Appendix of Distributed Asynchronous Variational Gaussian Processes

A Derivatives

A.1 Objective

As described in the paper, the objective function to be minimized is $-\mathcal{L} = \sum_{i=1}^{n} g_i + h$, where

$$\begin{aligned} g_i &= -\ln \mathcal{N}(y_n|\boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\mu}, \beta^{-1}) + \frac{\beta}{2}\boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\phi}_i + \frac{\beta}{2}\tilde{k}_{ii} \\ &= \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\beta + \frac{\beta}{2}\left(y_i^2 - 2y_i\boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\phi}_i + k_{ii} - \boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\phi}_i\right), \qquad (1) \\ h &= \mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w})) \\ &= \frac{1}{2}\left(-\ln|\boldsymbol{\Sigma}| - m + \mathrm{tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\right), \qquad (2) \end{aligned}$$

and we define $\beta = \sigma^{-2}$ and $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$.

A.2 Kernel

A common choice for the kernel is the anisotropic squared exponential covariance function:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = a_0^2 \exp\big(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \mathrm{diag}(\boldsymbol{\eta})(\boldsymbol{x}_i - \boldsymbol{x}_j)\big), \qquad (3)$$

in which the hyperparameters are the signal variance a_0 and the lengthscales $\boldsymbol{\eta} = \{1/a_k^2\}_{k=1}^d$, controlling how fast the covariance decays with the distance between inputs. Using this covariance function, we can prune input dimensions by shrinking the corresponding lengthscales based on the data (when $\eta_d = 0$, the d-th dimension becomes totally irrelevant to the covariance function value). This pruning is known as Automatic Relevance Determination (ARD) and therefore this covariance is also called the ARD squared exponential.

A.3 Derivative over $\ln\sigma$ $(\ln\beta^{-1/2})$

The derivative of g_i over $\ln\sigma$ is

$$\frac{\partial g_i}{\partial \ln\sigma} = 1 - \frac{1}{\sigma^2}(y_i^2 - 2y_i\boldsymbol{\phi}_i^T\boldsymbol{\mu} + \boldsymbol{\phi}_i^T(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{\phi}_i + k_{ii} - \boldsymbol{\phi}_i^T\boldsymbol{\phi}_i). \qquad (4)$$

A.4 Derivative over $\ln a_0$

The derivative of g_i over $\ln a_0$ is

$$\frac{\partial g_i}{\partial \ln a_0} = \frac{1}{\sigma^2}(-y_i\boldsymbol{\phi}_i^T\boldsymbol{\mu} + \boldsymbol{\phi}_i^T(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{\phi}_i + k_{ii} - \boldsymbol{\phi}_i^T\boldsymbol{\phi}_i). \qquad (5)$$

A.5 Derivative over $\boldsymbol{Z}$

By defining $\boldsymbol{L}$ the lower triangular Cholesky factor of $\boldsymbol{K}_{mm}^{-1}$, the derivative of g_i over $\boldsymbol{Z}$ is

$$\begin{aligned} \frac{\partial g_i}{\partial \boldsymbol{Z}} =& \frac{1}{\sigma^2}\big[\left((\boldsymbol{L}\boldsymbol{p}_i)\circ\boldsymbol{k}_m(\boldsymbol{x}_i)\right)\boldsymbol{x}_i^T\mathrm{diag}(\eta) - \left(((\boldsymbol{L}\boldsymbol{p}_i)\circ\boldsymbol{k}_m(\boldsymbol{x}_i))\mathbf{1}_d^T\right)\circ(\boldsymbol{Z}\mathrm{diag}(\boldsymbol{\eta})) \\ &- (\boldsymbol{T}_i + \boldsymbol{T}_i^T)\boldsymbol{Z}\mathrm{diag}(\boldsymbol{\eta}) + ((\boldsymbol{T}_i + \boldsymbol{T}_i^T)\mathbf{1}_m\boldsymbol{\eta}^T)\circ\boldsymbol{Z}\big], \qquad (6) \end{aligned}$$

where

$$\boldsymbol{p_i} = -\boldsymbol{\mu}y_i + (\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\phi}(\boldsymbol{x}_i), \qquad (7)$$

$$\boldsymbol{T_i} = \left[\boldsymbol{L}\left((\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{p}_i^T)\circ\boldsymbol{\Psi}\right)\boldsymbol{L}^T\right]\circ\boldsymbol{K}_{mm}. \qquad (8)$$

The symbol $\circ$ denotes the Hadamard product, and $\boldsymbol{\Psi}$ is an upper triangular matrix with diagonal elements all equal to 0.5 and strictly upper triangular elements all equal to 1, as follows:

$$\boldsymbol{\Psi} = \begin{bmatrix} 0.5 & 1 & \dots & 1 & 1 \\ 0 & 0.5 & \ddots & 1 & 1 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 0.5 & 1 \\ 0 & 0 & \dots & 0 & 0.5 \end{bmatrix}. \tag{9}$$

A.6 Derivative over $\ln \boldsymbol{\eta}$

The derivative of g_i over $\ln \boldsymbol{\eta}$ is

$$\begin{aligned} \frac{\partial g_i}{\partial \ln \boldsymbol{\eta}} =& \frac{1}{2\sigma^2} \Big\{ 2\mathbf{1}_m^{\mathrm{T}} \left[\boldsymbol{Z} \circ \left(\left((\boldsymbol{L}\boldsymbol{p}_i) \circ \boldsymbol{k}_m(\boldsymbol{x}_i) \right) \boldsymbol{x}_i^T \right) \right] - \mathbf{1}_m^{\mathrm{T}} \left((\boldsymbol{L}\boldsymbol{p}_i) \circ \boldsymbol{k}_m(\boldsymbol{x}_i) \right) (\boldsymbol{x}_i \circ \boldsymbol{x}_i)^{\mathrm{T}} - \left((\boldsymbol{L}\boldsymbol{p}_i) \circ \boldsymbol{k}_m(\boldsymbol{x}_i) \right)^{\mathrm{T}} (\boldsymbol{Z} \circ \boldsymbol{Z}) \\ & - \mathbf{1}_m^{\mathrm{T}} \left[\boldsymbol{Z} \circ ((\boldsymbol{T}_i + \boldsymbol{T}_i^{\mathrm{T}}) \boldsymbol{Z}) \right] + \mathbf{1}_m^{\mathrm{T}} \left[(\boldsymbol{T}_i + \boldsymbol{T}_i^T)(\boldsymbol{Z} \circ \boldsymbol{Z}) \right] \Big\} \circ \boldsymbol{\eta}. \end{aligned} \tag{10}$$

B Properties of the ELBO of ADVGP

By defining $\boldsymbol{U}$ as the upper triangular Cholesky factor of $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{\Sigma} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$, we have

Lemma B.1 *The gradient of g_i in Equation 1, ∇g_i, is Lipschitz continuous with respect to each element in $\boldsymbol{\mu}$ and $\boldsymbol{U}$.*

We can prove this by showing the first derivative of ∇g_i with respect to each element of $\boldsymbol{\mu}$ and $\boldsymbol{U}$ is bounded, which is constant in our case. As shown in our paper, the gradients of g_i with respect to $\boldsymbol{\mu}$ and $\boldsymbol{U}$ are:

$$\frac{\partial g_i}{\partial \boldsymbol{\mu}} = \frac{1}{\sigma^2}[-y_i \boldsymbol{\phi}_i + \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T \boldsymbol{\mu}], \tag{11}$$

$$\frac{\partial g_i}{\partial \boldsymbol{U}} = \frac{1}{\sigma^2} \mathrm{triu}[\boldsymbol{U} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T], \tag{12}$$

which are affine functions for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. Therefore, the first derivative of ∇g_i is constant.

Lemma B.2 *h in Equation 2 is a convex function with respect to $\boldsymbol{\mu}$ and $\boldsymbol{U}$.*

This can be proved by verifying that the Hessian matrices of h with respect to $\boldsymbol{\mu}$ and $\mathrm{vec}(\boldsymbol{U})$ are both positive semidefinite, where we denote $\mathrm{vec}(\cdot)$ as the operator that stacks the columns of a matrix as a vector. To show this, we first compute the partial derivatives of h with respect to $\boldsymbol{\mu}$ and $\boldsymbol{U}$ as

$$\frac{\partial h}{\partial \boldsymbol{\mu}} = \boldsymbol{\mu}, \tag{13}$$

$$\frac{\partial h}{\partial \boldsymbol{U}} = -\mathrm{diag}(\boldsymbol{U}^{-1}) + \boldsymbol{U}. \tag{14}$$

The Hessian matrix of h with respect to $\boldsymbol{\mu}$ is

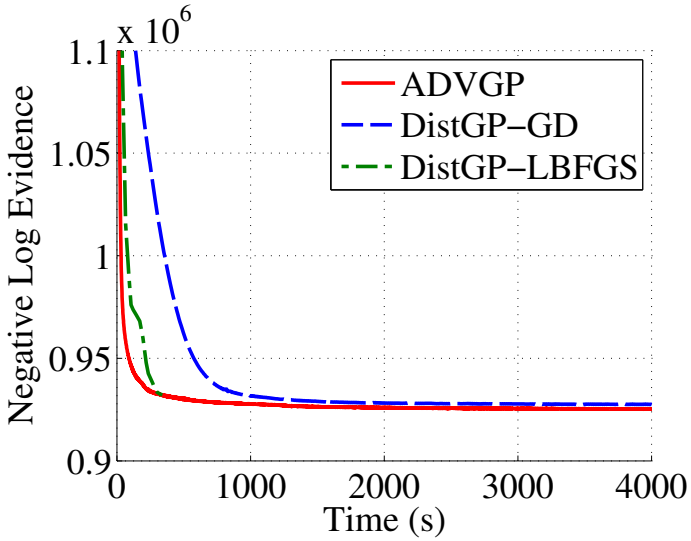
$$\boldsymbol{H}(\boldsymbol{\mu}) = \boldsymbol{I}_{m \times m} \succeq 0. \tag{15}$$

The Hessian matrix of h with respect to $\mathrm{vec}(\boldsymbol{U})$ is

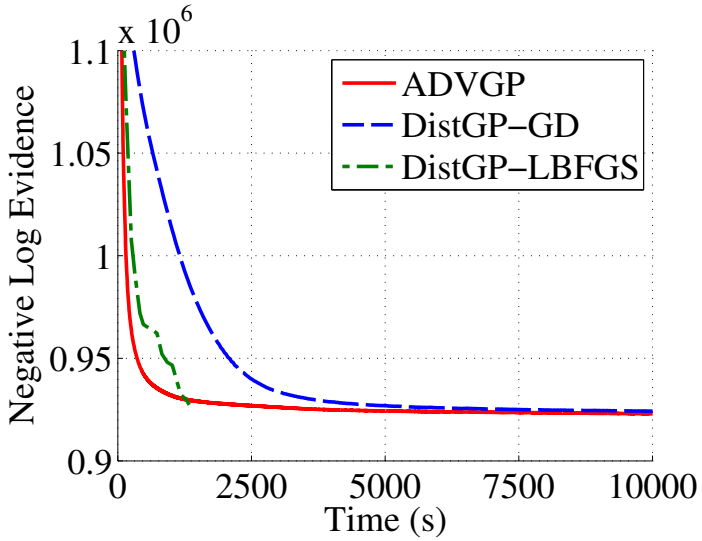
$$\boldsymbol{H}(\mathrm{vec}(\boldsymbol{U})) = \mathrm{diag}(\boldsymbol{h}) \succeq 0, \tag{16}$$

where $\boldsymbol{h} = [\frac{\partial h}{\partial U_{11}^2}, \dots, \frac{\partial h}{\partial U_{1m}^2}, \dots, \frac{\partial h}{\partial U_{m1}^2}, \dots, \frac{\partial h}{\partial U_{mm}^2}]$, and $\frac{\partial h}{\partial U_{ij}^2} = 1 + \delta(i,j) \frac{1}{U_{i,i}^2}$.

C Negative Log Evidences on US Flight Data



(A) $m = 100$

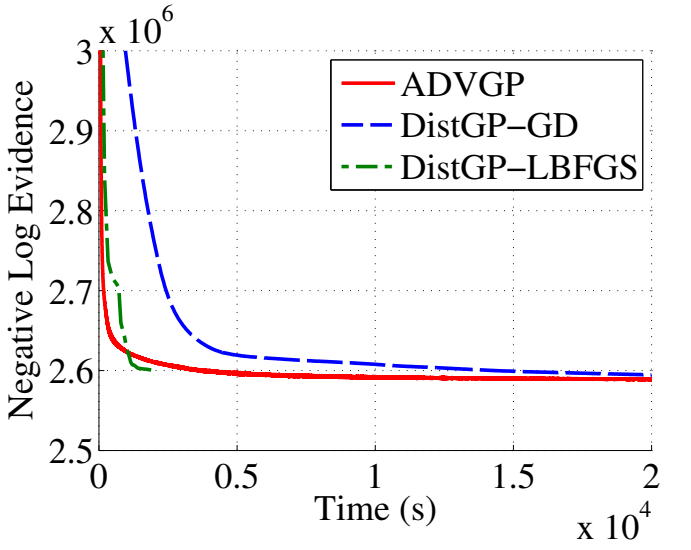


(B) $m = 200$

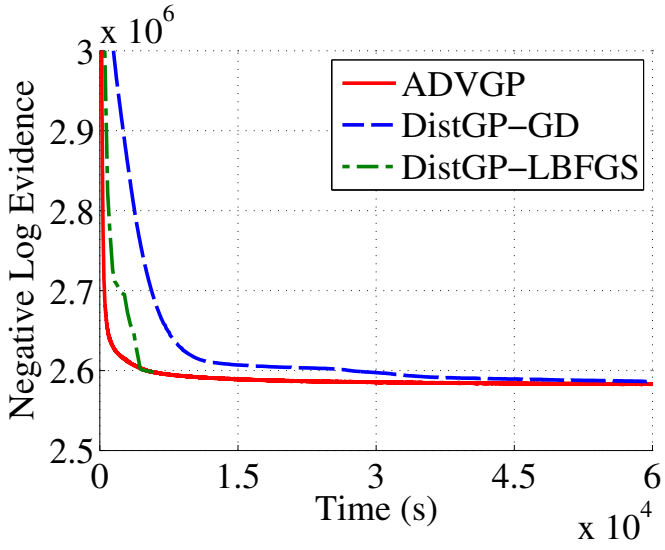
Figure C.1: Negative log evidences for 700K/100K US Flight data as a function of training time.

Method	$m = 100$	$m = 200$
ADVGP	**925236**	**922907**
DistGP-GD	927414	924208
DistGP-LBFGS	932179	927331

Table C.1: Negative log evidences for 700K/100K US Flight data.



(A) $m = 100$



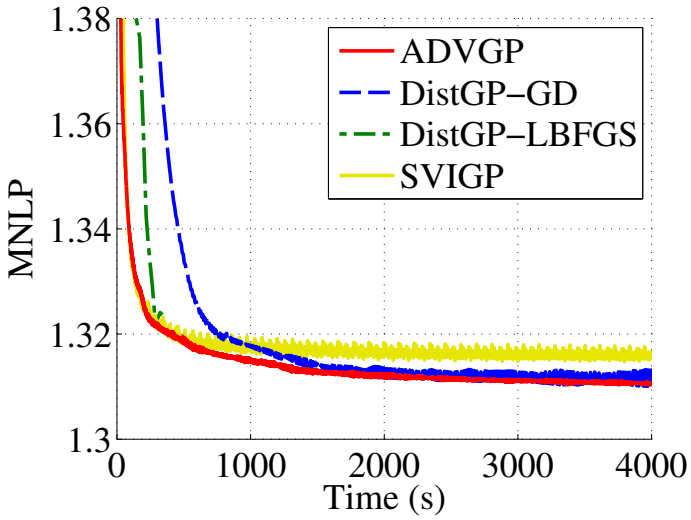
(B) $m = 200$

Figure C.2: Negative log evidences for 2M/100K US Flight data as a function of training time.

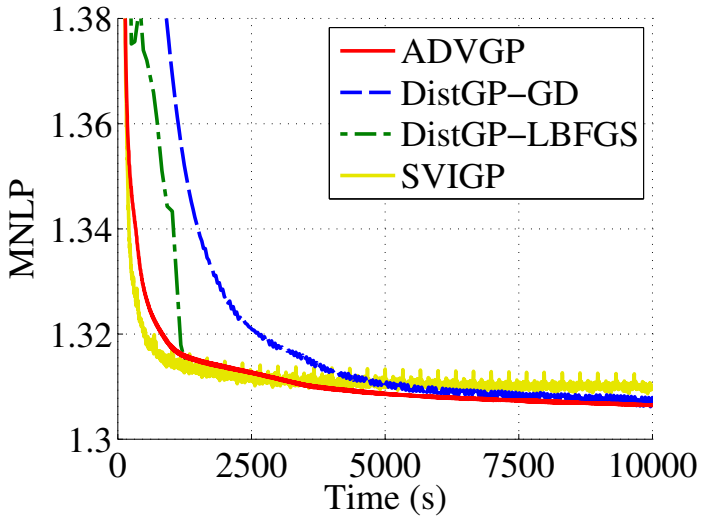
Method	$m = 100$	$m = 200$
ADVGP	$\mathbf{2.58921 \times 10^6}$	$\mathbf{2.58267 \times 10^6}$
DistGP-GD	2.59471×10^6	2.58601×10^6
DistGP-LBFGS	2.59971×10^6	2.59817×10^6

Table C.2: Negative log evidences for 2M/100K US Flight data.

D Mean Negative Log Predictive Likelihoods (MNLPs) on US Flight Data



(A) $m = 100$

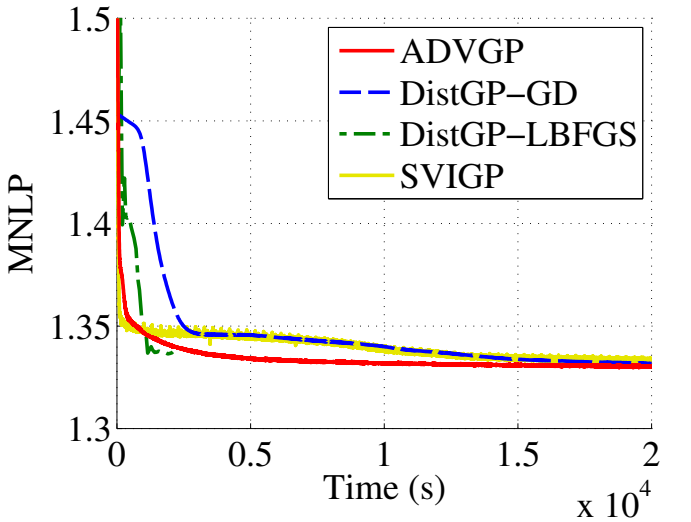


(B) $m = 200$

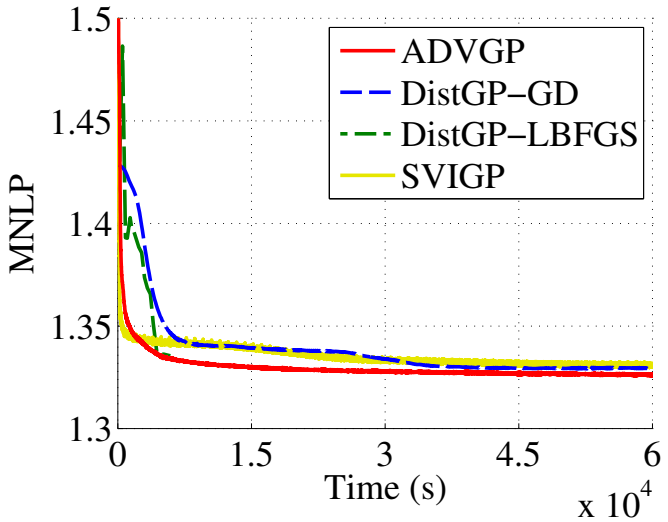
Figure D.1: Mean negative log predictive likelihoods for 700K/100K US Flight data as a function of training time.

Method	$m = 100$	$m = 200$
ADVGP	1.3106	1.3066
DistGP-GD	**1.3099**	**1.3062**
DistGP-LBFGS	1.3237	1.3136
SVIGP	1.3157	1.3096

Table D.1: Mean negative log predictive likelihoods for 700K/100K US Flight data.



(A) $m = 100$



(B) $m = 200$

Figure D.2: Mean negative log predictive likelihoods for 2M/100K US Flight data as a function of training time.

Method	$m = 100$	$m = 200$
ADVGP	**1.3301**	**1.3258**
DistGP-GD	1.3317	1.3297
DistGP-LBFGS	1.3380	1.3355
SVIGP	1.3335	1.3306

Table D.2: Mean negative log predictive likelihoods for 2M/100K US Flight data.