Scalable Training of Inference Networks for Gaussian-Process Models Appendix

A. Derivation using the Dual Representation of GP and Natural Gradient

We note that the stochastic mirror descent is equivalent to natural gradient updates for exponential family densities (Raskutti & Mukherjee, 2015; Khan & Lin, 2017). To show this, we derive the same adaptive Bayesian filter in eq. (6) using the dual representation of GP as a Gaussian measure and natural gradient.

A Gaussian process $\mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$ has a dual representation as a Gaussian measure ν on a separable Banach space $\mathcal{B}$ (Cheng & Boots, 2017; Mallasto & Feragen, 2017). There is an RKHS $\mathcal{H}$ that corresponds to a positive definite kernel k densely embedded in $\mathcal{B}$. The measure ν is constructed as follows. First define a *canonical Gaussian cylinder set measure* $\nu^{\mathcal{H}}$ on $\mathcal{H}$, denoted as $\mathcal{N}(f|\mu, \Sigma)$, where $\mu \in \mathcal{H}$ is the mean function, $\Sigma : \mathcal{H} \to \mathcal{H}$ is a bounded positive semi-definite linear operator. They satisfy

$$m(x) = \mu^\top k(x, \cdot),$$
$$\kappa(x, x') = k(x, \cdot)^\top \Sigma k(x', \cdot),$$

where $h^\top g$ denotes inner product in the RKHS: $h^\top g = \langle h, g \rangle_{\mathcal{H}}$. Let i be the inclusion map from $\mathcal{H}$ into $\mathcal{B}$. Then the measure ν is induced by $\nu^{\mathcal{H}}$ using this map. In the measure theory of infinite-dimensional space, ν is known as the *abstract Wiener measure*. The RKHS $\mathcal{H}$ is sometimes called the *Cameron-Martin space*.

Remark 1. *The intuition for this construction is that the canonical Gaussian cylinder set measure $\nu^{\mathcal{H}}$ is not a proper measure. In fact, we can show that countable additivity does not hold for this "measure" (Eldredge, 2016). The inclusion map i here radonifies $\nu^{\mathcal{H}}$ into a true measure ν. One way to think about this is that functions drawn from the Gaussian process fall outside of $\mathcal{H}$ with probability one (Kanagawa et al., 2018), but are contained in $\mathcal{B}$. Despite this, as pointed out in Cheng & Boots (2017), we can conveniently work with the canonical form $\mathcal{N}(f|\mu, \Sigma)$ and get correct results as long as the conclusion is independent of the dimension of f.*

It is easy to check that the canonical Gaussian measure which corresponds to the GP prior $\mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ is $p(f) = \mathcal{N}(f|0, I)$. Assuming the canonical form of the variational distribution is $q(f) = \mathcal{N}(f|\mu, \Sigma)$, we have stochastic approximation of the lower bound as:

$$\hat{\mathcal{L}}(q) = N\mathbb{E}_q \log p(y_n|f) - \mathrm{KL}\left[q(f)\|p(f)\right].$$

Now that we can interpret GPs in the form of canonical Gaussian measures, we can then write $q(f)$ and $p(f)$ in the exponential family form to simplify the derivation:

$$q(f) \propto \exp\{\lambda^\top t(f) - A(\lambda)\},$$
$$p(f) \propto \exp\{\lambda_0^\top t(f) - A(\lambda_0)\}.$$

where $t(f) = \{f, ff^\top\}$ denotes the sufficient statistics, $A(\lambda) = \frac{1}{2}\mu^\top \Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma|$ is the partition function. The natural parameters of $p(f)$ and $q(f)$ are $\lambda_0 = \{0, -\frac{1}{2}I\}$ and $\lambda = \{\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\}$, respectively. Let u denote the mean parameters of $q(f)$: $u = \mathbb{E}_q\left[t(f)\right] = \{\mu, \Sigma\}$. There is a dual relationship between the natural parameter λ and the mean parameter u: $u = \nabla_\lambda A(\lambda)$. The mapping ∇A is one-to-one when the exponential family is minimal (Wainwright et al., 2008). The stochastic natural gradient of the lower bound with respect to λ is defined as

$$\tilde{\nabla}_\lambda \hat{\mathcal{L}}(q) = F(\lambda)^{-1} \nabla_\lambda \hat{\mathcal{L}}(q),$$

where $F(\lambda) = \mathbb{E}_q\left[\nabla_\lambda \log q(f)\ \nabla_\lambda \log q(f)^\top\right]$ is the Fisher information matrix. The natural gradient of the KL divergence

term is

$$\begin{aligned}\tilde{\nabla}_\lambda \mathrm{KL}\left[q(f)\|p(f)\right] &= F(\lambda)^{-1}\nabla_\lambda \mathbb{E}_q\left[\log\frac{q(f)}{p(f)}\right] \\ &= F(\lambda)^{-1}\nabla_\lambda\left[(\lambda-\lambda_0)^\top u - A(\lambda) + A(\lambda_0)\right] \\ &= F(\lambda)^{-1}\left[u - \nabla_\lambda A(\lambda) + \nabla_\lambda u(\lambda-\lambda_0)\right] \\ &= F(\lambda)^{-1}\nabla_\lambda^2 A(\lambda)(\lambda-\lambda_0) \\ &= \lambda - \lambda_0,\end{aligned}$$

where we have used the fact that $\nabla_\lambda^2 A(\lambda) = F(\lambda)$. We can also derive a simplified form of the natural gradient of the conditional log likelihood term, by writing it as the gradient with respect to the mean parameter u:

$$\begin{aligned}\tilde{\nabla}_\lambda \mathbb{E}_q[\log p(y_n|f)] &= F(\lambda)^{-1}\nabla_\lambda \mathbb{E}_q\left[\log p(y_n|f)\right] \\ &= F(\lambda)^{-1}\nabla_\lambda u \cdot \nabla_u \mathbb{E}_q[\log p(y_n|f)] \\ &= F(\lambda)^{-1}\nabla_\lambda^2 A(\lambda)\cdot \nabla_u \mathbb{E}_q[\log p(y_n|f)] \\ &= \nabla_u \mathbb{E}_q[\log p(y_n|f)].\end{aligned}$$

So the natural gradient update can be written as

$$\begin{aligned}\lambda_{t+1} &= \lambda_t + \beta_t\left(N\nabla_u \mathbb{E}_q\left[\log p(y_n|f)\right] - \lambda_t + \lambda_0\right) \\ &= (1-\beta_t)\lambda_t + \beta_t\left(N\nabla_u \mathbb{E}_q[\log p(y_n|f)] + \lambda_0\right).\end{aligned}$$

Reinterpreting the above equation in the density space, we have

$$q_{t+1}(f) \propto q_t(f)^{1-\beta_t} p(f)^{\beta_t} \exp\{\langle \nabla_u \mathbb{E}_q[\log p(y_n|f)], t(f)\rangle\}^{N\beta_t}. \tag{15}$$

The likelihood $p(y_n|f)$ is said to be conjugate with the prior if it has a form as $p(y_n|f) \propto \exp\{\lambda(y_n)^\top t(f)\}$. For example, in GP regression, the likelihood is $p(y_n|f) = \mathcal{N}(y_n|f(\mathbf{x}_n), \sigma^2)$, which has an above form with the natural parameter $\lambda(y_n) = \{\frac{1}{\sigma^2} y_n k(\mathbf{x}_n,\cdot), -\frac{1}{2\sigma^2}k(\mathbf{x}_n,\cdot)k(\mathbf{x}_n,\cdot)^\top\}$. By plugging in $p(y_n|f) \propto \exp\{\lambda(y_n)^\top t(f)\}$, we can verify that eq. (15) is equivalent to

$$q_{t+1}(f) \propto q_t(f)^{1-\beta_t} p(f)^{\beta_t} p(y_n|f)^{N\beta_t},$$

which turns out to be the same adaptive Bayesian filter we get in eq. (6). As for the non-conjugate case, we can view the natural gradient update as the projection of the functional mirror descent update onto exponential families, by approximating the likelihood term with the exponential family $\exp\{\langle \nabla_u \mathbb{E}_q[\log p(y_n|f)], t(f)\rangle\}$.

B. Experiment Details and Additional Results

B.1. Synthetic Data

The full figures including FBNN with $M = 2, 5, 20$ are shown in Fig. 4. We can see that FBNN's problem of overestimating uncertainty is consistent with different values of M. We set $\beta_0 = 1, \xi = 0.1$ in this experiment.

B.2. Regression

For all regression experiments, we set the measurement points to be sampled from the empirical distribution of training data convolved with the prior RBF kernel if the prior hyperparameters are initialized by optimizing GP marginal likelihood on a random subset, otherwise we set $c(\mathbf{x})$ simply to be the training data distribution. We fix $c(\mathbf{x})$ and do not adapt it together with the prior kernel parameters when the prior hyperparameters are updated during training.

Benchmarks The GP we use has a RBF kernel with dimension-wise lengthscales, also known as *Automatic Relevance Determination* (ARD) (MacKay, 1996). The RFE inference network we use has 1000 features (hidden units). We initialize the lengthscales in the network using the lengthscales of the prior kernel, and initialize the frequencies (first-layer weights) with random samples from a standard Gaussian. We then train these frequencies and lengthscales as inference network parameters. For FBNN, we keep all settings (including the inference network) the same as GPNet except the training

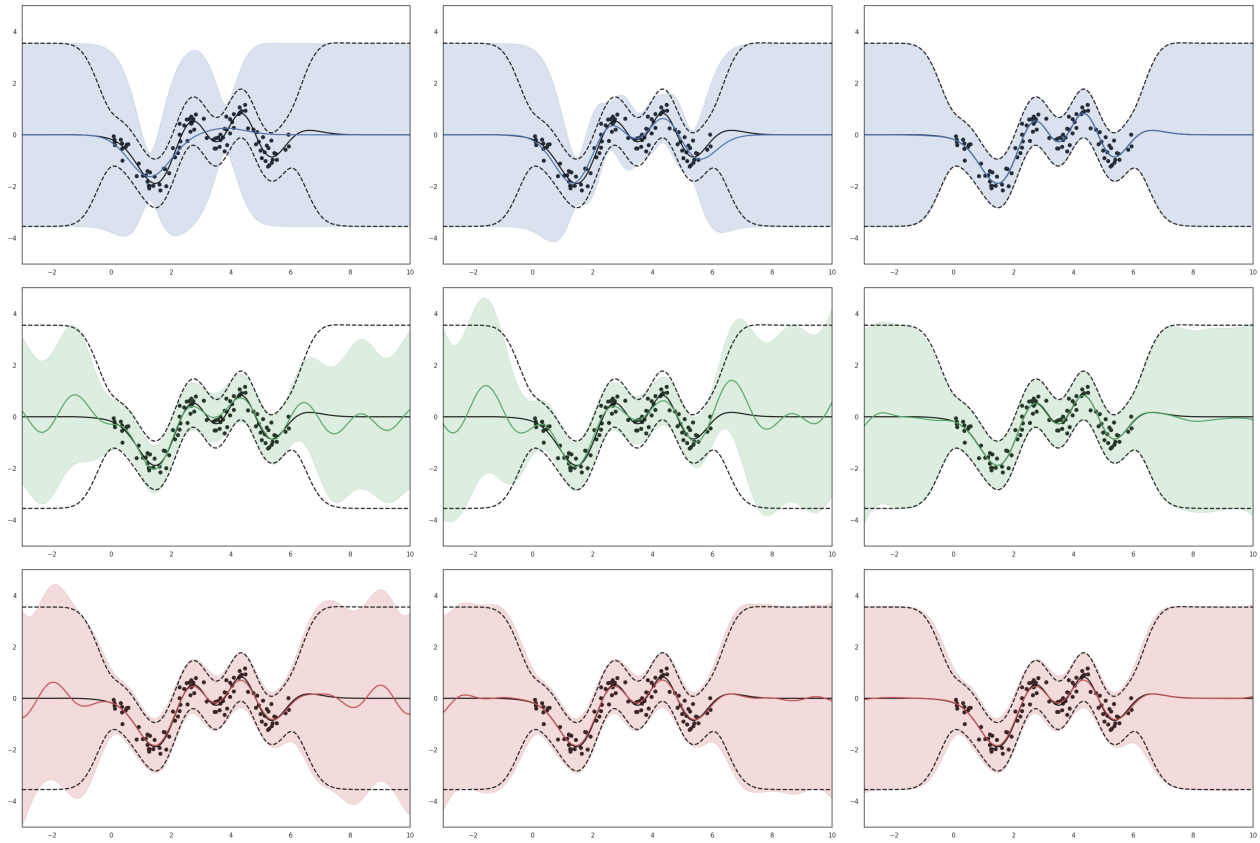

Figure 4. Posterior process on the Snelson dataset, where shaded areas correspond to intervals of ± 3 standard deviations, and dashed lines denote the ground truths. (a) From top to bottom: SVGP with $M \in \{2, 5, 20\}$ (left to right) inducing points; GPNet with $M \in \{2, 5, 20\}$ measurement points; FBNN with $M \in \{2, 5, 20\}$ measurement points.

objective used. We use 20 random splits for each dataset, where we keep 90% of the dataset as the training set and use the remaining 10% for test. The inputs and outputs of all data points are scaled to have nearly zero mean and unit variance using the mean and standard deviation calculated from training data. We use minibatch size 500, learning rate $\eta = 0.003$ for all the experiments. With each random seed we ran 10K iterations. For Boston, Concrete, Kin8nm and Protein we initialize the prior hyperparameters by maximizing GP marginal likelihood for 1K iterations on a randomly chosen subset of 1000 points. For Power and Wine we optimize the prior hyperparameters during training using the minibatch lower bound and the same learning rate as η, as described in section 3.3. We set $\beta_0 = 1$ and $\xi = 1$ except for Power and Protein we use $\beta_0 = 0.01, \xi = 0.1$ and $\beta_0 = 0.1, \xi = 0.1$, respectively. In Table 3 and Table 4 we list the mean and standard errors of all experiments.

Airline Delay We use the same type of GP prior and inference networks as in benchmark datasets above. We set $\beta_0 = 0.1, \xi = 0.1$. For all methods we train for 10K iterations with minibatch size 500 and learning rate $\eta = 0.003$. For GPNet we optimize the prior hyperparameters during training with the same learning rate as η, for which the objective is described in section 3.3, while we found that doing the same with FBNN seriously hurt its performance (leading to RMSE 27.186 for $M = 100$), therefore we did not update hyperparameters during training for FBNN. We initialize the prior hyperparameters by maximizing GP marginal likelihood for 1K iterations on a randomly chosen subset of 1000 points. We did this for both GPNet and FBNN, though we found that for GPNet this does not improve the performance.

B.3. Classification

CNN-GP Prior The prior is defined as follows. Let $\mathbf{Z}^{(\ell)}(\mathbf{x})$ denote the pre-activation output of the ℓ-th layer of the ConvNet. The shape of $\mathbf{Z}^{(\ell)}(\mathbf{x})$ is $C^{(\ell)} \times (H^{(\ell)} D^{(\ell)})$. Each row of it represents the flattened feature map in a channel. A

Table 3. Regression: RMSE.

DATA SET	N	D	SVGP, 100	GPNET, 100	SVGP, 500	GPNET, 500	FBNN, 500
BOSTON	506	12	2.897±0.132	2.786±0.142	3.023±0.187	2.754±0.143	2.775±0.141
CONCRETE	1030	8	5.768±0.094	5.301±0.127	5.075±0.119	5.050±0.132	5.089±0.117
ENERGY	768	8	0.469±0.014	0.493±0.022	0.439±0.015	0.461±0.014	0.459±0.013
KIN8NM	8192	8	0.086±0.001	0.080±0.001	0.074±0.000	0.067±0.000	0.072±0.000
POWER	9568	4	3.941±0.033	3.942±0.032	3.791±0.034	3.898±0.032	4.135±0.029
PROTEIN	45730	9	4.536±0.010	4.540±0.014	4.154±0.010	4.329±0.013	4.087±0.051
WINE	1599	11	0.625±0.009	0.614±0.010	0.626±0.009	0.627±0.009	0.633±0.008

Table 4. Regression: Test log likelihood.

DATA SET	N	D	SVGP, 100	GPNET, 100	SVGP, 500	GPNET, 500	FBNN, 500
BOSTON	506	12	-2.465±0.054	-2.421±0.049	-2.458±0.072	-2.429±0.055	-2.437±0.025
CONCRETE	1030	8	-3.166±0.015	-3.115±0.024	-3.027±0.023	-3.066±0.022	-3.046±0.029
ENERGY	768	8	-0.675±0.024	-1.060±0.008	-0.600±0.033	-0.847±0.013	-0.755±0.018
KIN8NM	8192	8	1.006±0.004	1.095±0.011	1.183±0.004	1.283±0.005	1.189±0.005
POWER	9568	4	-2.793±0.008	-2.794±0.007	-2.755±0.008	-2.783±0.008	-2.847±0.006
PROTEIN	45730	9	-2.932±0.002	-3.057±0.032	-2.841±0.002	-2.986±0.029	-2.821±0.014
WINE	1599	11	-0.949±0.014	-0.917±0.014	-0.949±0.015	-0.948±0.014	-0.961±0.013

hidden layer in the network makes the transformation:

$$Z^{(\ell+1)}_{j,g}(\mathbf{x}) = b^{(\ell)}_j + \sum_{i=1}^{C^{(\ell)}} \sum_{h=1}^{H^{(\ell)} D^{(\ell)}} W^{(\ell)}_{j,i,g,h} a(Z^{(\ell)}_{i,h}(\mathbf{x})),$$

where $\mathbf{W}_{ji}$ is the pseudo weight matrix that corresponds to the convolutional filter $\mathbf{U}_{ji}$. The elements of each row in $\mathbf{W}_{ji}$ are zero except where $\mathbf{U}_{ji}$ applies. b_j denotes the bias in the j-th channel. a is the ReLU activation function. Let x, y denote the positions within a filter, independent Gaussian priors are placed over $u^{(\ell)}_{j,i,x,y}$ and $b^{(\ell)}_j$ to form a Bayesian ConvNet:

$$u^{(\ell)}_{j,i,x,y} \sim \mathcal{N}(0, \sigma_w^2/C^{(\ell)}), \quad b_j \sim \mathcal{N}(0, \sigma_b^2).$$

By carefully taking the limit of hidden-layer widths, one can prove that each row in $\mathbf{Z}^{(\ell)}(\mathbf{x})$ form a multivariate Gaussian, and different rows (channels) are independent and identically distributed (i.i.d.), thus showing that the Bayesian ConvNet defines a GP (Garriga-Alonso et al., 2019). It is easy to show the prior mean function is zero: $\mathbb{E}[Z^{(\ell+1)}_{j,g}(\mathbf{x})] = 0$. To determine the prior covariance kernel of the output, we can follow a recursive procedure (For simplicity, we use $v^{(\ell)}_g(\mathbf{x}, \mathbf{x}')$ to denote the covariance between $Z^{(\ell)}_{j,g}(\mathbf{x})$ and $Z^{(\ell)}_{j,g}(\mathbf{x}')$):

$$v^{(\ell+1)}_g(\mathbf{x}, \mathbf{x}') = \sigma_b^2 + \sigma_w^2 \sum_{h \in g\text{-th patch}} s^{(\ell)}_h(\mathbf{x}, \mathbf{x}'),$$

$$s^{(\ell)}_h(\mathbf{x}, \mathbf{x}') = 1/2\pi \sqrt{v^{(\ell)}_g(\mathbf{x}, \mathbf{x}) v^{(\ell)}_g(\mathbf{x}', \mathbf{x}')} J_1(\theta^{(\ell)}_g),$$

where $J_1(\theta^{(\ell)}_g) = \sin\theta^{(\ell)}_g + (\pi - \theta^{(\ell)}_g)\cos\theta^{(\ell)}_g$ and $\theta^{(\ell)}_g = \arccos\left(v^{(\ell)}_g(\mathbf{x}, \mathbf{x}')/\sqrt{v^{(\ell)}_g(\mathbf{x}, \mathbf{x}) v^{(\ell)}_g(\mathbf{x}', \mathbf{x}')}\right)$. The prior convnet we used is a deep convolutional neural network with 6 residual blocks, each two of them operates on a different size of feature maps, with the first two on feature maps with the same size as the original image. There are strided convolution (stride=2) between the three groups.

Inference Networks The inference network we used has the same structure as the prior ConvNet, except the number of convolutional filters are [64, 64, 128, 128, 256, 256]. On top of it we have a fully-connected layer of size 512 and neural tangent kernels defined by a MLP with 100 hidden units for the output of each class.

We use batch size 64 and $M = 64$ measurement points in this experiment. We set $c(\mathbf{x})$ to be the empirical distribution of the training data. In implementation this simply means that we use two different shuffles of the training dataset, and pick a minibatch from each of them. Then we use one of the two minibatches as training points, and the other as measurement points. The learning rate is $\eta = 0.0003$. We set $\beta_0 = 0.01, \xi = 0.1$, and ran for 10K iterations. We did not update prior hyperparameters in this experiment.