Supplementary Material for "Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions"

1 Examples of Splitting for Variational-Gaussian Inference

We give detailed derivations for the splitting-examples shown in Section 3.1 in the main paper. As in the main paper, we denote the Gaussian posterior distribution by $q(\mathbf{z}|\boldsymbol{\lambda}) := \mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})$, so that $\boldsymbol{\lambda} = \{\mathbf{m}, \mathbf{V}\}$ with $\mathbf{m}$ being the mean and $\mathbf{V}$ being the covariance matrix.

1.1 Gaussian Process (GP) Models

Consider GP models for N input-output pairs $\{y_n, \mathbf{x}_n\}$ indexed by n. Let $z_n := f(\mathbf{x}_n)$ be the latent function drawn from a GP with a zero-mean function and a covariance function $\kappa(\mathbf{x}, \mathbf{x}')$. We denote the Kernel matrix obtained on the data $\mathbf{x}_n$ for all n by $\mathbf{K}$.

We use a non-Gaussian likelihood $p(y_n|z_n)$ to model the output, and assume that each y_n is independently sampled from this likelihood given $\mathbf{z}$. The joint-distribution over $\mathbf{y}$ and $\mathbf{z}$ is shown below:

$$p(\mathbf{y}, \mathbf{z}) = \prod_{n=1}^{N} p(y_n|z_n)\mathcal{N}(\mathbf{z}|0, \mathbf{K}) \tag{1}$$

The ratio required for the lower bound is shown below, along with the split, where non-Gaussian terms are in $\tilde{p}_d$ and Gaussian terms are in $\tilde{p}_e$:

$$\frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z}|\mathbf{m}, \mathbf{V})} = \underbrace{\prod_{n=1}^{N} p(y_n|z_n)}_{\tilde{p}_d(\mathbf{z}|\boldsymbol{\lambda})} \underbrace{\frac{\mathcal{N}(\mathbf{z}|0, \mathbf{K})}{\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})}}_{\tilde{p}_e(\mathbf{z}|\boldsymbol{\lambda})}. \tag{2}$$

By substituting in Eq. 1 of the main paper, we can obtain the lower bound $\underline{\mathcal{L}}$ after a few simplifica-

tions, as shown below:

$$\underline{\mathcal{L}}(\mathbf{m}, \mathbf{V}) := \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z}|\mathbf{m}, \mathbf{V})}\right], \tag{3}$$

$$= \mathbb{E}_{q(\mathbf{z})}\left[\sum_{n=1}^{N} \log p(y_n|z_n)\right] + \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{\mathcal{N}(\mathbf{z}|0, \mathbf{K})}{\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})}\right], \tag{4}$$

$$= \underbrace{\sum_{n=1}^{N} \mathbb{E}_q[\log p(y_n|z_n)]}_{-f(\boldsymbol{\lambda})} - \underbrace{\mathbb{D}_{KL}[\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V}) \,\|\, \mathcal{N}(\mathbf{z}|0, \mathbf{K})]}_{h(\boldsymbol{\lambda})}. \tag{5}$$

The assumption A2 is satisfied since the KL divergence is convex in both $\mathbf{m}$ and $\mathbf{V}$. This is clear from the expression of the KL divergence:

$$D_{KL}\left[\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})||\mathcal{N}(\mathbf{z}|0, \mathbf{K})\right] = \tfrac{1}{2}[-\log|\mathbf{V}\mathbf{K}^{-1}| + \mathrm{Tr}(\mathbf{V}\mathbf{K}^{-1}) + \mathbf{m}^T\mathbf{K}^{-1}\mathbf{m} - D] \tag{6}$$

where D is the dimensionality of $\mathbf{z}$. Convexity w.r.t. $\mathbf{m}$ follows from the fact that the above is quadratic in $\mathbf{m}$. Convexity w.r.t. $\mathbf{V}$ follows due to concavity of $\log|\mathbf{V}|$ (trace is linear, so does not matter).

Assumption A1 depends on the choice of the likelihood $p(y_n|z_n)$, but is usually satisfied. Simplest example is a Gaussian likelihood for which the function f takes the following form:

$$f(\mathbf{m}, \mathbf{V}) = \sum_{n=1}^{N} \mathbb{E}_q[-\log p(y_n|z_n)] = \sum_{n=1}^{N} \mathbb{E}_q[-\log \mathcal{N}(y_n|z_n, \sigma^2)] \tag{7}$$

$$= \sum_{n=1}^{N} \tfrac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left[(y_n - m_n)^2 + v_n\right] \tag{8}$$

where m_n is the n'th element of $\mathbf{m}$ and v_n is the n'th diagonal entry of $\mathbf{V}$. This clearly satisfies A1, since the objective is quadratic in $\mathbf{m}$ and linear in $\mathbf{V}$.

Here is an example where A1 is not satisfied: for Poisson likelihood $\log p(y_n|z_n) = \exp[y_n z_n - e^{z_n}]/y_n!$ with rate parameter equal to e^{z_n}, the function f takes the following form:

$$f(\mathbf{m}, \mathbf{V}) = \sum_{n=1}^{N} \mathbb{E}_q[-\log p(y_n|z_n)] = \sum_{n=1}^{N}[-y_n m_n + e^{m_n + v_n/2} + \log(y_n!)] \tag{9}$$

whose derivative is not Lipschitz continuous since exponential is not Lipschitz.

1.2 Generalized Linear Models (GLMs)

We now describe a split for Generalized linear models. We model the output y_n by using an exponential family distribution whose natural-parameter is equal to $\eta_n := \mathbf{x}_n^T\mathbf{z}$. Assuming a standard

Gaussian prior over $\mathbf{z}$, the joint distribution can be written as follows:

$$p(\mathbf{y}, \mathbf{z}) := \prod_{n=1}^{N} p(y_n|\mathbf{x}_n^T\mathbf{z})\mathcal{N}(\mathbf{z}|0, \mathbf{I}) \tag{10}$$

A similar split can be obtained by putting non-conjugate terms $p(y_n|\mathbf{x}_n^T\mathbf{z})$ in $\tilde{p}_d$ and the rest in $\tilde{p}_e$:

$$\frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})} = \underbrace{\prod_{n=1}^{N} p(y_n|\mathbf{x}_n^T\mathbf{z})}_{\tilde{p}_d(\mathbf{z}|\boldsymbol{\lambda})} \underbrace{\frac{\mathcal{N}(\mathbf{z}|0, \mathbf{I})}{\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})}}_{\tilde{p}_e(\mathbf{z}|\boldsymbol{\lambda})}.$$

The lower bound can be shown to be the following:

$$\underline{\mathcal{L}}(\mathbf{m}, \mathbf{V}) := \underbrace{\sum_{n=1}^{N} \mathbb{E}_q[\log p(y_n|\mathbf{x}_n^T\mathbf{z})]}_{-f(\boldsymbol{\lambda})} - \underbrace{\mathbb{D}_{KL}[\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V}) \,\|\, \mathcal{N}(\mathbf{z}|0, \mathbf{I})]}_{h(\boldsymbol{\lambda})}. \tag{11}$$

which is very similar to the GP case. Therefore, Assumptions A1 and A2 will follow with similar arguments.

1.3 Correlated Topic Model (CTM)

We consider text documents with a vocabulary size N. Let $\mathbf{z}$ be a length K real-valued vector which follows a Gaussian distribution shown in (12). Given $\mathbf{z}$, a topic t_n is sampled for the n'th word using a multinomial distribution shown in (13). Probability of observing a word in the vocabulary is then given by (14).

$$p(\mathbf{z}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{12}$$

$$p(t_n = k|\mathbf{z}) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}, \tag{13}$$

$$p(\text{Observing a word v}|t_n, \boldsymbol{\theta}) = \beta_{v,t_n}. \tag{14}$$

Here $\boldsymbol{\beta}$ is a $N \times K$ real-valued matrix with non-negative entries and columns that sum to 1. The parameter set for this model is given by $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}\}$. We can marginalize out t_n and obtain the data-likelihood given $\mathbf{z}$,

$$p(\text{Observing a word v}|\mathbf{z}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p(\text{Observing a word v}|t_n = k, \boldsymbol{\theta})p(t_n = k|\mathbf{z}), \tag{15}$$

$$= \sum_{k=1}^{K} \beta_{vk} \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}. \tag{16}$$

Given that we observe n'th word y_n times, we can write the following joint distribution:

$$p(\mathbf{y}, \mathbf{z}) := \prod_{n=1}^{N} \left[\sum_{k=1}^{K} \beta_{n,k} \frac{e^{z_k}}{\sum_j e^{z_j}} \right]^{y_n} \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{17}$$

We can then use the following split:

$$\frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})} = \underbrace{\prod_{n=1}^{N} \left[\sum_{k=1}^{K} \beta_{n,k} \frac{e^{z_k}}{\sum_j e^{z_j}} \right]^{y_n}}_{\tilde{p}_d(\mathbf{z}|\boldsymbol{\lambda})} \underbrace{\frac{\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})}}_{\tilde{p}_e(\mathbf{z}|\boldsymbol{\lambda})},$$

where $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ are parameters of the Gaussian prior and $\beta_{n,k}$ are parameters of K multinomials.

The lower bound is shown below:

$$\begin{aligned} \underline{\mathcal{L}}(\mathbf{m}, \mathbf{V}) := \sum_{n=1}^{N} y_n \left\{ \mathbb{E}_q \left[\log \left(\sum_{k=1}^{K} \beta_{n,k} e^{z_k} \right) \right] \right\} - W \mathbb{E}_q \left\{ \log \left[\sum_{j=1}^{K} e^{z_j} \right] \right\} \\ - \mathbb{D}_{KL}[\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V}) \,\|\, \mathcal{N}(\mathbf{z}|0, \mathbf{I})]. \end{aligned} \tag{18}$$

where $W = \sum_n y_n$ is the total number of words. The top line is the function $[-f(\boldsymbol{\lambda})]$ while the bottom line is $[-h(\boldsymbol{\lambda})]$.

There are two intractable expectations in f, each involving expectation of a log-sum-exp function. Wang and Blei (2013) use the Delta method and Laplace method to approximate these expectations. In contrast, in PG-SVI algorithm, we use Monte Carlo to approximate the gradient of these functions.

2 Proof of Proposition 1 and 2

We first prove the Proposition 2. Proposition 1 is obtained as a special case of it. Our proof technique is borrowed from Ghadimi et. al. (2014). We extend their results to general divergence functions.

We denote the proximal projection at $\boldsymbol{\lambda}_k$ with gradient $\mathbf{g}$ and step-size β by,

$$\mathcal{P}(\boldsymbol{\lambda}_k, \mathbf{g}, \beta) := \frac{1}{\beta}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}), \tag{19}$$

$$\text{where } \boldsymbol{\lambda}_{k+1} = \arg\min_{\boldsymbol{\lambda} \in \mathcal{S}} \; \boldsymbol{\lambda}^T \mathbf{g} + h(\boldsymbol{\lambda}) + \frac{1}{\beta} \mathbb{D}(\boldsymbol{\lambda} \,\|\, \boldsymbol{\lambda}_k). \tag{20}$$

The following lemma gives a bound on the norm of $\mathcal{P}(\boldsymbol{\lambda}_k, \mathbf{g}, \beta)$.

Lemma 1. *The following holds for any $\boldsymbol{\lambda}_k \in \mathcal{S}$, any real-valued vector $\mathbf{g}$ and $\beta > 0$.*

$$\mathbf{g}^T \mathcal{P}(\boldsymbol{\lambda}_k, \mathbf{g}, \beta) \geq \alpha ||\mathcal{P}(\boldsymbol{\lambda}_k, \mathbf{g}, \beta)||^2 + \frac{1}{\beta}[h(\boldsymbol{\lambda}_{k+1}) - h(\boldsymbol{\lambda}_k)] \tag{21}$$

Proof. The gradient of the right hand side of (20) is given as follows:

$$\mathbf{g} + \bigtriangledown h(\boldsymbol{\lambda}) + \frac{1}{\beta} \bigtriangledown_{\lambda} \mathbb{D}(\boldsymbol{\lambda} \,\|\, \boldsymbol{\lambda}_k). \tag{22}$$

We use this to derive the optimality condition of (20). For any $\boldsymbol{\lambda}$, the following holds from the optimality condition:

$$(\boldsymbol{\lambda} - \boldsymbol{\lambda}_{k+1})^T \left[\mathbf{g} + \bigtriangledown h(\boldsymbol{\lambda}_{k+1}) + \frac{1}{\beta} \bigtriangledown_{\lambda} \mathbb{D}(\boldsymbol{\lambda}_{k+1} \,\|\, \boldsymbol{\lambda}_k) \right] \geq 0. \tag{23}$$

Letting $\boldsymbol{\lambda} = \boldsymbol{\lambda}_k$,

$$(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1})^T \left[\mathbf{g} + \bigtriangledown h(\boldsymbol{\lambda}_{k+1}) + \frac{1}{\beta} \bigtriangledown_{\lambda} \mathbb{D}(\boldsymbol{\lambda}_{k+1} \,\|\, \boldsymbol{\lambda}_k) \right] \geq 0, \tag{24}$$

which implies,

$$\begin{aligned} \mathbf{g}^T(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}) &\geq \frac{1}{\beta}(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k)^T \bigtriangledown_{\lambda} \mathbb{D}(\boldsymbol{\lambda}_{k+1} \,\|\, \boldsymbol{\lambda}_k) + h(\boldsymbol{\lambda}_{k+1}) - h(\boldsymbol{\lambda}_k), & (25) \\ &\geq \frac{\alpha}{\beta} ||\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k||^2 + h(\boldsymbol{\lambda}_{k+1}) - h(\boldsymbol{\lambda}_k). & (26) \end{aligned}$$

The first line follows from Assumption A2 (convexity of h), and the second line follows from Assumption A6. □

Now, we are ready to prove Proposition 2:

Proof. Let $\tilde{g}_{\lambda,k} := \mathcal{P}(\boldsymbol{\lambda}_k, \nabla f(\boldsymbol{\lambda}_k), \beta_k)$. Since f is L-smooth (Assumption A1), for any $k = 0, 1, \ldots, t-1$ we have,

$$\begin{aligned} f(\boldsymbol{\lambda}_{k+1}) &\leq f(\boldsymbol{\lambda}_k) + \langle \bigtriangledown f(\boldsymbol{\lambda}_k), \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k \rangle + \frac{L}{2} \|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|^2, \\ &= f(\boldsymbol{\lambda}_k) - \beta_k \langle \bigtriangledown f(\boldsymbol{\lambda}_k), \tilde{g}_{\lambda,k} \rangle + \frac{L}{2} \beta_k^2 \|\tilde{g}_{\lambda,k}\|^2, \\ &\leq f(\boldsymbol{\lambda}_k) - \beta_k \alpha \|\tilde{g}_{\lambda,k}\|^2 - [h(\boldsymbol{\lambda}_{k+1}) - h(\boldsymbol{\lambda}_k)] + \frac{L}{2} \beta_k^2 \|\tilde{g}_{\lambda,k}\|^2. \end{aligned}$$

The second line follows from the definition of $\mathcal{P}$ and the last line is due to Lemma 1. Rearranging the terms we get:

$$\begin{aligned} -\underline{\mathcal{L}}(\boldsymbol{\lambda}_{k+1}) + \underline{\mathcal{L}}(\boldsymbol{\lambda}_k) &\leq -[\beta_k \alpha - \frac{L}{2} \beta_k^2] \|\tilde{g}_{\lambda,k}\|^2, \\ \Rightarrow \qquad \underline{\mathcal{L}}(\boldsymbol{\lambda}_{k+1}) - \underline{\mathcal{L}}(\boldsymbol{\lambda}_k) &\geq [\beta_k \alpha - \frac{L}{2} \beta_k^2] \|\tilde{g}_{\lambda,k}\|^2. \end{aligned}$$

Summing these term for all $k = 0, 1, \ldots t-1$, we get the following:

$$\underline{\mathcal{L}}(\boldsymbol{\lambda}_{t-1}) - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0) \geq \sum_{k=0}^{t-1} [\beta_k \alpha - \frac{L}{2}\beta_k^2] \|\tilde{g}_{\lambda,k}\|^2.$$

By noting that the global maximum of the lower bound always upper bounds any other value, we get $\underline{\mathcal{L}}(\boldsymbol{\lambda}_*) - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0) \geq \underline{\mathcal{L}}(\boldsymbol{\lambda}_{t-1}) - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0)$. Using this,

$$\underline{\mathcal{L}}(\boldsymbol{\lambda}_*) - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0) \geq \sum_{k=0}^{t-1} [\beta_k \alpha - \frac{L}{2}\beta_k^2] \|\tilde{g}_{\lambda,k}\|^2,$$

$$\Rightarrow \quad \min_{k=0,1,\ldots,t-1} \|\tilde{g}_{\lambda,k}\|^2 [\sum_{k=0}^{t-1} [\beta_k \alpha - \frac{L}{2}\beta_k^2]] \leq \underline{\mathcal{L}}(\boldsymbol{\lambda}_*) - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0).$$

Since we assume at least one of $\beta_k < 2\alpha/L$, we can divide by the summation term, to get the following:

$$\min_{k=0,1,\ldots,t-1} \|\tilde{g}_{\lambda,k}\|^2 \leq \frac{\underline{\mathcal{L}}(\boldsymbol{\lambda}_*) - \underline{\mathcal{L}}(\boldsymbol{\lambda}_1)}{\sum_{k=0}^{t-1} [\beta_k \alpha - \frac{L}{2}\beta_k^2]},$$

which proves the Proposition 2. □

Proposition 1 can be obtained by simply plugging in $\beta_k = \alpha/L$.

Proof.

$$\min_{k=0,1,\ldots,t-1} \|\tilde{g}_{\lambda,k}\|^2 \leq \frac{C_0}{\sum_{k=0}^{t-1} [\frac{\alpha^2}{L} - \frac{\alpha^2}{2L}]} = \frac{2C_0 L}{\alpha^2 t}.$$

Expanding the left hand side, we get the required result:

$$\min_{k=0,1,\ldots,t} \|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|^2 \leq \beta_k \frac{2C_0 L}{\alpha^2 t} = \frac{2C_0}{\alpha t}.$$

□

3 Proof of Proposition 3

We will first prove the following theorem, which gives a similar result to Proposition 2 but for a stochastic gradient $\widehat{\nabla} f$.

Theorem 1. *If we choose the step-size β_k such that $0 < \beta_k \leq 2\alpha_*/L$ with $\beta_k < 2\alpha_*/L$ for at least one k, then,*

$$\mathbb{E}_{R,\boldsymbol{\xi}} \left(\frac{\|\boldsymbol{\lambda}_{R+1} - \boldsymbol{\lambda}_R\|^2}{\beta_R} \right) \leq \frac{C_0 + \frac{c\sigma^2}{2} \sum_{k=0}^{t-1} \frac{\beta_k}{M_k}}{\sum_{k=0}^{t-1} \left(\alpha_* \beta_k - L\beta_k^2/2 \right)}. \tag{27}$$

where the expectation is taken over $R \in \{0, 1, 2, \ldots, t-1\}$ which is a discrete random variable drawn from the probability mass function

$$Prob(R = k) = \frac{\alpha_* \beta_k - L\beta_k^2/2}{\sum_{k=0}^{t-1} \left(\alpha_* \beta_k - L\beta_k^2/2\right)},$$

and over $\boldsymbol{\xi} := \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_{t-1}\}$ with $\boldsymbol{\xi}_k$ is the noise in the stochastic approximation $\widehat{\nabla} f$.

Proof. Let $\tilde{g}_{\lambda,k} := \mathcal{P}(\boldsymbol{\lambda}_k, \widehat{\nabla} f(\lambda_k), \beta_k)$, $\delta_k := \widehat{\nabla} f(\lambda_k) - \bigtriangledown f(\boldsymbol{\lambda}_k)$. Since f is L-smooth, for any $k = 0, 1, \ldots, t$ we have,

$$f(\boldsymbol{\lambda}_{k+1}) \leq f(\boldsymbol{\lambda}_k) + \langle \bigtriangledown f(\boldsymbol{\lambda}_k), \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k \rangle + \frac{L}{2} \|\lambda_{k+1} - \boldsymbol{\lambda}_k\|^2 \tag{28}$$

$$= f(\boldsymbol{\lambda}_k) - \beta_k \langle \bigtriangledown f(\boldsymbol{\lambda}_k), \tilde{g}_{\lambda,k} \rangle + \frac{L}{2} \beta_k^2 \|\tilde{g}_{\lambda,k}\|^2 \tag{29}$$

$$= f(\boldsymbol{\lambda}_k) - \beta_k \left\langle \widehat{\nabla} f(\lambda_k), \tilde{g}_{\lambda,k} \right\rangle + \frac{L}{2} \beta_k^2 \|\tilde{g}_{\lambda,k}\|^2 + \beta_k \langle \delta_k, \tilde{g}_{\lambda,k} \rangle \tag{30}$$

where we have used the definition of $\tilde{g}_{\lambda,k}$ and δ_k. Now using Lemma 1 on the second term and Cauchy-Schwarz for the last term, we get the following:

$$f(\boldsymbol{\lambda}_{k+1}) \leq f(\boldsymbol{\lambda}_k) - \left[\alpha \beta_k \|\tilde{g}_{\lambda,k}\|^2 + h(\boldsymbol{\lambda}_{k+1}) - h(\boldsymbol{\lambda}_k)\right] + \frac{L}{2} \beta_k^2 \|\tilde{g}_{\lambda,k}\|^2 + \beta_k \|\delta_k\| \|\tilde{g}_{\lambda,k}\| \tag{31}$$

After rearranging and using Young's inequality $\|\delta_k\| \|\tilde{g}_{\lambda,k}\| \leq (c/2)\|\delta_k\|^2 + 1/(2c)\|\tilde{g}_{\lambda,k}\|^2$ given a constant $c > 0$, we get

$$-\underline{\mathcal{L}}(\boldsymbol{\lambda}_{k+1}) \leq -\underline{\mathcal{L}}(\boldsymbol{\lambda}_k) - \alpha \beta_k \|\tilde{g}_{\lambda,k}\|^2 + \frac{L}{2} \beta_k^2 \|\tilde{g}_{\lambda,k}\|^2 + \frac{\beta_k}{2c} \|\tilde{g}_{\lambda,k}\|^2 + \frac{\beta_k c}{2} \|\delta_k\|^2 \tag{32}$$

$$= -\underline{\mathcal{L}}(\boldsymbol{\lambda}_k) - \left((\alpha - 1/(2c))\beta_k - \frac{L}{2}\beta_k^2\right) \|\tilde{g}_{\lambda,k}\|^2 + \frac{c\beta_k}{2} \|\delta_k\|^2 \tag{33}$$

Now considering $c > 1/(2\alpha)$, $\alpha_* = \alpha - 1/(2c)$ and $\beta_k \leq \frac{2\alpha_*}{L}$, and summing up both side for iteration $k = 0, 1 \ldots, t-1$, we obtain

$$\sum_{k=0}^{t-1} \left(\alpha_* \beta_k - \frac{L}{2}\beta_k^2\right) \|\tilde{g}_{\lambda,k}\|^2 \leq \underline{\mathcal{L}}^* - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0) + \sum_{k=0}^{t-1} \frac{c\beta_k}{2} \|\delta_k\|^2 \tag{34}$$

Now by taking expectation w.r.t. $\boldsymbol{\xi}$ on both side and using the fact that $\mathbb{E}_{\boldsymbol{\xi}} \|\delta_k\|^2 \leq \frac{\sigma^2}{M_k}$ by assumption $A3$ and $A4$, we get

$$\sum_{k=0}^{t-1} \left(\alpha_* \beta_k - \frac{L}{2}\beta_k^2\right) \mathbb{E}_{\boldsymbol{\xi}} \|\tilde{g}_{\lambda,k}\|^2 \leq C_0 + \frac{c\sigma^2}{2} \sum_{k=0}^{t} \frac{\beta_k}{M_k} \tag{35}$$

Noting that the expectation w.r.t. R and $\boldsymbol{\xi}$ can be written in terms of expectation w.r.t. $\boldsymbol{\xi}$ alone,

$$\mathbb{E}_{R,\xi}[\|\tilde{g}_{\lambda_k,R}\|^2] = \frac{\sum_{k=0}^{t}\left(\alpha_*\beta_k - \frac{L}{2}\beta_k^2\right)\mathbb{E}_{\xi}\|\tilde{g}_{\lambda,k}\|^2}{\sum_{k=0}^{t}\left(\alpha_*\beta_k - \frac{L}{2}\beta_k^2\right)} \tag{36}$$

whose numerator is basically the term in the left hand side of (35). Dividing (35) by $\sum_{k=0}^{t}\left(\alpha_*\beta_k - \frac{L}{2}\beta_k^2\right)$, we get the required result. □

By substituting $\beta_k = \gamma\alpha_*/L$ for $0 < \gamma < 2$ and $M_k = M$ in (27),

$$\mathbb{E}_{R,\boldsymbol{\xi}}\left(\|\boldsymbol{\lambda}_{R+1} - \boldsymbol{\lambda}_R\|^2\right) \leq \frac{\gamma\alpha_*}{L}\frac{C_0 + \frac{c\sigma^2}{2}\sum_{k=0}^{t-1}\frac{\beta_k}{M_k}}{\sum_{k=1}^{t-1}\left(\alpha_*\beta_k - L\beta_k^2/2\right)} \tag{37}$$

$$= \frac{\gamma\alpha_*}{L}\frac{C_0 + \frac{c\sigma^2\alpha_* t}{2LM}}{\frac{\alpha_*^2 t\gamma(2-\gamma)}{2L}} = \frac{1}{2-\gamma}\left(\frac{2C_0}{\alpha_* t} + \frac{\gamma c\sigma^2}{ML}\right) \tag{38}$$

The probability distribution of R also reduces to a uniform distribution with probability of each iteration being $1/t$. This proves Proposition 3.

4 Derivation of Closed-Form Updates for the GP Model

The PG-SVI iterations $\boldsymbol{\lambda}_{k+1} = \min_{\lambda\in\mathcal{S}} \boldsymbol{\lambda}^T\left[\widehat{\bigtriangledown} f(\boldsymbol{\lambda}_k)\right] + h(\boldsymbol{\lambda}) + \frac{1}{\beta_k}\mathbb{D}(\boldsymbol{\lambda}\|\boldsymbol{\lambda}_k)$ takes the following form for the GP model, as discussed in Section 6 of the main paper:

$$(\mathbf{m}_{k+1}, \mathbf{V}_{k+1}) = \arg\min_{\mathbf{m},\mathbf{V}\succ 0} \left(m_n\alpha_{n_k,k} + \tfrac{1}{2}v_n\gamma_{n_k,k}\right) + D_{KL}\left[\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})||\mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] + \frac{1}{\beta_k}D_{KL}\left[\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})||\mathcal{N}(\mathbf{z}|\mathbf{m}_k,\mathbf{V}_k)\right]. \tag{39}$$

where n_k is the example selected in k'th iteration. We will now show that its solution can be obtained in closed-form.

4.1 Full Update of $\mathbf{V}_{k+1}$

We first derive the full update of $\mathbf{V}_{k+1}$. The KL divergence between two Gaussian distributions is given as follows:

$$D_{KL}\left[\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})||\mathcal{N}(\mathbf{z}|0,\mathbf{K})\right] = -\tfrac{1}{2}[\log|\mathbf{V}\mathbf{K}^{-1}| - \mathrm{Tr}(\mathbf{V}\mathbf{K}^{-1}) - \mathbf{m}^T\mathbf{K}^{-1}\mathbf{m} + D] \tag{40}$$

Using this, we expand the last two terms of (39) to get the following,

$$
\begin{aligned}
&-\tfrac{1}{2}\left[\log|\mathbf{V}\mathbf{K}^{-1}| - \text{Tr}(\mathbf{V}\mathbf{K}^{-1}) - \mathbf{m}^T\mathbf{K}^{-1}\mathbf{m} + D\right] \\
&-\tfrac{1}{2}\frac{1}{\beta_k}\left[\log|\mathbf{V}\mathbf{K}^{-1}| - \text{Tr}\{\mathbf{V}\mathbf{V}_k^{-1}\} - (\mathbf{m}-\mathbf{m}_k)^T\mathbf{V}_k^{-1}(\mathbf{m}-\mathbf{m}_k) + D\right] \\
&= -\tfrac{1}{2}\left[\left(1+\frac{1}{\beta_k}\right)\log|\mathbf{V}| - \text{Tr}\{\mathbf{V}(\mathbf{K}^{-1}+\frac{1}{\beta_k}\mathbf{V}_k^{-1})\} - \mathbf{m}^T\mathbf{K}^{-1}\mathbf{m}\right. \\
&\qquad \left. -\frac{1}{\beta_k}(\mathbf{m}-\mathbf{m}_k)^T\mathbf{V}_k^{-1}(\mathbf{m}-\mathbf{m}_k) + \left(1+\frac{1}{\beta_k}\right)(D-\log|\mathbf{K}|)\right] \qquad (41)
\end{aligned}
$$

Taking derivative of (39) with respect to $\mathbf{V}$ at $\mathbf{V} = \mathbf{V}_{k+1}$ and setting it to zero, we get the following (here $\mathbf{I}_n$ is a matrix with all zeros, except the n'th diagonal element which is set to 1):

$$
\begin{aligned}
&\Rightarrow \quad -\left(1+\frac{1}{\beta_k}\right)\mathbf{V}_{k+1}^{-1} + \left(\mathbf{K}^{-1}+\frac{1}{\beta_k}\mathbf{V}_k^{-1}\right) + \gamma_{n_k,k}\mathbf{I}_{n_k} = 0 \qquad (42) \\
&\Rightarrow \quad \mathbf{V}_{k+1}^{-1} = \frac{1}{1+\beta_k}\mathbf{V}_k^{-1} + \frac{\beta_k}{1+\beta_k}\left(\mathbf{K}^{-1}+\gamma_{n_k,k}\mathbf{I}_{n_k}\right) \qquad (43) \\
&\Rightarrow \quad \mathbf{V}_{k+1}^{-1} = r_k\mathbf{V}_k^{-1} + (1-r_k)\left(\mathbf{K}^{-1}+\gamma_{n_k,k}\mathbf{I}_{n_k}\right) \qquad (44)
\end{aligned}
$$

which gives us the update of $\mathbf{V}_{k+1}$ for $r_k := 1/(1+\beta_k)$.

4.2 Avoiding a full update of $\mathbf{V}_{k+1}$

A full update will require storing the matrix $\mathbf{V}_{k+1}$. Fortunately, we can avoid storing the full matrix and still do an exact update. The key point here is to notice that to compute the stochastic gradient in the next iteration we only need one diagonal element of $\mathbf{V}_{k+1}$ rather than the whole matrix. Specifically, if we sample n_{k+1}'th example at the iteration $k+1$, then we need to compute $v_{n_{k+1},k+1}$ which is the n_{k+1}'th diagonal element of $\mathbf{V}_{k+1}$. This can be done by solving one linear equation, as we show in this section. Specifically, we show that the following updates can be used to compute $v_{n_{k+1},k+1}$:

$$
v_{n_{k+1},k+1} = \kappa_{n_{k+1},n_{k+1}} - \boldsymbol{\kappa}_{n_{k+1}}^T\left(\mathbf{K} + [\text{diag}(\widetilde{\boldsymbol{\gamma}}_k)]^{-1}\right)^{-1}\boldsymbol{\kappa}_{n_{k+1}}, \qquad (45)
$$

where $\widetilde{\boldsymbol{\gamma}}_k = r_k\widetilde{\boldsymbol{\gamma}}_{k-1} + (1-r_k)\gamma_{n_k,k}\mathbf{1}_{n_k}$ ($\mathbf{1}_n$ is a vector of all zeros except its n'th entry which is equal to 1). We start the recursion with $\widetilde{\boldsymbol{\gamma}}_0 = \epsilon$ where ϵ is a small positive number.

We will now show that $\mathbf{V}_k$ can be reparameterized in terms of a vector $\widetilde{\boldsymbol{\gamma}}_k$ which contains accumulated weighted sum of the gradient $\gamma_{n_j,j}$, for all $j \leq k$. To show this, we recursively substitute the update of $\mathbf{V}_j$ for $j < k+1$, as shown below (recall that n_k is the example selected at the k'th iteration). The second line is obtained by substituting the full update of $\mathbf{V}_k$ by using (44). The third line is obtained after a few simplifications. The fourth line is obtained by substituting the

update of $\mathbf{V}_{k-1}$ and a few simplifications.

$$\mathbf{V}_{k+1}^{-1} = r_k \mathbf{V}_k^{-1} + (1 - r_k) \left[\mathbf{K}^{-1} + \gamma_{n_k,k} \mathbf{I}_{n_k}\right] \tag{46}$$

$$= r_k \left[r_{k-1} \mathbf{V}_{k-1}^{-1} + (1 - r_{k-1}) \left(\mathbf{K}^{-1} + \gamma_{n_{k-1},k-1} \mathbf{I}_{n_{k-1}}\right)\right] + (1 - r_k) \left[\mathbf{K}^{-1} + \gamma_{n_k,k} \mathbf{I}_{n_k}\right] \tag{47}$$

$$= r_k r_{k-1} \mathbf{V}_{k-1}^{-1} + (1 - r_k r_{k-1}) \mathbf{K}^{-1} + \left[r_k (1 - r_{k-1}) \gamma_{n_{k-1},k-1} \mathbf{I}_{n_{k-1}} + (1 - r_k) \gamma_{n_k,k} \mathbf{I}_{n_k}\right]$$

$$= r_k r_{k-1} r_{k-2} \mathbf{V}_{k-2}^{-1} + (1 - r_k r_{k-1} r_{k-2}) \mathbf{K}^{-1}$$

$$+ \left[r_k r_{k-1} (1 - r_{k-2}) \gamma_{n_{k-2},k-2} \mathbf{I}_{n_{k-2}} + r_k (1 - r_{k-1}) \gamma_{n_{k-1},k-1} \mathbf{I}_{n_{k-1}} + (1 - r_k) \gamma_{n_k,k} \mathbf{I}_{n_k}\right] \tag{48}$$

This update expresses $\mathbf{V}_{k+1}$ in terms of $\mathbf{V}_{k-2}$, $\mathbf{K}$, and gradients of the data example selected at $k, k-1$, and $k-2$. Continuing in this fashion until $k = 0$, we can write the update as follows:

$$\mathbf{V}_{k+1}^{-1} = t_k \mathbf{V}_0^{-1} + (1 - t_k) \mathbf{K}^{-1} + [r_k r_{k-1} \dots r_3 r_2 (1 - r_1) \gamma_{n_1,1} \mathbf{I}_{n_1}$$

$$+ r_k r_{k-1} \dots r_4 r_3 (1 - r_2) \gamma_{n_2,2} \mathbf{I}_{n_2} + r_k r_{k-1} \dots r_5 r_4 (1 - r_3) \gamma_{n_3,3} \mathbf{I}_{n_2} + \dots$$

$$+ r_k r_{k-1} (1 - r_{k-2}) \gamma_{n_{k-2},k-2} \mathbf{I}_{n_{k-2}} + r_k (1 - r_{k-1}) \gamma_{n_{k-1},k-1} \mathbf{I}_{n_{k-1}} + (1 - r_k) \gamma_{n_k,k} \mathbf{I}_{n_k}] \tag{49}$$

where t_k is the product of $r_k, r_{k-1}, \dots, r_0$. We can write the updates more compactly by defining the accumulation of the gradients $\gamma_{n_j,j}$ for all $j \leq k$ by a vector $\widetilde{\gamma}_k$,

$$\mathbf{V}_{k+1}^{-1} = t_k \mathbf{V}_0^{-1} + (1 - t_k) \mathbf{K}^{-1} + \text{diag}(\widetilde{\gamma}_k) \tag{50}$$

The vector $\widetilde{\gamma}_k$ can be obtained by using a recursion. We illustrate this below, where we have grouped the terms in (49) to show the recursion for $\widetilde{\gamma}_k$ (here $\mathbf{1}_n$ is a vector with all zero entries except n'th entry which is set to 1):

$$\begin{aligned}
&r_k r_{k-1} \dots r_6 r_5 r_4 r_3 r_2 (1 - r_1) \gamma_{n_1,1} \mathbf{1}_{n_1} &&= \widetilde{\gamma}_1 \\
+&r_k r_{k-1} \dots r_6 r_5 r_4 r_3 (1 - r_2) \gamma_{n_2,2} \mathbf{1}_{n_2} &&= \widetilde{\gamma}_2 \\
+&r_k r_{k-1} \dots r_6 r_5 r_4 (1 - r_3) \gamma_{n_3,3} \mathbf{1}_{n_3} &&= \widetilde{\gamma}_3 \\
+&r_k r_{k-1} \dots r_6 r_5 (1 - r_4) \gamma_{n_4,4} \mathbf{1}_{n_4} &&= \widetilde{\gamma}_4 \\
+&r_k r_{k-1} \dots r_6 (1 - r_5) \gamma_{n_5,5} \mathbf{1}_{n_5} &&= \widetilde{\gamma}_5 \\
&\vdots
\end{aligned}$$

Therefore, $\widetilde{\gamma}_k$ can be recursively updated as follows:

$$\widetilde{\gamma}_k = r_k \widetilde{\gamma}_{k-1} + (1 - r_k) \gamma_{n_k,k} \mathbf{1}_{n_k} \tag{51}$$

with an initialization $\widetilde{\gamma}_0 = \epsilon$ where ϵ is a small constant to avoid numerical issues.

If we set $\mathbf{V}_0 = \mathbf{K}$, then the formula simplifies to the following:

$$\mathbf{V}_{k+1}^{-1} = \mathbf{K}^{-1} + \text{diag}(\widetilde{\gamma}_k) \tag{52}$$

which is completely specified by $\widetilde{\gamma}_k$, eliminating the need to compute and store $\mathbf{V}_{k+1}$.

The n_{k+1}'th diagonal element can be obtained by using Matrix Inversion Lemma, which gives us the update (45).

4.3 Update of $\mathbf{m}$

Taking derivative of (39) with respect to $\mathbf{m}$ at $\mathbf{m} = \mathbf{m}_{k+1}$ and setting it to zero, we get the following (here $\mathbf{1}_n$ is a vector with all zero entries except n'th entry which is set to 1):

$$\Rightarrow \quad -\mathbf{K}^{-1}\mathbf{m}_{k+1} - \frac{1}{\beta_k}\mathbf{V}_k^{-1}(\mathbf{m}_{k+1} - \mathbf{m}_k) - \alpha_{n_k,k}\mathbf{1}_{n_k} = 0 \tag{53}$$

$$\Rightarrow \quad -\left[\mathbf{K}^{-1} + \frac{1}{\beta_k}\mathbf{V}_k^{-1}\right]\mathbf{m}_{k+1} + \frac{1}{\beta_k}\mathbf{V}_k^{-1}\mathbf{m}_k - \alpha_{n_k,k}\mathbf{1}_{n_k} = 0 \tag{54}$$

$$\Rightarrow \quad \mathbf{m}_{k+1} = \left[\mathbf{K}^{-1} + \frac{1}{\beta_k}\mathbf{V}_k^{-1}\right]^{-1}\left[\frac{1}{\beta_k}\mathbf{V}_k^{-1}\mathbf{m}_k - \alpha_{n_k,k}\mathbf{1}_{n_k}\right] \tag{55}$$

$$\Rightarrow \quad \mathbf{m}_{k+1} = \left[(1-r_k)\mathbf{K}^{-1} + r_k\mathbf{V}_k^{-1}\right]^{-1}\left[-(1-r_k)\alpha_{n_k,k}\mathbf{1}_{n_k} + r_k\mathbf{V}_k^{-1}\mathbf{m}_k\right] \tag{56}$$

where the last step is obtained using the fact that $1/\beta_k = r_k/(1-r_k)$.

We simplify as shown below. The second line is obtained by adding and subtracting $(1 - r_k)\mathbf{K}^{-1}\mathbf{m}_k$ in the square bracket at the right. In the the third line, we take $\mathbf{m}_k$ out. The fourth line is obtained by plugging in the updates of $\mathbf{V}_k^{-1} = \mathbf{K}^{-1} + \text{diag}(\widetilde{\boldsymbol{\gamma}}_k)$. The fifth line is obtained by using Matrix-Inversion lemma, and the sixth line is obtained by taking $\mathbf{K}^{-1}$ out of the right-most term.

$$\mathbf{m}_{k+1} = \left[(1-r_k)\mathbf{K}^{-1} + r_k\mathbf{V}_k^{-1}\right]^{-1}\left[-(1-r_k)\alpha_{n_k,k}\mathbf{1}_{n_k} + r_k\mathbf{V}_k^{-1}\mathbf{m}_k\right] \tag{57}$$

$$= \left[(1-r_k)\mathbf{K}^{-1} + r_k\mathbf{V}_k^{-1}\right]^{-1}\left[(1-r_k)\{-\mathbf{K}^{-1}\mathbf{m}_k - \alpha_{n_k,k}\mathbf{1}_{n_k}\} + \{(1-r_k)\mathbf{K}^{-1} + r_k\mathbf{V}_k^{-1}\}\mathbf{m}_k\right]$$

$$= \mathbf{m}_k + (1-r_k)\left[(1-r_k)\mathbf{K}^{-1} + r_k\mathbf{V}_k^{-1}\right]^{-1}\left(-\mathbf{K}^{-1}\mathbf{m}_k - \alpha_{n_k,k}\mathbf{1}_{n_k}\right) \tag{58}$$

$$= \mathbf{m}_k - (1-r_k)\left[\mathbf{K}^{-1} + r_k\text{diag}(\widetilde{\boldsymbol{\gamma}}_{k-1})\right]^{-1}\left(\mathbf{K}^{-1}\mathbf{m}_k + \alpha_{n_k,k}\mathbf{1}_{n_k}\right) \tag{59}$$

$$= \mathbf{m}_k - (1-r_k)\left[\mathbf{K} - \mathbf{K}\left(\mathbf{K} + \text{diag}(r_k\widetilde{\boldsymbol{\gamma}}_{k-1})^{-1}\right)^{-1}\mathbf{K}\right]\left(\mathbf{K}^{-1}\mathbf{m}_k + \alpha_{n_k,k}\mathbf{1}_{n_k}\right) \tag{60}$$

$$= \mathbf{m}_k - (1-r_k)\left[\mathbf{I} - \mathbf{K}\left(\mathbf{K} + \text{diag}(r_k\widetilde{\boldsymbol{\gamma}}_{k-1})^{-1}\right)^{-1}\right](\mathbf{m}_k + \alpha_{n_k,k}\boldsymbol{\kappa}_{n_k}) \tag{61}$$

$$= \mathbf{m}_k - (1-r_k)(\mathbf{I} - \mathbf{K}\mathbf{B}_k^{-1})(\mathbf{m}_k + \alpha_{n_k,k}\boldsymbol{\kappa}_{n_k}) \tag{62}$$

where $\mathbf{B}_k := \mathbf{K} + [\text{diag}(r_k\widetilde{\boldsymbol{\gamma}}_{k-1})]^{-1}$.

Since $r_k\widetilde{\boldsymbol{\gamma}}_{k-1}$ and $\widetilde{\boldsymbol{\gamma}}_k$ differ only slightly (by the new example gradient γ_{n_k}, we can instead use the following approximate update:

$$\mathbf{m}_{k+1} = \mathbf{m}_k - (1-r_k)(\mathbf{I} - \mathbf{K}\mathbf{A}_k^{-1})(\mathbf{m}_k + \alpha_{n_k,k}\boldsymbol{\kappa}_{n_k}) \tag{63}$$

where $\mathbf{A}_k := \mathbf{K} + [\text{diag}(\widetilde{\boldsymbol{\gamma}}_k)]^{-1}$.

5 Closed-Form Updates for GLMs

We rewrite the lower bound as

$$-\underline{\mathcal{L}}(\mathbf{m},\mathbf{V}) := \underbrace{\sum_{n=1}^{N} f_n(\widetilde{m}_n,\widetilde{v}_n)}_{f(\boldsymbol{m},\boldsymbol{V})} + \underbrace{\mathbb{D}_{KL}[\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})\,\|\,\mathcal{N}(\mathbf{z}|0,\mathbf{I})]}_{h(\mathbf{m},\mathbf{V})} \tag{64}$$

where $f_n(\widetilde{m}_n,\widetilde{v}_n) := -\mathbb{E}_q[\log p(y_n|\mathbf{x}_n^T\mathbf{z})]$ with $\widetilde{m}_n := \mathbf{x}_n^T$ and $\widetilde{v}_n := \mathbf{x}_n^T\mathbf{V}\mathbf{x}_n$. We can compute a stochastic approximation to the gradient of f by randomly selecting an example n_k (choosing $M=1$) and using a Monte Carlo gradient approximation to the gradient of f_{n_k}. Similar to GP, we define the following as our gradients of function f_n:

$$\alpha_{n_k,k} := N\nabla_{\widetilde{m}_{n_k}} f_{n_k}(\widetilde{m}_{n_k},\widetilde{v}_{n_k}), \quad \gamma_{n_k,k} := 2N\nabla_{\widetilde{v}_{n_k}} f_{n_k}(\widetilde{m}_{n_k},\widetilde{v}_{n_k}) \tag{65}$$

The PG-SVI iteration can be written as follows:

$$\begin{aligned}(\mathbf{m}_{k+1},\mathbf{V}_{k+1}) = \arg\min_{\mathbf{m},\mathbf{V}\succ 0}\ & (\widetilde{m}_n\alpha_{n_k,k} + \tfrac{1}{2}\widetilde{v}_n\gamma_{n_k,k}) + D_{KL}\left[\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})||\mathcal{N}(\mathbf{z}|0,\mathbf{I})\right] \\ & + \frac{1}{\beta_k} D_{KL}\left[\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})||\mathcal{N}(\mathbf{z}|\mathbf{m}_k,\mathbf{V}_k)\right].\end{aligned} \tag{66}$$

Using a similar derivation to the GP model, we can show that the following updates will give us the solution:

$$\begin{aligned}\widetilde{\boldsymbol{\gamma}}_k &= r_k\widetilde{\boldsymbol{\gamma}}_{k-1} + (1-r_k)\gamma_{n_k,k}\mathbf{1}_{n_k}, \\ \widetilde{\mathbf{m}}_{k+1} &= \widetilde{\mathbf{m}}_k - (1-r_k)(\mathbf{I} - \mathbf{K}\mathbf{A}_k^{-1})(\mathbf{m}_k + \alpha_{n_k,k}\boldsymbol{\kappa}_{n_k}), \\ \widetilde{v}_{n_{k+1},k+1} &= \kappa_{n_{k+1},n_{k+1}} - \boldsymbol{\kappa}_{n_{k+1}}^T\mathbf{A}_k^{-1}\boldsymbol{\kappa}_{n_{k+1}},\end{aligned} \tag{67}$$

where $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ and $\widetilde{\mathbf{m}}_k := \mathbf{X}^T\mathbf{m}$.

6 Description of the Dataset for Binary GP Classification

	Sonar	Ionosphere	USPS
# of data points	208	351	1,781
# of features	60	34	256
# of training data points	165	280	884

7 Description of Algorithms for Binary GP Classification

We give implementation details of all the algorithms used for binary GP- classification experiment. For all methods, we compute a stochastic estimate of the gradient by using a mini-batch size of

5, 5, and 20 for the three datasets: Sonar, Ionosphere, and USPS-3vs5 respectively. Similarly, the number of MC samples used are 2000, 500, and 2000.

For GD, SGD, and all the adaptive methods, $\boldsymbol{\lambda} := \{\mathbf{m}, \mathbf{L}\}$ where $\mathbf{L}$ is the Cholesky factor of $\mathbf{V}$. The algorithmic parameters of these methods is given in Table 1. Below, we give details of their updates.

For the GD method, we use the following update:

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \alpha \nabla \underline{\mathcal{L}}(\boldsymbol{\lambda}_k), \tag{68}$$

where α is a fixed step-size.

For the SGD method, we use a stochastic gradient, instead of the exact gradient:

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \alpha_k \mathbf{g}_k, \tag{69}$$

where $\alpha_k = (k+1)^{-\kappa}$ is the step-size and $\mathbf{g}_k := -\widehat{\nabla} \underline{\mathcal{L}}(\boldsymbol{\lambda}_k)$.

We use the following updates for ADAGRAD:

$$\mathbf{s}_k = \mathbf{s}_{k-1} + (\mathbf{g}_k \odot \mathbf{g}_k), \tag{70}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \alpha_0 \left[\frac{1}{\sqrt{\mathbf{s}_k + \epsilon}}\right] \odot \mathbf{g}_k. \tag{71}$$

where α_0 is a fixed step-size and ϵ is a small constant used to avoid numerical errors.

We use the following update for RMSprop:

$$\mathbf{s}_k = \rho \mathbf{s}_{k-1} + (1-\rho)(\mathbf{g}_k \odot \mathbf{g}_k), \tag{72}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \alpha_0 \left[\frac{1}{\sqrt{\mathbf{s}_k + \epsilon}}\right] \odot \mathbf{g}_k, \tag{73}$$

where α_0 is a fixed step-size and ρ is the decay factor.

We use the following updates for ADADELTA:

$$\mathbf{s}_k = \rho \mathbf{s}_{k-1} + (1-\rho)(\mathbf{g}_k \odot \mathbf{g}_k), \tag{74}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \mathbf{g}_k^{AD}, \quad \text{where } \mathbf{g}_k^{AD} = \alpha_0 \left(\frac{\sqrt{\boldsymbol{\delta}_k + \epsilon}}{\sqrt{\mathbf{s}_k + \epsilon}}\right) \odot \mathbf{g}_k, \tag{75}$$

$$\boldsymbol{\delta}_{k+1} = \rho \boldsymbol{\delta}_k + (1-\rho)\left(\mathbf{g}_k^{AD} \odot \mathbf{g}_k^{AD}\right). \tag{76}$$

where again α_0 is a fixed step-size, and ρ is the decay factor.

Finally, the updates for ADAM are shown below:

$$\boldsymbol{\mu}_k = \rho_\mu \boldsymbol{\mu}_{k-1} + (1-\rho_\mu)\mathbf{g}_k, \tag{77}$$

$$\mathbf{s}_k = \rho_s \mathbf{s}_{k-1} + (1-\rho_s)(\mathbf{g}_k \odot \mathbf{g}_k), \tag{78}$$

$$\mathbf{g}_{s,k} = \sqrt{\frac{\mathbf{s}_k}{1-\rho_s^k}}, \tag{79}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \alpha_0 \left[\frac{1}{\mathbf{g}_{s,k} + \epsilon}\right] \odot \left[\frac{\boldsymbol{\mu}_k}{1-\rho_\mu^k}\right]. \tag{80}$$

Table 1: Algorithmic parameters for Binary GP classification experiment (Figure 2 in the main paper). N is the number of training examples.

Parameter	Sonar	Ionosphere	USPS
SGD			
κ	0.8	0.51	0.6
$\alpha_0 \times N$	1200	25	800
ADAGRAD			
α_0	4.5	4	8
RMSprop			
α_0	0.1	0.04	0.1
ρ	0.9	0.9999	0.9
ADADELTA			
α_0	1.0	0.1	1.0
$1 - \rho$	5×10^{-10}	10^{-11}	10^{-12}
ADAM			
α_0	0.04	0.25	2.5
ρ_μ	0.9	0.9	0.9
ρ_s	0.999	0.999	0.999
PG-SVI			
$\beta_k \times N$	0.2	2.0	2.5

where α_0 is a fixed step-size and ρ_μ, ρ_s are decay factors.