Variational Tempering
Supplementary Material

Stephan Mandt
Columbia University
sm3976@columbia.edu

James McInerney
Columbia University
jm4181@columbia.edu

Farhan Abrol
Princeton University
fabrol@cs.princeton.edu

Rajesh Ranganath
Princeton University
rajeshr@cs.princeton.edu

David Blei
Columbia University
david.blei@columbia.edu

A Tempered Partition Functions

Variational Tempering requires that we precompute the tempered partition functions for a finite set of pre-specified temperatures:

$$C(T) = \int d\beta\, p(\beta) \prod_{i=1}^{N} \left(\int dx_i dz_i p(x_i, z_i|\beta)^{1/T} \right). \quad (1)$$

We first show how to reduce this to an integral only over the globals. Because the size of the remaining integration is independent of N, it is tractable by Monte-Carlo integration with a few hundred to thousand samples.

A.1 Generic model.

Here we consider a generic latent variable model of the SVI class, i.e. containing local and global hidden variables. The following calculation reduces the original integral over global and local variables to an integral over the global variables alone:

$$\begin{aligned} C(T) &= \int dx\, dz\, d\beta\; p(\beta) p(z,x|\beta)^{1/T} \quad (2) \\ &= \int d\beta\; p(\beta) \prod_{i=1}^{N} \left(\int dx_i\, dz_i\, p(x_i, z_i|\beta)^{1/T} \right) \\ &= \int d\beta\; p(\beta) \prod_{i=1}^{N} \Big(\int dz_i\, dx_i \\ &\qquad \times \exp\{T^{-1}\beta t(x_i, z_i) - T^{-1} a_l(\beta)\} \Big). \end{aligned}$$

We now use the following identity:

$$\begin{aligned} & \int dz_i\, dx_i \exp\{T^{-1}\beta t(x_i, z_i) - T^{-1} a_l(\beta)\} \\ = & \int dz_i\, dx_i\, e^{T^{-1}\beta t(x_i,z_i) - T^{-1} a_l(\beta) + a_l(T^{-1}\beta) - a_l(T^{-1}\beta)} \\ = & \; e^{-T^{-1} a_l(\beta) + a_l(T^{-1}\beta)} \underbrace{\int dz_i\, dx_i\, e^{T^{-1}\beta t(x_i,z_i) - a_l(T^{-1}\beta)}}_{=1} \\ = & \; \exp\{-T^{-1} a_l(\beta) + a_l(T^{-1}\beta)\}. \end{aligned}$$

Note that the integral is independent of i, as all data points contribute the same amount to the tempered partition function. Combining the last 2 equations yields

$$C(T) = \int d\beta\; p(\beta) \exp\{-NT^{-1} a_l(\beta) + N a_l(T^{-1}\beta)\}.$$

The complexity of computing the remaining integral does not depend on the number of data points, and therefore it is tractable with simple Monte-Carlo integration. We approximate the integral as is

$$\log C(T) \approx \log \frac{1}{N_s} \sum_{\beta \sim p(\beta)} \exp\left\{-NT^{-1} a_l(\beta) + N a_l(T^{-1}\beta)\right\}. \quad (3)$$

For the models under consideration, we found that typically less than 100 samples suffice. In more complicated setups, more advances methods to estimate the Monte Carlo integral can be used, such as annealed importance sampling. While we typically precompute the tempered partition function for about 100 values of T, the corresponding computation could easily be incorporated into the variational tempering algorithm.

Analytic approximation. Instead of precomputing the log partition function, one could also use an analytic approximation for sparse priors with a low variance (this approximation was not used in the paper). In this case, we can

MAP-approximate the $\beta-$integral, which results in

$$\log C(T) \quad \approx \quad N\left(T^{-1}a_l(\beta^*) - a_l(T^{-1}\beta^*)\right). \qquad (4)$$

For large data, this approximation gets better and yields an analytic result.

B Latent Dirichlet Allocation

B.1 Tempered partition function

We now demonstrate the calculation of the tempered partition function on the example of Latent Dirichlet Allocation (FFM). We use the multinomial representation of LDA where the topic assignments are integrated out,

$$p(w,\beta,\theta) = p(\beta)p(\theta)\prod_{nd}\left(\sum_k \theta_{dk}\beta_{kw_{dn}}\right). \qquad (5)$$

The probability that word w_{dn} is the multinomial parameter $\sum_k \theta_{dk}\beta_{kw_{dn}}$. In this formulation, LDA relates to probabilistic matrix factorization models.

LDA uses Dirichlet priors $p(\theta) = \prod_d \mathrm{Dir}(\theta_d|\alpha)$ and $p(\beta) = \prod_k \mathrm{Dir}(\beta_k|\eta)$ for the global variational parameters β and the per-document topic proportions θ.

The inner "integral" over w_{dn} is just the sum over the multinomial mean parameters,

$$\int dw_{dn}\; p(w_{dn}|\theta_d,\beta)^{1/T} \quad = \quad \sum_{v=1}^{V}\left(\sum_k \theta_{dk}\beta_{kv}\right)^{1/T}. \qquad (6)$$

The tempered partition function for LDA is therefore

$$C(T) = \int d\beta\, p(\beta)\prod_{d=1}^{D}\int d\theta_d\, p(\theta_d)\left(\sum_v\left(\sum_k \theta_{dk}\beta_{kv}\right)^{1/T}\right)^{N_d}$$
$$\approx \int d\beta\, p(\beta)\left(\int d\theta\; p(\theta)\left(\sum_v\left(\sum_k \theta_k\beta_{kv}\right)^{1/T}\right)^{N}\right)^{D},$$

where as usual $N = W/D$ is the approximate number of words per document. The corresponding Monte-Carlo approximation for the log partition function is

$$\log C(T) \quad \approx \quad \log\frac{1}{N_\beta}\sum_{\beta\sim p(\beta)}\exp\{D\log\frac{1}{N_\theta} \qquad (7)$$
$$\times\sum_{\theta\sim p(\theta)}\exp(N\log\sum_v(\sum_k \theta_k\beta_{kv})^{1/T}))\}.$$

N_θ and N_β are the number of samples from $p(\theta)$ and $p(\beta)$, respectively. To bound the log partition function we can now apply Jensen's inequality twice: Once for the concave logarithm, and once in the other direction for the convex functions $x \to x^D$ and $x \to x^N$:

$$\log C(T) \geq N\cdot D\int d\beta\, p(\beta)\int d\theta\, p(\theta)\log\sum_v\left(\sum_k \theta_k\beta_{kv}\right)^{1/T}, \qquad (8)$$

$$\log C(T) \leq N\cdot D\log\int d\beta\, p(\beta)\int d\theta\, p(\theta)\sum_v\left(\sum_k \theta_k\beta_{kv}\right)^{1/T}.$$

We see that the log partition function scales with the total number of observed words $N\times D$, This is conceptually important because otherwise $\log C(T)$ would have no effect on the updates in the limit of large data sets.

B.2 Variational updates

In its formulation with the local assignment variables z_n, the LDA model is

$$p(w,z,\beta,\theta) = p(\beta)p(\theta)\prod_{n,d,k}\exp\{z_{dnk}(\log\theta_{dk} + \log\beta_{kw_{dn}})\}.$$

Let N_{tot} be overall the number of words in the corpus, N_d the number of words in document d, D the number of documents, and K the number of topics. We have that $\sum_{d=1}^{D} N_d = N_{tot}$. The tempered model becomes

$$p(w,z,\beta,\theta,y) = p(\beta)p(\theta)\prod_m \times$$
$$\exp\{y_m\left[\sum_{n,d,k} z_{dnk}\frac{\log\theta_{dk}+\log\beta_{kw_{dn}}}{T_m} - \log C(T_m)\right]\}.$$

where m indexes temperatures.

Variational updates. We obtain the following optimal variational distributions from the complete conditionals (all up to constants). We replaced sums over word indices n by sums over the vocabulary indices v, weighted with word counts n_{dv}:

$$\log q^*(z_{dvk}) = z_{dvk}n_{dv}\mathbb{E}[1/T_y](\mathbb{E}[\log\theta_{dk}] + \mathbb{E}[\log\beta_{kv}]), \qquad (9)$$

$$\log q^*(y_m) = y_m\left[\frac{1}{T_m}\sum_{vk} n_{dv}\,\mathbb{E}[z_{dvk}](\mathbb{E}[\log\theta_{dk}] + \mathbb{E}[\log\beta_{kv}]) - \mathbb{E}[\log C(T_y)]\right],$$

$$\log q^*(\theta_{dk}) = \log\theta_{dk}\left(\sum_v n_{dv}\,\mathbb{E}[1/T_y]\mathbb{E}[z_{dvk}] + \alpha\right),$$

$$\log q^*(\beta_{kv}) = \log\beta_{kv}\left(\sum_d n_{dv}\,\mathbb{E}[1/T_y]\mathbb{E}[z_{dvk}] + \eta\right).$$

C Tempered Partition function for the Factorial Mixture Model

We apply variational tempering to the factorial mixture model (FMM) as described in the main paper,

$$p(\mathbf{X},\mathbf{Z},\mu,\pi|\alpha,\mu_0) \quad = \quad p(\mathbf{X}|\mathbf{Z},\mu,\sigma_n)p(\mathbf{Z}|\pi)p(\mu|\sigma_\mu).$$

For convenience, we define the assigned cluster means for each data point:

$$\mu_n(\mathbf{Z}) \quad = \quad \sum_k \mathbf{Z}_{nk}\mu_k. \tag{10}$$

The data generating distribution for the FMM is now a product over $D-$dimensional Gaussians:

$$p(\mathbf{X}|\mathbf{Z},\mu,\sigma_n) \quad = \quad \prod_n \mathcal{N}(\mathbf{X}_n;\mu_n(\mathbf{Z}),\sigma_n \mathbb{1}_D) \tag{11}$$

The local conditional distribution also involves the prior $\prod_n p(\mathbf{Z}_n|\pi)$ of hidden assignments,

$$p(\mathbf{X},\mathbf{Z}|\mu,\sigma_n) \quad = \quad \prod_n \mathcal{N}(\mathbf{X}_n;\mu_n(\mathbf{Z}),\sigma_n \mathbb{1}_D) p(\mathbf{Z}_n|\pi).$$

Here is the tempered local conditional:

$$p(\mathbf{X},\mathbf{Z}|\mu,\pi)^{1/T} = \prod_n \frac{1}{\sqrt{(2\pi\sigma_n)^D}}$$
$$\times \exp\left\{-\frac{1}{2\sigma_n T}\sum_n (\mathbf{X}_n-\mu_n)^\top(\mathbf{X}_n-\mu_n) + \frac{1}{T}\log p(\mathbf{Z}_n|\pi)\right\}.$$

When computing the tempered partition function, we need to integrate out all variables, starting with the locals. We can easily integrate out $\mathbf{X}$; this removes the dependence on μ which only determines the *means* of the Gaussians:

$$\begin{aligned} C(T,\mathbf{Z},\pi) \quad &= \quad \int d^D\mathbf{X}\; p(\mathbf{X},\mathbf{Z}|\mu,\pi)^{1/T} \qquad (12)\\ &= \quad \left(\prod_{n=1}^N \frac{\sqrt{(2\pi\sigma_n T)^D}}{\sqrt{(2\pi\sigma_n)^D}}\right)\prod_n p(\mathbf{Z}_n|\pi)^{1/T} \\ &= \quad \sqrt{T}^{ND}\prod_n p(\mathbf{Z}_n|\pi)^{1/T}. \end{aligned}$$

Hence, integrating the tempered Gaussians removes the $\mu-$dependence and gives an analytic contribution $\sqrt{T}^{ND}$ to the tempered partition function. It remains to compute

$$C(T) \quad = \quad \sqrt{T}^{ND}\int d\mathbf{Z}\prod_n p(\mathbf{Z}_n|\pi)^{1/T}. \tag{13}$$

Since the Bernoulli variables are discrete, the last marginalization yields

$$\begin{aligned} &\sum_{\{\mathbf{Z}_{nk}\}}\prod_{nk}\pi_k^{\mathbf{Z}_{nk}/T}(1-\pi_k)^{(1-\mathbf{Z}_{nk})/T} \\ = \quad &\prod_{nk}\sum_{\{\mathbf{Z}_{nk}=\pm 1\}}\pi_k^{\mathbf{Z}_{nk}/T}(1-\pi_k)^{(1-\mathbf{Z}_{nk})/T} \\ = \quad &\prod_k\left(\pi_k^{1/T}+(1-\pi_k)^{1/T}\right)^N. \qquad (14) \end{aligned}$$

Finally, the log tempered partition function is

$$\begin{aligned} \log C(T) \quad &= \quad \frac{1}{2}ND\log(T) + N\sum_k \log\left(\pi_k^{1/T}+(1-\pi_k)^{1/T}\right) \\ &= \quad \frac{1}{2}ND\log(T) + NK\log\left(\pi^{1/T}+(1-\pi)^{1/T}\right). \end{aligned}$$

In the last line we used that the hyperparameters $\pi_k \equiv \pi$ are isotropic in $K-$space.