leq s(\boldsymbol{x}_i)$. We further define $S = \frac{1}{N}\sum_{n=1}^{N} s(\boldsymbol{x}_n) = 12$

Having proven Lemma 5.1, we move on to proving Theorem 3.1. The proof is similar to the proof of lightweight K-Means coresets (Bachem et al., 2017) and is based on Theorem 5 of Li et al. (2001), which we repeat for completeness:

Theorem 5.2 (Theorem 5 of Li et al. (2001))**.** *Let $\alpha > 0$, $\nu > 0$ and $\delta > 0$. Let $\mathcal{X}$ be a countably infinite domain and let q be any probability distribution over $\mathcal{X}$. Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to $[0, 1]$ with $\operatorname{Pdim}(\mathcal{F}) = d'$. Denote by C a sample of M points from $\mathcal{X}$ sampled independently according to q. Then, for $M \in \Omega\left(\frac{1}{\alpha^2\nu}(d' \log \frac{1}{\nu} + \ln \frac{1}{\delta})\right)$, with probability at least $1 - \delta$ it holds that*

$$\forall f \in \mathcal{F}: \quad d_\nu\left(\sum_{\boldsymbol{x}\in\mathcal{X}} q(\boldsymbol{x})f(\boldsymbol{x}),\ \frac{1}{|C|}\sum_{\boldsymbol{x}\in C} f(\boldsymbol{x})\right) \leq \alpha$$

where $d_\nu(a, b) = \frac{|a-b|}{a+b+\nu}$.

Now we are ready to prove Theorem 3.1.

Proof. Applying the theorem to our setting, we choose $f(\boldsymbol{x}) = \frac{\phi(\boldsymbol{x}|\boldsymbol{\theta})}{\tilde{f}(\boldsymbol{\theta})s(\boldsymbol{x})}$ and $q(\boldsymbol{x}) = \frac{s(\boldsymbol{x})}{N \cdot S}$. By the definition of $s(\boldsymbol{x})$ and S, f maps onto $[0, 1]$ and $\sum_{\boldsymbol{x}\in\mathbf{X}} q(\boldsymbol{x}) = 1$, so they satisfy the requirements of the theorem. We first note that since f maps onto $[0, 1]$, both $\sum_{\boldsymbol{x}\in\mathbf{X}} q(\boldsymbol{x})f(\boldsymbol{x})$ and $\frac{1}{|\mathbf{C}|}\sum_{\boldsymbol{x}\in\mathbf{C}} f(\boldsymbol{x})$

are upper bounded by 1. Hence we instantiate $\alpha = \varepsilon/36$, $\nu = 1/2$ and by multiplying with the denominator that is upper bounded by 3, the theorem guarantee can be transformed into:

$$\left| \sum_{\boldsymbol{x} \in \mathbf{X}} q(\boldsymbol{x}) f(\boldsymbol{x}) - \frac{1}{|\mathbf{C}|} \sum_{\boldsymbol{x} \in \mathbf{C}} f(\boldsymbol{x}) \right| \leq \frac{\varepsilon}{12} \tag{2}$$

We now have

$$\sum_{\boldsymbol{x} \in \mathbf{X}} q(\boldsymbol{x}) f(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{x} \in \mathbf{X}} \phi(\boldsymbol{x}|\boldsymbol{\theta})}{\tilde{f}(\boldsymbol{\theta}) \cdot S \cdot N} \tag{3}$$

and

$$\frac{1}{|\mathbf{C}|} \sum_{\boldsymbol{x} \in \mathbf{C}} f(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{x} \in \mathbf{C}} \frac{S \cdot N}{|\mathbf{C}| \cdot s(\boldsymbol{x})} \phi(\boldsymbol{x}|\boldsymbol{\theta})}{\tilde{f}(\boldsymbol{\theta}) \cdot S \cdot N} \tag{4}$$

Denoting $\gamma(\boldsymbol{x}) = \frac{S \cdot N}{|\mathbf{C}| \cdot s(\boldsymbol{x})}$, inserting Equations 3 and 4 into 2 and multiplying by $\tilde{f}(\boldsymbol{\theta}) \cdot S \cdot N$ we get:

$$\left| \sum_{\boldsymbol{x} \in \mathbf{X}} \phi(\boldsymbol{x}|\boldsymbol{\theta}) - \sum_{\boldsymbol{x} \in \mathbf{C}} \gamma(\boldsymbol{x}) \phi(\boldsymbol{x}|\boldsymbol{\theta}) \right| \leq \varepsilon \tilde{f}(\boldsymbol{\theta}) \cdot N \tag{5}$$

where we used the fact that $S = 12$. We arrived to the claimed error bound in the theorem.

There are two more steps to complete the proof. First, we need to argue that the sampling distribution and weighting scheme in Algorithm 1 agrees with $q(.)$ and $\gamma(.)$ This is easily seen by pattern matching.

Second, in order do claim the required coreset size, we need to find the pseudo-dimension of f. This was derived in Lucic et al. (2017) and turns out to be $\mathcal{O}\left(D^4 T^4\right)$. Since we chose $\alpha = \varepsilon/36$ and $\nu = 1/2$, the required coreset size by Theorem 5.2 is $\Omega\left(\frac{D^4 T^4 + \log \frac{1}{\delta}}{\varepsilon^2}\right)$. Our proof is now complete. □

5.2 Proof of Corollary 3.1.1

Proof. From Theorem 3.1 we know that under the lemma's assumption

$$\begin{aligned} |\mathbb{E}[\phi(\mathbf{X}|\boldsymbol{\theta})] - \mathbb{E}[\phi(\mathbf{C}|\boldsymbol{\theta})]| &= |\mathbb{E}[\phi(\mathbf{X}|\boldsymbol{\theta}) - \phi(\mathbf{C}|\boldsymbol{\theta})]| \\ &\leq \mathbb{E}[|\phi(\mathbf{X}|\boldsymbol{\theta}) - \phi(\mathbf{C}|\boldsymbol{\theta})|] \\ &\leq \varepsilon \mathbb{E}[\phi(\mathbf{X}|\boldsymbol{\theta})] + \varepsilon \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{\mu}||^2 \cdot \sum_{t=1}^{T} \mathbb{E}[\mathrm{Tr}(\boldsymbol{\Lambda}_{\boldsymbol{t}})] \end{aligned}$$

where for the last inequality we used Theorem 3.1 and the linearity of expectation.

As for the second part,

$$\begin{aligned} |f_{\mathbf{X}}(\boldsymbol{\theta}) - f_{\mathbf{C}}(\boldsymbol{\theta})| &= \left| \mathbb{E}\left[\ln \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} - \ln \frac{p(\mathbf{C}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right] \right| \\ &= |\mathbb{E}\left[\ln p(\mathbf{X}|\boldsymbol{\theta}) - \ln p(\mathbf{C}|\boldsymbol{\theta})\right]| \\ &= |\mathbb{E}[\phi(\mathbf{X}|\boldsymbol{\theta})] - \mathbb{E}[\phi(\mathbf{C}|\boldsymbol{\theta})]| \end{aligned}$$

Having reduced the problem to the previous one, we proved the claims of the lemma. □

5.3 Lower bounding

As optimizing $f_{\mathbf{C}}(q_{\boldsymbol{\theta}})$ with CAVI is intractable, we search for a suitable lower bound, which turns out to be the $\mathrm{ELBO}_{\mathbf{C}}(g)$.

$$\begin{aligned} f_{\mathbf{C}}(q_{\boldsymbol{\theta}}) &= \int q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{C}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \,\mathrm{d}\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \ln \frac{\int p(\mathbf{C}, \boldsymbol{\theta}, \boldsymbol{Z}) \,\mathrm{d}\boldsymbol{Z}}{q(\boldsymbol{\theta})} \,\mathrm{d}\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \ln \left(\int \frac{p(\mathbf{C}, \boldsymbol{\theta}, \boldsymbol{Z})}{q(\boldsymbol{\theta}) q(\mathbf{Z})} q(\mathbf{Z}) \,\mathrm{d}\boldsymbol{Z} \right) \mathrm{d}\boldsymbol{\theta} \\ &\geq \iint q(\boldsymbol{\theta}) q(\mathbf{Z}) \ln \frac{p(\mathbf{C}, \boldsymbol{\theta}, \boldsymbol{Z})}{q(\boldsymbol{\theta}) q(\mathbf{Z})} \,\mathrm{d}\boldsymbol{Z} \,\mathrm{d}\boldsymbol{\theta} \\ &= \mathrm{ELBO}_{\mathbf{C}}(g) \end{aligned}$$

where the inequality follows from Jensen's inequality. If the latent variables are discrete, the second integral is replaced by a sum.

However, our coreset retains the guarantees under more sophisticated inference methods such as black-box variational inference, which can operate even without introducing latent variables.

5.4 Weighted CAVI for DPMM

Algorithm 2 Weighted CAVI - DPMM

Input: coreset $\mathbf{C} = \{(\gamma_1, \mathbf{x}_1), ..., (\gamma_N, \mathbf{x}_N)\}$; prior params $\alpha_0, \boldsymbol{\mu}_0, \beta_0, \mathbf{W}_0, \nu_0$
Initialize $\boldsymbol{\phi}$
repeat
  for $t = 1$ **to** T **do**
    $\rho_{t,1} = 1 + \sum_{n=1}^{N} \phi_{n,t}$
    $\rho_{t,2} = \alpha_0 + \sum_{n=1}^{N} \sum_{j=t+1}^{T} \phi_{n,j}$
    $N_t = \sum_{n=1}^{N} \phi_{nt}$
    $\overline{\mathbf{x}}_t = \frac{1}{N_t} \sum_{n=1}^{N} \phi_{nt} \mathbf{x}_n$
    $\mathbf{S}_t = \frac{1}{N_t} \sum_{n=1}^{N} \phi_{nt} (\mathbf{x}_n - \overline{\mathbf{x}}_t)(\mathbf{x}_n - \overline{\mathbf{x}}_t)^\intercal$
    $\beta_t = \beta_0 + N_t$
    $\mathbf{m}_t = \frac{1}{\beta_t}(\beta_0 \mathbf{m}_0 + N_t \overline{\mathbf{x}}_t)$
    $\mathbf{W}_t^{-1} = \mathbf{W}_0^{-1} + N_t \mathbf{S}_t + \frac{\beta_0 N_t}{\beta_0 + N_t} \cdot \cdot (\overline{\mathbf{x}}_t - \boldsymbol{\mu}_0)(\overline{\mathbf{x}}_t - \boldsymbol{\mu}_0)^\intercal$
    $\nu_t = \nu_0 + N_t$
    $\mathbb{E}[\ln v_t] = \psi(\rho_{t,1}) - \psi(\rho_{t,1} + \rho_{t,2})$
    $\mathbb{E}[\ln(1 - v_t)] = \psi(\rho_{t,2}) - \psi(\rho_{t,1} + \rho_{t,2})$
    $\mathbb{E}[\ln |\boldsymbol{\Lambda}_t|] = \sum_{d=1}^{D} \Psi\left(\frac{\nu_t + 1 - d}{2}\right) + \ln |\mathbf{W}_t|$
  end for
  for $n = 1$ **to** N **do**
    for $t = 1$ **to** T **do**
      $\phi_{nt} = \exp\Big(\mathbb{E}[\ln v_t] + \sum_{i=1}^{t-1} \mathbb{E}[\ln(1 - v_i)] + \mathbb{E}[\ln |\boldsymbol{\Lambda}_t|] -$
        $- \frac{D}{2\beta_t} - \frac{\nu_t}{2}(\mathbf{x}_n - \mathbf{m}_t)^\intercal \mathbf{W}_t (\mathbf{x}_n - \mathbf{m}_t)\Big)$
    end for
    renormalize $\boldsymbol{\phi}_n$
    $\boldsymbol{\phi}_n = \boldsymbol{\gamma}_n \boldsymbol{\phi}_n$ ▷ weight modification
  end for
until convergence
return $\{(\rho_{t,1}, \rho_{t,2}), \mathbf{m}_t, \mathbf{W}_t, \beta_t, \nu_t, \boldsymbol{\phi}_{\cdot t}\}_{t=1}^{T}$

The CAVI can be naturally extended to weighted points with a single modification, which can be motivated as follows. A weighted point $(\gamma, \mathbf{x})$ can be considered as observing $\mathbf{x}$ γ times (possibly

fractional). Thus by creating γ_n copies of point $\mathbf{x}_n$, we reduced the problem to the unweighted case. In practice, we do not need to create actual copies since, for a specific point, the copies have identical ϕ vectors (resembling the responsibilities in EM) due to sharing the same location. As a result, we can concatenate them by weighting in the analogous "E" step, whereas the "M" step remains unaffected. Since we reduced our weighted scenario to the unweighted case, we are guaranteed that the resulting algorithm monotonically increases the optimization objective.

5.5 Datasets

We give a short presentation of data sets that are used throughout the experiments:

- `KDD` (kdd, 2004) — data set used for Protein Homology Prediction KDD competition. It contains 145,751 observations with 74 features that measure the match between a protein and a native sequence.
- `MSYP` Bertin-Mahieux and Ellis (2011) — contains features of songs used for prediction of the year of song release. Contains 515,345 points and 25 selected features.
- `CSN` (Faulkner et al., 2011) — cellphone accelerometer data for earthquake detection, processed into 80,000 observations and 17 features.
- `MNIST` (LeCun et al., 1998) — 70,000 low resolution images of handwritten characters transformed using PCA with whitening and retaining 10 dimensions.