Supplementary Material For Conjugate-Computation Variational Inference

A Definition of Conjugacy

The following definition is taken from Chapter 2 of Gelman et al. (2014). Suppose $\mathcal{F}$ is the class of data distributions $p(\mathbf{y}|\mathbf{z})$ parameterized by $\mathbf{z}$, and $\mathcal{P}$ is the class of prior distributions for $\mathbf{z}$, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\mathbf{z}|\mathbf{y}) \in \mathcal{P}, \quad \forall p(\cdot|\boldsymbol{\theta}) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P} \tag{21}$$

B Variational Inference in the GP Model and Issues with the SGD Algorithm

To derive the lower bound we substitute the joint-distribution (4) in the lower bound (3) and simplify:

$$\begin{aligned}
\mathcal{L}(q) &:= \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{z}) - \log q(\mathbf{z})] && (22)\\
&= \mathbb{E}_q\left[\sum_{n=1}^{N} \log p(y_n|z_n) + \log \mathcal{N}(\mathbf{z}|0, \mathbf{K}) - \log \mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})\right] && (23)\\
&= \sum_n \mathbb{E}_q\left[\log p(y_n|z_n)\right] - \mathbb{D}_{KL}[\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V}) \,\|\, \mathcal{N}(\mathbf{z}|0, \mathbf{K})] && (24)\\
&= \sum_n \mathbb{E}_q\left[\log p(y_n|z_n)\right] + \tfrac{1}{2}\left[\log|\mathbf{V}| - \text{Tr}\left(\mathbf{K}^{-1}\mathbf{V}\right) - \mathbf{m}^T\mathbf{K}^{-1}\mathbf{m}\right] + \text{constant} && (25)
\end{aligned}$$

We can see the special structure of the lower bound. The first term here might be intractable, but the second term (the KL divergence term) and its gradients have a closed-form expression when q is Gaussian. Therefore we do not need stochastic-gradient approximations for this term. A naive SGD implementation might ignore this.

There are at least three alternate parameterizations of the posterior $\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})$ in this case. We could use the natural parameters $\{\mathbf{V}^{-1}\mathbf{m}, -\frac{1}{2}\mathbf{V}^{-1}\}$, or the mean parameters $\{\mathbf{m}, \mathbf{V} + \mathbf{m}\mathbf{m}^T\}$, or simply use $\{\mathbf{m}, \mathbf{V}\}$ itself. Different parameterization lead to different updates whose computational efficiency differ drastically. For example, if we choose to update the inverse of covariance $\mathbf{V}^{-1}$, we get the following updates:

$$\mathbf{V}_{t+1}^{-1} = \mathbf{V}_t^{-1} + \frac{\rho_t}{2}\left[\frac{\partial}{\partial \mathbf{V}^{-1}} \sum_n \mathbb{E}_q\left[\log p(y_n|z_n)\right]\Big|_{V=V_t} - \tfrac{1}{2}\mathbf{V}_t + \tfrac{1}{2}\mathbf{V}_t\mathbf{K}^{-1}\mathbf{V}_t\right] \tag{26}$$

On the other hand, if we choose to update the covariance $\mathbf{V}$ instead, we get the following update:

$$\mathbf{V}_{t+1} = \mathbf{V}_t + \frac{\rho_t}{2}\left[\frac{\partial}{\partial \mathbf{V}} \sum_n \mathbb{E}_q\left[\log p(y_n|z_n)\right]\Big|_{V=V_t} + \tfrac{1}{2}\mathbf{V}_t^{-1} - \tfrac{1}{2}\mathbf{K}^{-1}\right] \tag{27}$$

The two updates are quite different. The second update involves less computation than the first one because the last term in the first update involves multiplication of three matrices. Both of these steps require explicitly forming the matrix $\mathbf{V}$ and $\mathbf{V}^{-1}$, which might be infeasible for large N (e.g. a million data points). In addition, they both compute inverse of $\mathbf{K}$ which might be ill-conditioned.

The above parameterization requires $O(N^2)$ memory, however, it is well known that for the GP model, there are only $O(N)$ free parameters (Opper and Archambeau, 2009). Choosing any of the three parameterizations discussed earlier will lead to an algorithm that is an order of magnitude slower than the best option.

Our CVI method completely avoids this re-parameterization issue by expressing the gradient steps as a conjugate computation step. Our updates naturally only have $O(N)$ free variational parameter which are obtained by using stochastic-gradients of the non-conjugate terms $\mathbb{E}_q[\log p(y_n|z_n)]$. We can reduce the number of gradients to be computed in each iteration t to $O(1)$ by using a doubly-stochastic scheme.

B.1 Stochastic Gradients with respect to the Mean Parameters

In this section, we explain the computation of the gradient of $f_n = \mathbb{E}_q[\log p(y_n|z_n)]$ with respect to the following mean parameter of the Gaussian distribution $q(z_n) = \mathcal{N}(z_n|m_n, V_{nn})$:

$$\mu_n^{(1)} := m_n, \quad \mu_n^{(2)} := V_{nn} + m_n^2 \tag{28}$$

According to (Opper and Archambeau, 2009), the gradient with respect to the mean, m_n and the variance, V_{nn} are:

$$\frac{\partial f_n}{\partial m_n} = \mathbb{E}_q\left[\frac{\partial f_n}{\partial z_n}\right], \quad \frac{\partial f_n}{\partial V_{nn}} = \tfrac{1}{2}\mathbb{E}_q\left[\frac{\partial^2 f_n}{\partial z_n^2}\right] \tag{29}$$

Therefore, we can easily approximate these gradients by using the Monte Carlo method. By using the chain rule, we can express the gradient with respect to the mean parameters in terms of the gradients with respect to m_n and V_{nn} and then use Monte Carlo. We derive these expressions below.

For notation simplicity, we drop n from now and refer to m_n and V_{nn} as m and v, respectively. We first express m and v in terms of the mean parameters: $m = \mu^{(1)}$ and $v = \mu^{(2)} - (\mu^{(1)})^2$. By using the chain rule, we express the gradient with respect to the mean parameters in terms of the gradients with respect to m and v:

$$\frac{\partial f}{\partial \mu^{(1)}} = \frac{\partial f}{\partial m}\frac{\partial m}{\partial \mu^{(1)}} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial \mu^{(1)}} = \frac{\partial f}{\partial m} - 2\frac{\partial f}{\partial v}m \tag{30}$$

$$\frac{\partial f}{\partial \mu^{(2)}} = \frac{\partial f}{\partial m}\frac{\partial m}{\partial \mu^{(2)}} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial \mu^{(2)}} = \frac{\partial f}{\partial v} \tag{31}$$

C Basics of Exponential Families

We summarize a few results regarding exponential family. Details of these results can be found in Chapter 3 of Wainwright and Jordan (2008). We assume that $q(\mathbf{z}|\boldsymbol{\lambda})$ takes the following exponential form:

$$q(\mathbf{z}|\boldsymbol{\lambda}) = h(\mathbf{z}) \exp\left\{\langle \boldsymbol{\lambda}, \boldsymbol{\phi}(\mathbf{z})\rangle - A(\boldsymbol{\lambda})\right\} \tag{32}$$

where $\boldsymbol{\phi} := [\phi_1, \phi_2, \ldots, \phi_M]$ is a vector of sufficient statistics, $\boldsymbol{\lambda} := [\lambda_1, \lambda_2, \ldots, \lambda_M]^T$ is a vector of natural parameters, $\langle \mathbf{a}, \mathbf{b}\rangle$ is an inner product, and $A(\boldsymbol{\lambda})$ is the log-partition function. The set of natural parameters is denotes by $\Omega := \{\boldsymbol{\lambda} \in \mathbb{R}^M | A(\boldsymbol{\lambda}) < \infty\}$.

We call the above representation *minimal* when there does not exist a nonzero vector $\mathbf{a} \in \mathbb{R}^M$ such that the linear combination $\langle \mathbf{a}, \boldsymbol{\phi}\rangle$ is equal to a constant. Minimal representation implies that each distribution $q(\mathbf{z}|\boldsymbol{\lambda})$ has a unique natural parametrization $\boldsymbol{\lambda}$.

We define the mean parameter associated with a sufficient statistic ϕ_m as follows:

$$\mu_m := \mathbb{E}_q\left[\phi_m(\mathbf{z})\right] \tag{33}$$

We denote the vector of parameter by $\boldsymbol{\mu}$. The set of valid mean parameters is defined as shown below:

$$\mathcal{M} := \{\boldsymbol{\mu} \in \mathbb{R}^M | \exists\ p\ s.t.\ \mathbb{E}_q[\phi_m(\mathbf{z})] = \mu_m,\ \forall m\} \tag{34}$$

It is easy to show that $A(\boldsymbol{\lambda})$ is convex, and the mean parameter can be obtained by simply differentiating it, i.e., $\boldsymbol{\mu} = \nabla A(\boldsymbol{\lambda})$. The mapping ∇A is one-to-one and onto iff the representation is minimal. This property allows us to switch back and forth between Ω and $\mathcal{M}$.

Since ∇A is convex, we can find its convex conjugate as follows:

$$A^*(\boldsymbol{\mu}) := \sup_{\lambda \in \Omega}\left\{\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle - A(\boldsymbol{\lambda})\right\} \tag{35}$$

It is easy to see that $\boldsymbol{\lambda} = \nabla A^*(\boldsymbol{\mu})$, therefore the pair of operators $(\nabla A, \nabla A^*)$ lets us switch back and forth between Ω and $\mathcal{M}$.

Bregman divergences associated with functions A and A^* is defined as follows:

$$\mathbb{B}_{A^*}(\boldsymbol{\lambda}_1 \| \boldsymbol{\lambda}_2) := A(\boldsymbol{\lambda}_1) - A(\boldsymbol{\lambda}_2) - \langle \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2, \nabla_\lambda A(\boldsymbol{\lambda}_2)\rangle \tag{36}$$

$$\mathbb{B}_{A^*}(\boldsymbol{\mu}_1 \| \boldsymbol{\mu}_2) := A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \nabla_\mu A^*(\boldsymbol{\mu}_2)\rangle \tag{37}$$

D Proof of Theorem 1

To simplify the notation, we will refer to $\tilde{p}_c(\mathbf{y}, \mathbf{z})$ and $\tilde{p}_{nc}(\mathbf{y}, \mathbf{z})$ by simply $\tilde{p}_c$ and $\tilde{p}_{nc}$ respectively. Similarly, we will refer to $q(\mathbf{z}|\boldsymbol{\lambda})$ and $q(\mathbf{z}|\boldsymbol{\lambda}_t)$ by q and q_t respectively. Using this notation and the split of the joint distribution given in Assumption 2, the variational lower bound can be written as follows:

$$\widetilde{\mathcal{L}}(\boldsymbol{\mu}) = \mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q[\log \tilde{p}_{nc}] + \mathbb{E}_q[\log(\tilde{p}_c/q)] \tag{38}$$

We prove Theorem 1 by proving several lemmas. We start with the following lemma which shows that the linear approximation of the second term (the conjugate part) in (38) simplifies to the term itself plus a KL divergence term.

Lemma 1. *For the conjugate part of the lower bound, we have the following property:*

$$\langle \boldsymbol{\mu}, \nabla_\mu \mathbb{E}_q[\log(\tilde{p}_c/q)]|_{\mu=\mu_t} \rangle = \mathbb{E}_q\left[\log(\tilde{p}_c/q)\right] + \mathbb{E}_q\left[\log(q/q_t)\right] + c \tag{39}$$

where c is a constant that does not depend on $\boldsymbol{\mu}$ (or $\boldsymbol{\lambda}$).

Proof. By substituting the definitions of $\tilde{p}_c$ and q, we get the following:

$$\langle \boldsymbol{\mu}, \nabla_\mu \mathbb{E}_q[\log(\tilde{p}_c/q)]|_{\mu=\mu_t} \rangle = \langle \boldsymbol{\mu}, \nabla_\mu \mathbb{E}_q[\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\eta} - \boldsymbol{\lambda} \rangle + A(\boldsymbol{\lambda})]|_{\mu=\mu_t} \rangle = \langle \boldsymbol{\mu}, \nabla_\mu [\langle \boldsymbol{\mu}, \boldsymbol{\eta} - \boldsymbol{\lambda} \rangle + A(\boldsymbol{\lambda})]|_{\mu=\mu_t} \rangle \tag{40}$$

We derive the gradient w.r.t. $\boldsymbol{\mu}$ below:

$$\nabla_\mu \left[\langle \boldsymbol{\mu}, \boldsymbol{\eta} - \boldsymbol{\lambda} \rangle + A(\boldsymbol{\lambda})\right] = \boldsymbol{\eta} - \boldsymbol{\lambda} - \langle \boldsymbol{\mu}, \nabla_\mu \boldsymbol{\lambda} \rangle + \nabla_\mu A(\boldsymbol{\lambda}) = \boldsymbol{\eta} - \boldsymbol{\lambda} - \mathbf{C}_\lambda^{-1}\boldsymbol{\mu} + \mathbf{C}_\lambda^{-1}\boldsymbol{\mu} = \boldsymbol{\eta} - \boldsymbol{\lambda} \tag{41}$$

where $\mathbf{C}_\lambda$ is the Fisher-information matrix and we use the fact that the gradient w.r.t. $\boldsymbol{\mu}$ is equal to $\mathbf{C}_\lambda^{-1}$ times the gradient w.r.t. $\boldsymbol{\lambda}$ (this is explained in Appendix F. Substituting this back,

$$\begin{aligned}
\langle \boldsymbol{\mu}, \nabla_\mu \mathbb{E}_q[\log(\tilde{p}_c/q)]|_{\mu=\mu_t} \rangle &= \langle \boldsymbol{\mu}, \boldsymbol{\eta} - \boldsymbol{\lambda}_t \rangle && (42)\\
&= \mathbb{E}_q\left[\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\eta} - \boldsymbol{\lambda}_t \rangle + A(\boldsymbol{\lambda}_t)\right] + c && (43)\\
&= \mathbb{E}_q\left[\log(\tilde{p}_c/q_t)\right] + c && (44)\\
&= \mathbb{E}_q\left[\log(\tilde{p}_c/q)\right] + \mathbb{E}_q\left[\log(q/q_t)\right] + c && (45)
\end{aligned}$$

□

The following lemma shows that the Bregman divergence is equal to the KL divergence which has a convenient form.

Lemma 2. *For all q and q_t satisfying Assumption 1, we have the following relationships:*

$$\mathbb{B}_{A^*}(\boldsymbol{\mu} \| \boldsymbol{\mu}_t) = \mathbb{B}_A(\boldsymbol{\lambda}_t \| \boldsymbol{\lambda}) = \mathbb{E}_q[\log(q/q_t)] \tag{46}$$

Proof. The following equivalence holds between the two Bregman divergences defined using A and A^* (see Raskutti and Mukherjee (2015), for example): $\mathbb{B}_A(\boldsymbol{\lambda}_t \| \boldsymbol{\lambda}) = \mathbb{B}_{A^*}(\boldsymbol{\mu} \| \boldsymbol{\mu}_t)$. The last equality can be proved as follows:

$$\begin{aligned}
\mathbb{E}_q[\log(q/q_t)] &= \mathbb{E}_q\left[\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\lambda} \rangle - A(\boldsymbol{\lambda}) - \langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\lambda}_t \rangle + A(\boldsymbol{\lambda}_t)\right] && (47)\\
&= A(\boldsymbol{\lambda}_t) - A(\boldsymbol{\lambda}) - \langle \boldsymbol{\lambda}_t - \boldsymbol{\lambda}, \nabla A(\boldsymbol{\lambda}) \rangle && (48)\\
&= \mathbb{B}_A(\boldsymbol{\lambda}_t \| \boldsymbol{\lambda}) = \mathbb{B}_{A^*}(\boldsymbol{\mu} \| \boldsymbol{\mu}_t) && (49)
\end{aligned}$$

□

Denoting the gradient of the non-conjugate term by $\mathbf{g}_t := \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]|_{\mu=\mu_t}$, the following lemma shows that using Lemma 1 and 2 we can get a closed-form solution for (10).

Lemma 3. *The solution of* (10) *is equal to the mean $\boldsymbol{\mu}_{t+1}$ of the following distribution:*

$$q_{t+1} \propto \left\{ e^{\langle \boldsymbol{\phi}(\mathbf{z}), \mathbf{g}_t \rangle} \tilde{p}_c \right\}^{\beta_t} (q_t)^{1-\beta_t} \tag{50}$$

Proof. Using (38), we get the following expression for the first term in (10) which we simplify in the second line using Lemma 1:

$$\langle \boldsymbol{\mu}, \widehat{\nabla}_\mu \widetilde{\mathcal{L}}(\boldsymbol{\mu}_t)\rangle = \langle \boldsymbol{\mu}, \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}] + \widehat{\nabla}_\mu \mathbb{E}_q[\log(\tilde{p}_c/q)]\rangle \tag{51}$$

$$= \langle \boldsymbol{\mu}, \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]\rangle + \mathbb{E}_q[\log(\tilde{p}_c/q)] + \mathbb{E}_q[\log(q/q_t)] + c \tag{52}$$

Plugging this in (10) and using Lemma 2, we get the following objective function:

$$\langle \boldsymbol{\mu}, \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]\rangle + \mathbb{E}_q[\log(\tilde{p}_c/q)] + \mathbb{E}_q[\log(q/q_t)] - \frac{1}{\beta_t}\mathbb{E}_q[\log(q/q_t)] \tag{53}$$

$$= \langle \boldsymbol{\mu}, \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]\rangle + \mathbb{E}_q[\log(\tilde{p}_c/q)] - \frac{1-\beta_t}{\beta_t}\mathbb{E}_q[\log(q/q_t)] \tag{54}$$

$$= \mathbb{E}_q\left[\langle \boldsymbol{\phi}(\mathbf{z}), \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]\rangle + \log(\tilde{p}_c/q) - \frac{1-\beta_t}{\beta_t}\log(q/q_t)\right] \tag{55}$$

$$= \mathbb{E}_q\left[\log \frac{\exp\left\{\langle \boldsymbol{\phi}(\mathbf{z}), \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]\rangle\right\} \tilde{p}_c q_t^{(1-\beta_t)/\beta_t}}{q^{1+(1-\beta_t)/\beta_t}}\right] \tag{56}$$

$$= \mathbb{E}_q\left[\log \frac{\exp\left\{\langle \boldsymbol{\phi}(\mathbf{z}), \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]\rangle\right\} \tilde{p}_c q_t^{(1-\beta_t)/\beta_t}}{q^{1/\beta_t}}\right] \tag{57}$$

$$= \frac{1}{\beta_t}\mathbb{E}_q\left[\log \frac{\left(\exp\left\{\langle \boldsymbol{\phi}(\mathbf{z}), \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]\rangle\right\} \tilde{p}_c\right)^{\beta_t} q_t^{(1-\beta_t)}}{q}\right] \tag{58}$$

The numerator is an unnormalized exponential family distribution which takes the same exponential family form as q (note that the base measure $h(\mathbf{z})$ is present in both $\tilde{p}_c$ and q_t which sums to $h(\mathbf{z})$ due to convex combination). The normalizing constant of this distribution does not depend on $\boldsymbol{\mu}$, therefore the minimum is obtained when the numerator is equal to the denominator (minimum of the KL divergence). This proves the lemma. □

Finally, the following lemma uses recursion to express the solution as a Bayesian inference in a conjugate model.

Lemma 4. *Given the conditions of Theorem* (1)*, the distribution q_{t+1} is equal to the posterior distribution of the following model: $q_{t+1} \propto \exp(\langle \boldsymbol{\phi}(\mathbf{z}), \widetilde{\boldsymbol{\lambda}}_t\rangle)\, \tilde{p}_c$.*

Proof. Denote the gradient of the non-conjugate term by $\mathbf{g}_t := \widehat{\nabla}_\mu \mathbb{E}_q[\log \tilde{p}_{nc}]|_{\mu=\mu_t}$, If we initialize $q_1 \propto \tilde{p}_c$ and $\widetilde{\boldsymbol{\lambda}}_0 := 0$, we can apply recursion to express q_{t+1} as a conjugate model. We demonstrate this for q_1, q_2, and q_3 below:

$$q_1 \propto (\tilde{p}_c)^{\beta_0}(\tilde{p}_c)^{1-\beta_0} = \tilde{p}_c \tag{59}$$

$$q_2 \propto \exp\langle \boldsymbol{\phi}(\mathbf{z}),\, \beta_1 \mathbf{g}_1\rangle (\tilde{p}_c)^{\beta_1}(q_1)^{1-\beta_1} \tag{60}$$

$$= \exp\langle \boldsymbol{\phi}(\mathbf{z}),\, \beta_1 \mathbf{g}_1\rangle (\tilde{p}_c)^{\beta_1}\tilde{p}_c^{1-\beta_1}$$

$$= \exp\langle \boldsymbol{\phi}(\mathbf{z}),\, \beta_1 \mathbf{g}_1\rangle \tilde{p}_c \tag{61}$$

$$= \exp\langle \boldsymbol{\phi}(\mathbf{z}),\, \widetilde{\boldsymbol{\lambda}}_1\rangle \tilde{p}_c \tag{62}$$

$$q_3 \propto \exp\langle \boldsymbol{\phi}(\mathbf{z}),\, \beta_2 \mathbf{g}_2\rangle (\tilde{p}_c)^{\beta_2}(q_2)^{1-\beta_2} \tag{63}$$

$$= \exp\langle \boldsymbol{\phi}(\mathbf{z}),\, \beta_2 \mathbf{g}_2 + (1-\beta_2)\widetilde{\boldsymbol{\lambda}}_1\rangle \tilde{p}_c \tag{64}$$

$$= \exp\langle \boldsymbol{\phi}(\mathbf{z}),\, \widetilde{\boldsymbol{\lambda}}_2\rangle \tilde{p}_c \tag{65}$$

Proceeding as above, we get the required result.

□

E Examples of CVI

E.1 Example: Generalized Linear Model

A GLM assumes the following joint distribution:

$$p(\mathbf{y}, \mathbf{z}) := \underbrace{\left[\prod_{n=1}^{N} p(y_n|\widetilde{\mathbf{x}}_n^T\mathbf{z})\right]}_{\tilde{p}_{nc}(\mathbf{y},\mathbf{z})} \mathcal{N}(\mathbf{z}|0, \delta\mathbf{I}) \tag{66}$$

where $\widetilde{\mathbf{x}}_n = [1, \mathbf{x}_n^T]^T$. For a Gaussian distribution $q := \mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})$, the data terms $p(y_n|\widetilde{\mathbf{x}}_n^T\mathbf{z})$ are the non-conjugate terms. We define $\eta_n := \widetilde{\mathbf{x}}_n^T\mathbf{z}$ and use its mean parameters in a similar way as GPs to obtain the natural parameter approximations $\widetilde{\lambda}_{n,t}^{(1)}$ and $\widetilde{\lambda}_{n,t}^{(2)}$ of the data term $p(y_n|\widetilde{\mathbf{x}}_n^T\mathbf{z})$. In step 3 of Algorithm 1, these are updated as follows:

$$\widetilde{\lambda}_{n,t}^{(i)} = (1-\beta_t)\widetilde{\lambda}_{n,t-1}^{(i)} + \beta_t \widehat{\nabla}_{\mu_n^{(i)}} \mathbb{E}_{q_t}[\log p(y_n|\eta_n)]|_{\mu=\mu_t}$$

where $\mu_n^{(i)}$ is the i'th mean parameter of $q(\eta_n)$.

Using the above parameters for the approximations, we can write Step 4 as a conjugate computation in the following Bayesian linear regression:

$$q_{t+1} \propto \left[\prod_{n=1}^{N} \mathcal{N}(\tilde{y}_{n,t}|\widetilde{\mathbf{x}}_n^T\mathbf{z}, \tilde{\sigma}_{n,t}^2)\right] \mathcal{N}(\mathbf{z}|0, \delta\mathbf{I})$$

where $\tilde{y}_{n,t} = \tilde{\sigma}_{n,t}^2 \lambda_{n,t}^{(1)}$, $\tilde{\sigma}_{n,t}^2 = -1/(2\widetilde{\lambda}_{n,t}^{(2)})$,

E.2 Example: Kalman Filters with GLM Likelihoods

We seek a Gaussian approximation $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})$ to the following time-series model (we denote time by k to differentiate it from the iteration t):

$$p(\mathbf{y}, \mathbf{z}) = \mathcal{N}(z_0|0, 1) \prod_{k=1}^{K} \mathcal{N}(z_k|z_{k-1}, \sigma^2) \underbrace{\prod_{k=1}^{K} p(y_k|z_k)}_{\tilde{p}_c(\mathbf{y},\mathbf{z})} \tag{67}$$

The likelihood terms $p(y_k|z_k)$ are non-conjugate to q and by using our method we can approximate them by $\mathcal{N}(\tilde{y}_{k,t}|z_{k-1}, \tilde{\sigma}_{k,t}^2)$ where $\tilde{y}_{k,t} = \tilde{\sigma}_{k,t}^2 \lambda_{k,t}^{(1)}$, $\tilde{\sigma}_{k,t}^2 = -1/(2\widetilde{\lambda}_{k,t}^{(2)})$, and $\widetilde{\lambda}_{k,t}^{(i)}$ are updated as follows:

$$\widetilde{\lambda}_{k,t}^{(i)} = (1-\beta_t)\widetilde{\lambda}_{k,t-1}^{(i)} + \beta_t \widehat{\nabla}_{\tilde{\mu}_k^{(i)}} \mathbb{E}_{q_t}[\log p(y_k|z_k)]|_{\mu=\mu_t}$$

with $\tilde{\mu}_k^{(i)}$ being the i'th mean parameter of $q(z_k)$.

E.3 Example: A Gamma Distribution Model

We consider a simple non-conjugate Gamma distribution model discussed by Knowles (2012). We use the following definition of the Gamma distribution: $\text{Ga}(x|\alpha, \beta) \propto x^{\alpha-1}e^{-x\beta}$, where x, α, and β are all non-negative scalars.

Given a Gamma distributed scalar observation y, we place a Gamma prior on the shape parameter z, as shown below:

$$p(y, z) = \underbrace{\text{Ga}(y|z, 1)}_{\tilde{p}_{nc}(y,z)} \text{Ga}(z|a, b) \tag{68}$$

The rate of the likelihood is fixed to 1, and the prior parameters a and b are known. Our goal is to find the posterior distribution $p(z|y)$ which we will approximate with a Gamma distribution: $q(z) = \text{Ga}(z|\alpha, \beta)$. Clearly, the likelihood is non-conjugate to q.

The sufficient statistics and mean parameters of a Gamma distribution are as follows:

$$\begin{aligned}\phi_1(z) = z,\ \mu_1 &:= \mathbb{E}_q[\phi_1(z)] = \psi(\alpha) - \log\beta \\ \phi_2(z) = \log z,\ \mu_2 &:= \mathbb{E}_q[\phi_2(z)] = \alpha/\beta\end{aligned} \tag{69}$$

where ψ is the digamma function. Using these in the CVI updates we get the following update:

$$q_{t+1} \propto e^{\left[z\widetilde{\lambda}_t^{(1)} + (\log z)\widetilde{\lambda}_t^{(2)}\right]}\mathrm{Ga}(z|a,b) \tag{70}$$

where $\widetilde{\lambda}_t^{(i)}$ are updated as follows for $i = 1, 2$:

$$\widetilde{\lambda}_t^{(i)} = (1-\beta_t)\widetilde{\lambda}_{t-1}^{(i)} + \beta_t \widehat{\nabla}_{\tilde{\mu}_i}\mathbb{E}_{q_t}[\log p(y|z)]|_{\mu=\mu_t} \tag{71}$$

The approximated term is conjugate to the Gamma distribution and therefore it is straightforward to compute the posterior parameters.

F Gradient with respect to $\boldsymbol{\mu}$ for exponential family

For some distributions in the exponential family, it may be difficult to directly compute the gradient with respect to $\boldsymbol{\mu}$. We propose to express the gradient w.r.t. $\boldsymbol{\mu}$ in terms of the Fisher information matrix and the gradient w.r.t. the natural parameter, by using the chain rule. Given the following function of interest $f(\boldsymbol{\mu}) = \mathbb{E}_{q(\mathbf{z})}[h(\mathbf{z})]$, we can formally express this as follows:

$$\frac{\partial f}{\partial \boldsymbol{\mu}} = \left[\frac{\partial^2 A(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2}\right]^{-1}\frac{\partial f}{\partial \boldsymbol{\lambda}} \tag{72}$$

Since each of these quantities can be written as expectations, as shown below, we can use the re-parametrization trick Kingma and Welling (2013) along with the Monte Carlo method to approximate them.

$$\frac{\partial^2 A(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\lambda}} = \frac{\partial \mathbb{E}_{q(\mathbf{z})}[\boldsymbol{\phi}(\mathbf{z})]}{\partial \boldsymbol{\lambda}} \tag{73}$$

$$\frac{\partial f}{\partial \boldsymbol{\lambda}} = \frac{\partial \mathbb{E}_{q(\mathbf{z})}[h(\mathbf{z})]}{\partial \boldsymbol{\lambda}} \tag{74}$$

where $\boldsymbol{\phi}(\mathbf{z})$ is the sufficient statistics of $q(\mathbf{z})$.

G Derivation of the CVI Algorithm for Mean-Field

We rewrite the objective function which naturally splits over i:

$$\max_{\mu}\sum_{i=1}^{M}\left[\left\langle \boldsymbol{\mu}_i, \widehat{\nabla}_{\mu_i}\widetilde{\mathcal{L}}(\boldsymbol{\mu}_t)\right\rangle - \frac{1}{\beta_t}\mathbb{B}_{A^*}(\boldsymbol{\mu}_i\|\boldsymbol{\mu}_{i,t})\right] \tag{75}$$

We can optimize each $\boldsymbol{\mu}_i$ parallely or use a doubly-stochastic method to optimize.

In the following, $\boldsymbol{\mu}_{/i}$ denotes the mean-parameter vector without $\boldsymbol{\mu}_i$.

To optimize with respect to a $\boldsymbol{\mu}_i$, we need to express the lower bound as a function of $\boldsymbol{\mu}_i$. By using Assumption 4, the lower bound with respect to $\boldsymbol{\mu}_i$ can be expressed as a sum over non-conjugate and conjugate parts. We show this below in (76) which is obtained by replacing the joint distribution by the conditional of $\mathbf{z}_i$. The second step afterwards is obtained by substituting (18) from Assumption 4. The third step is obtained by using the definition of $q_i(\mathbf{z}_i|\boldsymbol{\lambda}_i)$ given in (17) in

Assumption 3. The fourth step is obtained by taking the expectation inside.

$$\widetilde{L}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_{/i}) = \mathbb{E}_q\left[\log p(\mathbf{z}_i|\mathbf{x}_{/i}) - \log q_i(\mathbf{z}_i|\boldsymbol{\lambda}_i)\right] + \text{constant} \tag{76}$$

$$= \mathbb{E}_q\left[\log h_i(\mathbf{z}_i) + \sum_{a\in\mathbb{N}_i} \log \tilde{p}_{nc}^{a,i}(\mathbf{z}_i, \mathbf{x}_{a/i}) + \sum_{a\in\mathbb{N}_i} \langle \boldsymbol{\phi}_i(\mathbf{z}_i), \boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i})\rangle - \log q_i(\mathbf{z}_i|\boldsymbol{\lambda}_i)\right] + \text{constant} \tag{77}$$

$$= \mathbb{E}_q\left[\sum_{a\in\mathbb{N}_i} \log \tilde{p}_{nc}^{a,i}(\mathbf{z}_i, \mathbf{x}_{a/i}) + \sum_{a\in\mathbb{N}_i} \langle \boldsymbol{\phi}_i(\mathbf{z}_i), \boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i}) - \boldsymbol{\lambda}_i\rangle + A_i(\boldsymbol{\lambda}_i)\right] + \text{constant} \tag{78}$$

$$= \sum_{a\in\mathbb{N}_i} \mathbb{E}_q\{\log \tilde{p}_{nc}^{a,i}(\mathbf{z}_i, \mathbf{x}_{a/i})\} + \langle \boldsymbol{\mu}_i, \sum_{a\in\mathbb{N}_i} \mathbb{E}_{q_{/i}}\{\boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i})\} - \boldsymbol{\lambda}_i\rangle + A_i(\boldsymbol{\lambda}_i) + \text{constant} \tag{79}$$

This is similar to (38) since the first term is non-conjugate while the rest of the terms correspond to conjugate parts in the model. We rewrite this below by using the notation $\tilde{\boldsymbol{\eta}}_{a,i} := \mathbb{E}_{q_{/i,t}}\{\boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i})\}$:

$$\widetilde{L}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_{/i}) = \sum_{a\in\mathbb{N}_i} \mathbb{E}_q[\log \tilde{p}_{nc}^{a,i}] + \langle \boldsymbol{\mu}_i, \sum_{a\in\mathbb{N}_i} \widetilde{\boldsymbol{\eta}}_{ai} - \boldsymbol{\lambda}_i\rangle + A_i(\boldsymbol{\lambda}_i) + \text{constant} \tag{80}$$

$$= \sum_{a\in\mathbb{N}_i} \mathbb{E}_q[\log \tilde{p}_{nc}^{a,i}] + \mathbb{E}_{q_i}[\log(\tilde{p}_c^i/q_i)] + \text{constant} \tag{81}$$

where $\tilde{p}_c^i$ is a conjugate factor whose natural parameter is equal to $\sum_{a\in\mathbb{N}_i} \widetilde{\boldsymbol{\eta}}_{ai}$. Therefore, we can simply use Lemma 1 to 3 to simplify.

Using the results of Lemma 3, we get the following expression:

$$q_{i,t+1} \propto \left[\exp\left\{\left\langle \boldsymbol{\phi}_i(\mathbf{z}_i), \sum_{a\in\mathbb{N}_i} \widehat{\nabla}_{\mu_i} \mathbb{E}_q[\log \tilde{p}_{nc}^{a,i}]|_{\mu=\mu_t}\right\rangle\right\} \tilde{p}_c^i\right]^{\beta_t} (q_{i,t})^{1-\beta_t} \tag{82}$$

$$= \left[\exp\left\{\left\langle \boldsymbol{\phi}_i(\mathbf{z}_i), \sum_{a\in\mathbb{N}_i} \left[\widehat{\nabla}_{\mu_i} \mathbb{E}_q[\log \tilde{p}_{nc}^{a,i}]|_{\mu=\mu_t} + \widetilde{\boldsymbol{\eta}}_{ai}\right]\right\rangle\right\}\right]^{\beta_t} (q_{i,t})^{1-\beta_t} \tag{83}$$

We define the natural parameter of the approximation term in the exponential:

$$\widetilde{\boldsymbol{\lambda}}_{i,t} = \sum_{a\in\mathbb{N}_i} \left[\tilde{\boldsymbol{\eta}}_{ai} + \widehat{\nabla}_{\mu_i} \mathbb{E}_{q_t}[\log \tilde{p}_{nc}^{a,i}]|_{\mu=\mu_t}\right] \tag{84}$$

The natural parameter of q_{t+1} is obtained by taking a convex combination of $\widetilde{\boldsymbol{\lambda}}_{i,t}$ and the natural parameter of q_t, i.e., $\boldsymbol{\lambda}_{i,t}$:

$$\boldsymbol{\lambda}_{i,t+1} = \beta_t \widetilde{\boldsymbol{\lambda}}_{i,t} + (1-\beta_t)\boldsymbol{\lambda}_{i,t} \tag{85}$$

G.1 Equivalence to NC-VMP

We can show that NC-VMP is equivalent to our method under these conditions: the gradients w.r.t. the mean are exact and the step-size is set to 1, i.e., $\beta_t = 1$. We now present a formal proof.

We rewrite the lower bound w.r.t. $\boldsymbol{\mu}_i$ shown in (80):

$$\widetilde{L}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_{/i}) = \sum_{a\in\mathbb{N}_i} \mathbb{E}_q[\log \tilde{p}_{nc}^{a,i}] + \left\langle \boldsymbol{\mu}_i, \sum_{a\in\mathbb{N}_i} \widetilde{\boldsymbol{\eta}}_{ai} - \boldsymbol{\lambda}_i\right\rangle + A_i(\boldsymbol{\lambda}_i) + \text{constant} \tag{86}$$

By taking the derivative w.r.t. $\boldsymbol{\mu}_i$ using (41), we get the first line below.

$$\nabla_{\mu_i} \widetilde{L}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_{/i}) = \sum_{a\in\mathbb{N}_i} \nabla_{\mu_i} \mathbb{E}_q[\log \tilde{p}_{nc}^{a,i}] + \sum_{a\in\mathbb{N}_i} \widetilde{\boldsymbol{\eta}}_{ai} - \boldsymbol{\lambda}_i \tag{87}$$

We define the conjugate factor with natural parameter $\widetilde{\boldsymbol{\eta}}_{ai}$ by $\tilde{p}_c^{ai}$. We use the property that the gradient of a conjugate-exponential term, such as $\mathbb{E}_q[\log \tilde{p}_c^{ai}]$ w.r.t. $\boldsymbol{\mu}_i$ is equal to the term itself. We derived this while proving Lemma 1 in

Appendix D (although it is easy to prove by simply substituting the definition of $\tilde{p}_c^{ai}$). Therefore in the second term, we can simply substitute the gradient of $\mathbb{E}_q[\log \tilde{p}_c^{ai}]$ to get the following:

$$\nabla_{\mu_i} \widetilde{L}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_{/i}) = \sum_{a \in \mathbb{N}_i} \nabla_{\mu_i} \mathbb{E}_q[\log \tilde{p}_{nc}^{a,i}] + \sum_{a \in \mathbb{N}_i} \nabla_{\mu_i} \mathbb{E}_q[\log \tilde{p}_c^{a,i}] - \boldsymbol{\lambda}_i \tag{88}$$

$$= \sum_{a \in \mathbb{N}_i} \nabla_{\mu_i} \mathbb{E}_q[\log p(\mathbf{x}_a | \mathbf{x}_{\text{pa}_a})] - \boldsymbol{\lambda}_i \tag{89}$$

where the last line is obtain by using Assumption 4.

We also note that the derivative of the Bregman divergence term $\mathbb{B}_{A_i^*}(\boldsymbol{\mu}_i \| \boldsymbol{\mu}_{i,t})$ is equal to $\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{i,t}$.

$$\nabla_{\mu_i} \mathbb{B}_{A_i^*}(\boldsymbol{\mu}_i \| \boldsymbol{\mu}_{i,t}) = \nabla_{\mu_i} \left[A_i^*(\boldsymbol{\mu}_i) - A_i^*(\boldsymbol{\mu}_{i,t}) - \langle \boldsymbol{\mu}_i - \boldsymbol{\mu}_{i,t}, \nabla A_i^*(\boldsymbol{\mu}_{i,t}) \rangle \right] \tag{90}$$

$$= \nabla_{\mu_i} A_i^*(\boldsymbol{\mu}_i) - \nabla_{\mu_i} A_i^*(\boldsymbol{\mu}_{i,t}) \tag{91}$$

$$= \boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{i,t} \tag{92}$$

When we use β_t = 1, mirror descent reduces to the following:

$$\max_{\mu_i} \left\langle \boldsymbol{\mu}_i, \widehat{\nabla}_{\mu_i} \widetilde{\mathcal{L}}(\boldsymbol{\mu}_t) \right\rangle - \mathbb{B}_{A^*}(\boldsymbol{\mu}_i \| \boldsymbol{\mu}_{i,t}) \tag{93}$$

Taking the derivative w.r.t. $\boldsymbol{\mu}_i$ and setting it to zero, we get:

$$\boldsymbol{\lambda}_{i,t+1} = \sum_{a \in \mathbb{N}_i} \nabla_{\mu_i} \mathbb{E}_q[\log p(\mathbf{x}_a | \mathbf{x}_{\text{pa}_a})]|_{\mu=\mu_t} = \sum_{a \in \mathbb{N}_i} \mathbf{C}_{i,t}^{-1} \nabla_{\lambda_i} \mathbb{E}_q[\log p(\mathbf{x}_a | \mathbf{x}_{\text{pa}_a})]|_{\lambda=\lambda_t} \tag{94}$$

where $\mathbf{C}_{i,t}$ is the Fisher information matrix of $q_{i,t}$. This is exactly the message used in NC-VMP.

H Dataset Details

Datasets for Bayesian logistic regression is available at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`, for gamma factor model can be found at `https://github.com/davidaknowles/gamma_sgvb`, and for Gaussian-process classification can be obtained from `https://github.com/emtiyaz/prox-grad-svi`.

For all experiments, we first use grid search to tune model hyper-parameters and then fix them during our experiments. The statistics of the datasets and the model hyper-parameters used are given in Table 1.

Table 1: A list of models and datasets. N_{Train} is the number of training data. K is the number of factors. The last column shows the values of hyperparameters. The details of the hyperparameters can be found in Appendix E. For GP classification σ_f and l are hyperparameters of the squared-exponential kernel.

Model	Dataset	N	D	N_{Train}	Hyperparameters
Bayesian Logistic Regression	a1a	32,561	123	1,605	$\delta = 2.8072$
	a7a	32,561	123	16,100	$\delta = 5.0$
	Colon-cancer	62	2000	31	$\delta = 596.3623$
	Australian-scale	690	14	345	$\delta = 10^{-5}$
	Breast-cancer-scale	683	10	341	$\delta = 1.0$
	Covtype-binary-scale	581,012	54	290,506	$\delta = 0.002$
Gamma Factor Model	Cytof	522,656	40	300,000	$\sigma^2 = 0.1$, $K = 40$, $a = b = 1.0$
Gamma Matrix Factorization	MNIST	70,000	784	60,000	$a^{(z0)} = b^{(z0)} = a^{(w0)} = 0.1$ $b^{(z0)} = 0.3$, $K = 100$
Gaussian Process Classification	USPS3vs5	1,781	256	884	$\log(\sigma_f) = 5.0$, $\log(l) = 2.5$

I Algorithmic Details and Additional Results

In this section, we include 3 additional methods in our comparisons. We compare to a method called PG-SVI which is similar to the PG-exact method but uses stochastic gradients are used. Similarly, we also compare to a method called

CVI-exact which is similar to CVI but uses exact gradients. For GP classification, we compare to expectation propagation (EP).

Table 2 gives the details of algorithmic parameters used in our experiments.

Table 2: Algorithmic Parameters and Model Parameters

Model	Datasets	step size	MC samples
CVI-exact, PG-exact, CVI, S&K Alg2, S&K FG ($\beta = \frac{w}{1+w}$)			
Bayesian Logistic Regression	Colon-cancer	$w = 0.3$	10
	Australian-scale	$w = 0.4$	10
	a1a	$w = 0.4$	10
	a7a	$w = 0.4$	10
	Breast-cancer-scale	$w = 0.3$	10
	Covtype-scale	$w = 0.3$	10
Knowles, CVI, where w_0 denotes the initial step size in Knowles (Ada-delta)			
Gamma Factor Model	Cytof	$w_0 = 10.0$ (Knowles) $\beta = 5 \times 10^{-5}$ (CVI)	50
ADAM, CVI, where w_0 denotes the initial step size in ADAM			
Gamma Matrix Factorization	MNIST	$w_0 = 0.5$ (ADAM) $\beta = 0.02$ (CVI)	10
CVI-exact, PG-exact, CVI, PG-SVI ($\beta = \frac{w}{1+w}$)			
Gaussian Process Classification	USPS3vs5	$w = 1.0$ (CVI-exact, PG-exact) $w = 0.3$ (CVI,PC-SVI)	100

I.1 Additional Results

We compare Bayesian logistic regression on seven real datasets. The results are summarized in Table 3. All methods reach the same performance. Chol is the slowest method. When $D > N$ S&K-FG is supposed to perform better than S&K-Alg2, but the situation is reversed when $N > D$. PG-Exact and CVI-exact are expected to have the same performance. CVI is expected to be a faster than them because stochastic gradients might be cheaper to compute. It is also expected to perform well for both $N > D$ regime and $D > N$ regime.

Additional results for the gamma factor model and gamma matrix factorization model are in Table 4 and 5 respectively.

For GP Classification, we present results below where we compare our method (CVI) to the following methods: expectation propagation (EP), explicit optimization with LBFGS using Cholskey factorization (Chol), Proximal gradient methods (PG-SVI). For PG-SVI and CVI, we use MC approximation to compute gradient while for CVI-exact, we use exact gradient. Figure 2 shows the result of Gaussian Process Classification.

J Details of the Gamma Factor Model

We consider the model discussed by Knowles (2015). In this model, observations $\mathbf{y}_i \in \mathbb{R}^D$, $i = 1 \dots N$ are modeled as

$$p(\mathbf{Y}, \mathbf{Z}|\sigma^2, a, b) = p(\mathbf{Y}|\mathbf{Z})p(\mathbf{Z}) = \left[\prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{Z}, \sigma^2)\right]\left[\prod_{j=1}^{D}\prod_{k=1}^{K} p(Z_{jk}|a, b)\right] \tag{95}$$

where each column of $\mathbf{Y}$ follows $p(\mathbf{y}_i|\mathbf{Z}, \sigma^2) = \mathcal{N}(\mathbf{y}_i|0, \mathbf{Z}\mathbf{Z}^T + \sigma^2 I)$ and each element of $\mathbf{Z}$ follows $p(Z_{jk}|a, b) = \text{Ga}(Z_{jk}|a, b)$ with the following parameterization $\text{Ga}(x|\alpha, \beta) \propto x^{\alpha-1}e^{-x\beta}$.

This is a non-conjugate model since the data term $p(\mathbf{y}|\mathbf{Z})$ is not conjugate to the prior $p(\mathbf{Z})$. We choose the following mean-field approximation:

$$q(\mathbf{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{D} q(Z_{j,k}).$$

where each factor is a Gamma distribution.

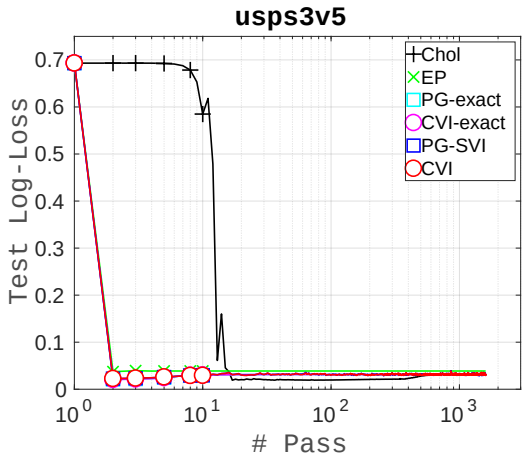

Figure 2: Comparison on Gaussian Process Classification.

K Details of the Gamma Matrix Factorization

Given the data matrix $\mathbf{X}$ of size $V \times N$, the Gamma matrix-factorization assumes the following joint-distribution:

$$p(\mathbf{X}, \mathbf{W}, \mathbf{Z}) = \prod_{i=1}^{V} \left[\prod_{j=1}^{N} p(X_{i,j} | \mathbf{w}_i^T \mathbf{z}_j) \right] \times \left[\prod_{i=1}^{V} \prod_{k=1}^{K} \text{Ga}(w_{k,i} | a_{k,i}^{(w0)}, b_{k,i}^{(w0)}) \right] \left[\prod_{j=1}^{N} \prod_{k=1}^{K} \text{Ga}(z_{k,j} | a_{k,j}^{(z0)}, b_{k,j}^{(z0)}) \right] \tag{96}$$

where $\mathbf{w}_i, \mathbf{z}_j$ are K dimensional latent vectors, $\mathbf{W}$ and $\mathbf{Z}$ are $K \times V$ and $K \times N$ matrices respectively. The likelihood term $p(X_{i,j} | \mathbf{w}_i^T \mathbf{z}_j)$ is a Poisson distribution. We use the following gamma posterior:

$$q(\mathbf{W}, \mathbf{Z}) = \left[\prod_{i=1}^{V} \prod_{k=1}^{K} \text{Ga}(w_{k.i} | a_{k,i}^{(w)}, b_{k,i}^{(w)}) \right] \left[\prod_{j=1}^{N} \prod_{k=1}^{K} \text{Ga}(z_{k,j} | a_{k,j}^{(z)}, b_{k,j}^{(z)}) \right] \tag{97}$$

References

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35.

Bouchard, G. (2007). Efficient bounds for the softmax and applications to approximate inference in hybrid models. In *NIPS 2007 Workshop on Approximate Inference in Hybrid Models*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Genkin, A., Lewis, D. D., and Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304.

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

Honkela, A. and Valpola, H. (2004). Unsupervised variational Bayesian learning of nonlinear models. In *Advances in Neural Information Processing Systems*, pages 593–600.

Jaakkola, T. and Jordan, M. (1996). A variational approach to Bayesian logistic regression problems and their extensions. In *International conference on Artificial Intelligence and Statistics*.

Khan, M. E. (2012). *Variational Learning for Latent Gaussian Models of Discrete Data*. PhD thesis, University of British Columbia.

Khan, M. E., Aravkin, A. Y., Friedlander, M. P., and Seeger, M. (2013). Fast dual variational inference for non-conjugate latent Gaussian models. In *International Conference on Machine Learning*.

Khan, M. E., Babanezhad, R., Lin, W., Schmidt, M., and Sugiyama, M. (2016). Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Khan, M. E., Baque, P., Flueret, F., and Fua, P. (2015). Kullback-Leibler Proximal Variational Inference. In *Advances in Neural Information Processing Systems*.

Khan, M. E., Marlin, B., Bouchard, G., and Murphy, K. (2010). Variational Bounds for Mixed-Data Factor Analysis. In *Advances in Neural Information Processing Systems*.

Khan, M. E., Mohamed, S., and Murphy, K. (2012). Fast Bayesian inference for non-conjugate Gaussian process regression. In *Advances in Neural Information Processing Systems*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Knowles, D. A. (2012). *Bayesian non-parametric models and inference for sparse and hierarchical latent structure*. PhD thesis, University of Cambridge.

Knowles, D. A. (2015). Stochastic gradient variational Bayes for gamma approximating distributions. *arXiv preprint arXiv:1509.01631*.

Knowles, D. A. and Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*, pages 1701–1709.

Kuss, M. and Rasmussen, C. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704.

Marlin, B., Khan, M., and Murphy, K. (2011). Piecewise bounds for estimating Bernoulli-logistic latent Gaussian models. In *International Conference on Machine Learning*.

Minka, T. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Mohamed, S., Ghahramani, Z., and Heller, K. A. (2009). Bayesian exponential family pca. In *Advances in neural information processing systems*, pages 1089–1096.

Nemirovskii, A., Yudin, D. B., and Dawson, E. R. (1983). Problem complexity and method efficiency in optimization.

Opper, M. and Archambeau, C. (2009). The Variational Gaussian Approximation Revisited. *Neural Computation*, 21(3):786–792.

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

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. (2015). Deep exponential families. In *International conference on Artificial Intelligence and Statistics*.

Raskutti, G. and Mukherjee, S. (2015). The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press.

Salimans, T., Knowles, D. A., et al. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882.

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1–2:1–305.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(1):1005–1031.

Winn, J. and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694.

Table 3: A summary of the results obtained on Bayesian logistic regression. In all columns, a lower value implies better performance. We report total time of convergence.

Dataset	Methods	Neg-Log-Lik	Log Loss	Time
a1a ($N > D$)	Chol	591.4	0.49	0.82s
	S&K Alg2	590.5	0.49	0.07s
	S&K FG	590.5	0.49	0.09s
	PG-exact	591.6	0.49	0.15s
	CVI-exact	590.5	0.49	0.10s
	CVI	590.4	0.49	0.10s
a7a ($N > D$)	Chol	5,418.1	0.47	17.79s
	S&K Alg2	5,416.4	0.47	0.74s
	S&K FG	5,416.3	0.47	1.19s
	PG-exact	5,418.0	0.47	1.35s
	CVI-exact	5,416.3	0.47	1.17s
	CVI	5,416.3	0.47	0.95s
Colon-cancer ($D > N$)	Chol	18.26	0.694	93.229s
	S&K Alg2	18.26	0.693	6.142s
	S&K FG	18.26	0.693	0.026s
	PG-exact	18.25	0.696	0.052s
	CVI-exact	18.26	0.698	0.012s
	CVI	18.26	0.698	0.021s
Australian-scale ($N > D$)	Chol	191.62	0.473	0.193s
	S&K Alg2	190.99	0.480	0.013s
	S&K FG	190.95	0.479	0.034s
	PG-exact	191.57	0.479	0.056s
	CVI-exact	191.14	0.480	0.020s
	CVI	191.30	0.478	0.011s
Breast-cancer-scale ($N > D$)	Chol	34.21	0.139	0.110s
	S&K Alg2	34.20	0.139	0.014s
	S&K FG	34.15	0.137	0.036s
	PG-exact	34.18	0.138	0.063s
	CVI-exact	34.24	0.138	0.032s
	CVI	34.15	0.140	0.021s
Covtype-scale ($N > D$) but N is large	Chol	149,641	0.7404	198.1932s
	S&K Alg2	149,623	0.7403	56.7972s
	S&K FG	149,612	0.7403	20.309s
	PG-exact	149,615	0.7403	42.6777s
	CVI-exact	149,615	0.7403	39.5720s
	CVI	149,616	0.7403	14.3319s

Table 4: Results obtained on Gamma factor model, a lower value implies better performance. CVI is much faster than Knowles method.

Dataset	Methods	Log Loss	Time
Cytof	Knowles	52.25	210.03s
	CVI	52.52	50.91s

Table 5: Results obtained on Gamma Matrix Factorization, a lower value implies better performance. CVI outperforms ADAM.

Dataset	Methods	Test Loss	Time
MNIST	ADAM	0.000125	1776.83s
	CVI	0.000119	1692.64s