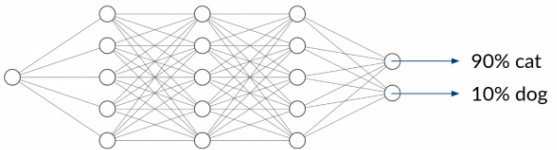
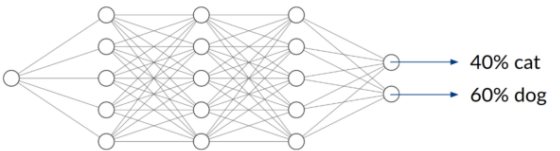
Variational Linearized Laplace Approximation for Bayesian Deep Learning

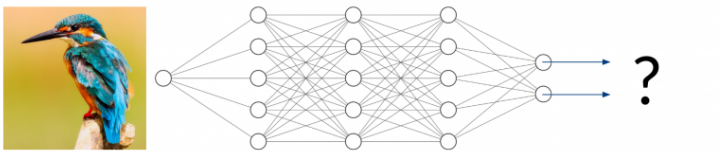
Luis Antonio Ortega Andrés

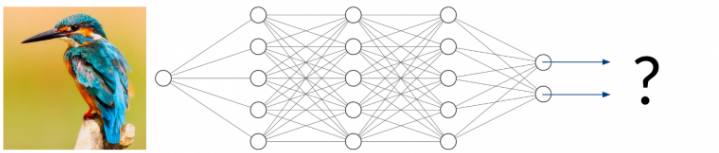
March 21, 2023

Autonomous University of Madrid


90% cat
10% dog

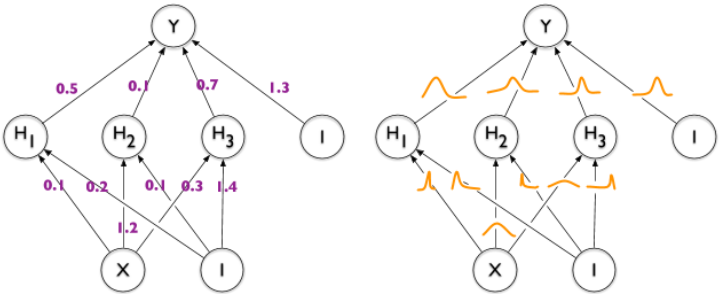

40% cat
60% dog





Deep learning methods are unable to quantify the uncertainty of their predictions!

Straight-forward solution: Using a Bayesian model.



Making predictions requires the posterior over the parameters of the model $\boldsymbol{\theta}$:

$$P(y^\star|\mathbf{x}^\star, \mathcal{D}) = \int P(y^\star|\mathbf{x}^\star, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{D}) \, d\boldsymbol{\theta}$$

where $P(\boldsymbol{\theta}|\mathcal{D})$ is **intractable for complex models**.

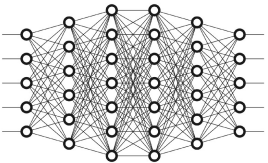
Approximate $P(\boldsymbol{\theta}|\mathcal{D})$ by something simpler $Q(\boldsymbol{\theta})$.

Approximate $P(\boldsymbol{\theta}|\mathcal{D})$ by something simpler $Q(\boldsymbol{\theta})$.

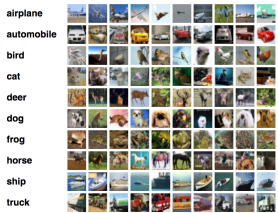
Poor performance of the model in many cases.

Laplace Approximation for Deep Learning

Deep Model



Dataset



Optimal parameter $\hat{\boldsymbol{\theta}}$ using deep learning techniques.

Given the optimal parameters $\hat{\boldsymbol{\theta}}$, we aim to make **uncertainty estimation** on the pre-trained model.

Given the optimal parameters $\hat{\boldsymbol{\theta}}$, we aim to make **uncertainty estimation** on the pre-trained model.

Laplace approximation (LA) of $P(\boldsymbol{\theta}|\mathcal{D})$ centered on $\hat{\boldsymbol{\theta}}$ as

$$P(\boldsymbol{\theta}|\mathcal{D}) \approx Q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma}^{-1} = -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log P(\boldsymbol{\theta}|\mathcal{D})_{|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} (\log P(\mathcal{D}|\boldsymbol{\theta}) + \log P(\boldsymbol{\theta}))_{|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

Given the optimal parameters $\hat{\boldsymbol{\theta}}$, we aim to make **uncertainty estimation** on the pre-trained model.

Laplace approximation (LA) of $P(\boldsymbol{\theta}|\mathcal{D})$ centered on $\hat{\boldsymbol{\theta}}$ as

$$P(\boldsymbol{\theta}|\mathcal{D}) \approx Q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma}^{-1} = -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log P(\boldsymbol{\theta}|\mathcal{D})_{|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log P(\mathcal{D}|\boldsymbol{\theta}) + \frac{1}{\sigma_0^2}\boldsymbol{I}_P$$

Given the optimal parameters $\hat{\boldsymbol{\theta}}$, we aim to make **uncertainty estimation** on the pre-trained model.

Laplace approximation (LA) of $P(\boldsymbol{\theta}|\mathcal{D})$ centered on $\hat{\boldsymbol{\theta}}$ as

$$P(\boldsymbol{\theta}|\mathcal{D}) \approx Q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma})$$

where applying the generalized Gauss-Newton matrix approximation

$$\boldsymbol{\Sigma}^{-1} \approx \tilde{\boldsymbol{\Sigma}}^{-1} = \sum_{n=1}^{N} \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_n) \Lambda(\mathbf{x}_n, y_n) \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_n)^T + \frac{1}{\sigma_0^2} \boldsymbol{I}_P \,.$$

with

$$\mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_n) = \nabla_{\boldsymbol{\theta}} g(\mathbf{x}_n, \boldsymbol{\theta})_{|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad \Lambda(\mathbf{x}_n, y_n) = -\nabla^2_{\mathbf{gg}} \log P(y_n|\mathbf{g})_{|\mathbf{g}=g(\mathbf{x}_n,\hat{\boldsymbol{\theta}})}$$

The Laplace approximation of the deep model has poor performance.

The Laplace approximation of the deep model has poor performance.

$$\Downarrow$$

There is a shift between the posterior and the predictive distribution.

$$Q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Sigma}}) \quad P(y^\star|\mathbf{x}^\star, \mathcal{D}) \approx \mathbb{E}_{Q(\boldsymbol{\theta})}\left[P(y^\star|g(\mathbf{x}^\star, \boldsymbol{\theta}))\right]$$

The Laplace approximation of the deep model has poor performance.

$$\Downarrow$$

There is a shift between the posterior and the predictive distribution.

$$Q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Sigma}}) \quad P(y^\star|\mathbf{x}^\star, \mathcal{D}) \approx \mathbb{E}_{Q(\boldsymbol{\theta})}\left[P(y^\star|g(\mathbf{x}^\star, \boldsymbol{\theta}))\right]$$

The GGN approximation, is the true posterior of a linearized model

$$g^{lin}(\mathbf{x}, \boldsymbol{\theta}) = g(\mathbf{x}, \hat{\boldsymbol{\theta}}) + \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_n)^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

Which means, that predictions should be made over the linearized model:

$$P(y^\star|\mathbf{x}^\star, \mathcal{D}) \approx \mathbb{E}_{Q(\boldsymbol{\theta})}\left[P(y^\star|g^{lin}(\mathbf{x}^\star, \boldsymbol{\theta}))\right]$$

The **linearized Laplace approximation** (LLA) is equivalent to a **Gaussian Process**.

The **linearized Laplace approximation** (LLA) is equivalent to a **Gaussian Process**.

Prior distribution

$$P(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$$

The **linearized Laplace approximation** (LLA) is equivalent to a **Gaussian Process**.

Prior distribution

$$P(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$$

$$\Downarrow$$

Gaussian Process

$$m(\mathbf{x}) = g^{lin}(\mathbf{x}, \boldsymbol{m})$$
$$K(\mathbf{x}, \mathbf{x}') = \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{x})^T \boldsymbol{S} \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}')$$

Using LLA approximate posterior

$$Q(\boldsymbol{\theta}) = \mathcal{N}(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Sigma}})$$

$$\Downarrow$$

Gaussian Process

$$m^\star(\mathbf{x}) = g^{lin}(\mathbf{x}, \hat{\boldsymbol{\theta}})$$

$$K^\star(\mathbf{x}, \mathbf{x}') = \sigma_0^2 \left(\kappa(\mathbf{x}, \mathbf{x}') - \kappa(\mathbf{x}, \mathbf{X}) \left(\frac{1}{\sigma_0^2} \boldsymbol{\Lambda}_{\mathbf{X},\mathbf{y}}^{-1} + \kappa(\mathbf{X}, \mathbf{X}) \right)^{-1} \kappa(\mathbf{X}, \mathbf{x}') \right).$$

where $\kappa(\mathbf{x}, \mathbf{x}') = \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{x})^T \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}')$.

Main idea: Approximate the exact GP posterior using a sparse approach with inducing points.

Main inconvenience: Using a sparse GP **changes the predictive mean**, loosing the pre-trained solution.

Solution: Use a Generalized Sparse GP in the RKHS.

Dual representation of Gaussian Processes

An RKHS $\mathcal{H}$ is a Hilbert space of functions satisfying the reproducing property: $\forall \mathbf{x} \in \mathcal{X}\ \exists \phi_{\mathbf{x}} \in \mathcal{H}$ such that $\forall f \in \mathcal{H}, f(\mathbf{x}) = \langle \phi_{\mathbf{x}}, f \rangle$.

Dual representation of Gaussian Processes

An RKHS $\mathcal{H}$ is a Hilbert space of functions satisfying the reproducing property: $\forall \mathbf{x} \in \mathcal{X}\ \exists \phi_{\mathbf{x}} \in \mathcal{H}$ such that $\forall f \in \mathcal{H}, f(\mathbf{x}) = \langle \phi_{\mathbf{x}}, f \rangle$.

A **Gaussian process** $\mathcal{GP}(m, K)$ has a dual representation in a RKHS as: there exists $\mu \in \mathcal{H}$ and a linear semi-definite positive operator $\Sigma : \mathcal{H} \to \mathcal{H}$ such that, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $\exists \phi_{\mathbf{x}}, \phi_{\mathbf{x}'}$, verifying

$$m(\mathbf{x}) = \langle \phi_{\mathbf{x}}, \mu \rangle \,, \quad K(\mathbf{x}, \mathbf{x}') = \langle \phi_{\mathbf{x}}, \Sigma(\phi_{\mathbf{x}'}) \rangle$$

Dual representation of Gaussian Processes

An RKHS $\mathcal{H}$ is a Hilbert space of functions satisfying the reproducing property: $\forall \mathbf{x} \in \mathcal{X}\ \exists \phi_{\mathbf{x}} \in \mathcal{H}$ such that $\forall f \in \mathcal{H}, f(\mathbf{x}) = \langle \phi_{\mathbf{x}}, f \rangle$.

A **Gaussian process** $\mathcal{GP}(m, K)$ has a dual representation in a RKHS as: there exists $\mu \in \mathcal{H}$ and a linear semi-definite positive operator $\Sigma : \mathcal{H} \to \mathcal{H}$ such that, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $\exists \phi_{\mathbf{x}}, \phi_{\mathbf{x}'}$, verifying

$$m(\mathbf{x}) = \langle \phi_{\mathbf{x}}, \mu \rangle \,, \quad K(\mathbf{x}, \mathbf{x}') = \langle \phi_{\mathbf{x}}, \Sigma(\phi_{\mathbf{x}'}) \rangle$$

As an abuse of notation, we write $f \sim \mathcal{N}(\mu, \Sigma)$, which is a Gaussian measure in the RKHS.

Theorem (Cheng and Boots, 2016). Using a sparse GP approximation with variational distribution $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$ is equivalent to restricting the mean and covariance functions of the dual representation in the RKHS to

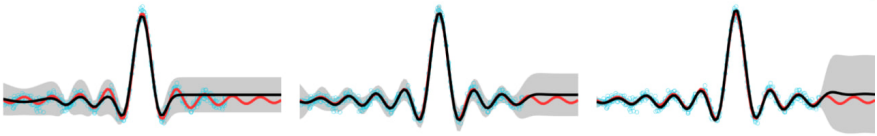
$$\tilde{\mu} = \Phi_{\mathbf{Z}}(\boldsymbol{a}) \quad \text{and} \quad \tilde{\Sigma} = I + \Phi_{\mathbf{Z}} \boldsymbol{A} \Phi_{\mathbf{Z}}^T ,$$

where $\Phi_{\mathbf{Z}} : \mathbb{R}^M \to \mathcal{H}$ is defined as $\Phi_{\mathbf{Z}}(\boldsymbol{a}) = \sum_{m=1}^M a_m \phi_{\mathbf{z}_m}$, $\boldsymbol{a} \in \mathbb{R}^M$ and $\Phi_{\mathbf{Z}} \boldsymbol{A} \Phi_{\mathbf{Z}}^T = \sum_{i=1}^M \sum_{j=1}^M \phi_{\mathbf{z}_i} A_{i,j} \phi_{\mathbf{z}_j}^T$, $\boldsymbol{A} \in \mathbb{R}^{M \times M}$ such that $\tilde{\Sigma} \geq 0$.

Theorem (Cheng and Boots, 2016). Using a sparse GP approximation with variational distribution $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$ is equivalent to restricting the mean and covariance functions of the dual representation in the RKHS to

$$\tilde{\mu} = \Phi_{\mathbf{Z}_\alpha}(\boldsymbol{a}) \quad \text{and} \quad \tilde{\Sigma} = I + \Phi_{\mathbf{Z}_\beta} \boldsymbol{A} \Phi_{\mathbf{Z}_\beta}{}^T,$$

where $\Phi_{\mathbf{Z}} : \mathbb{R}^M \to \mathcal{H}$ is defined as $\Phi_{\mathbf{Z}_\alpha}(\boldsymbol{a}) = \sum_{m=1}^M a_m \phi_{\mathbf{z}_m}$, $\boldsymbol{a} \in \mathbb{R}^M$ and $\Phi_{\mathbf{Z}_\beta} \boldsymbol{A} \Phi_{\mathbf{Z}_\beta}{}^T = \sum_{i=1}^M \sum_{j=1}^M \phi_{\mathbf{z}_i} A_{i,j} \phi_{\mathbf{z}_j}^T$, $\boldsymbol{A} \in \mathbb{R}^{M \times M}$ such that $\tilde{\Sigma} \geq 0$.



(a) $M = 10$ (b) $M_\alpha = 100, M_\beta = 10$ (c) $M = 100$

Comparison between models with shared and decoupled basis.

- (a)(c) denote the models with shared basis of size M.
- (b) denotes the model of decoupled basis with size (M_α, M_β).

For any function $t \in \mathcal{H}$, and $\epsilon > 0$, there exists a set of points $\mathbf{Z}_\alpha \subset \mathcal{X}$ and coefficients $\boldsymbol{a}$, such that

$$d_{\mathcal{H}}(t, h_\epsilon) \leq \epsilon, \text{ with } h_\epsilon = \Phi_{\mathbf{Z}_\alpha}(\boldsymbol{a}) .$$

For any function $t \in \mathcal{H}$, and $\epsilon > 0$, there exists a set of points $\mathbf{Z}_\alpha \subset \mathcal{X}$ and coefficients $\boldsymbol{a}$, such that

$$d_{\mathcal{H}}(t, h_\epsilon) \leq \epsilon, \text{ with } h_\epsilon = \Phi_{\mathbf{Z}_\alpha}(\boldsymbol{a})\,.$$

We can fit the mean of the posterior distribution of the sparse decoupled GP to any function in the Hilbert space ($g(\cdot, \hat{\boldsymbol{\theta}})$).

For any function $t \in \mathcal{H}$, and $\epsilon > 0$, there exists a set of points $\mathbf{Z}_\alpha \subset \mathcal{X}$ and coefficients $\boldsymbol{a}$, such that

$$d_{\mathcal{H}}(t, h_\epsilon) \leq \epsilon, \text{ with } h_\epsilon = \Phi_{\mathbf{Z}_\alpha}(\boldsymbol{a})\,.$$

We can fit the mean of the posterior distribution of the sparse decoupled GP to any function in the Hilbert space ($g(\cdot, \hat{\boldsymbol{\theta}})$).

If $g(\cdot, \hat{\boldsymbol{\theta}}) \notin \mathcal{H}$, we can use that $\mathcal{H}$ is dense in the span of the Gaussian Process.

Proposition. If $g(\cdot, \hat{\boldsymbol{\theta}}) \in \mathcal{H}$, $\forall \epsilon > 0$, there exists a set of M_α inducing points $\mathbf{Z}_\alpha \subset \mathcal{X}$ and $\boldsymbol{a} \in \mathbb{R}^{M_\alpha}$ such that the dual representation in the RKHS of the corresponding sparse Gaussian process defined by

$$\tilde{\mu} = \Phi_{\mathbf{Z}_\alpha}(\boldsymbol{a}) \quad \text{and} \quad \tilde{\Sigma} = (I + \Phi_{\mathbf{Z}_\beta} \boldsymbol{A} \Phi_{\mathbf{Z}_\beta}^T)^{-1} \,,$$

corresponds to a posterior approximation $\mathcal{GP}(m^\star, K^\star)$ with mean and covariance functions defined as

$$\begin{aligned} m^\star(\mathbf{x}) &= h_\epsilon(\mathbf{x}) \,, \\ K^\star(\mathbf{x}, \mathbf{x}') &= K(\mathbf{x}, \mathbf{x}') - K_{\mathbf{x}, \mathbf{Z}_\beta} (\boldsymbol{A}^{-1} + K_{\mathbf{Z}_\beta})^{-1} K_{\mathbf{Z}_\beta, \mathbf{x}'} \,, \end{aligned}$$

where $\mathbf{Z}_\beta \subset \mathcal{X}$ is a set of M_β inducing points, $\boldsymbol{A} \in \mathbb{R}^{M_\beta \times M_\beta}$ such that $\tilde{\Sigma} \geq 0$ and h_ϵ verifies $d_{\mathcal{H}}(g(\cdot, \hat{\boldsymbol{\theta}}), h_\epsilon) \leq \epsilon$.

Optimizing the ELBO in the Hilbert space:

$$\begin{aligned}\max_{q(f),\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta}}(q(f)) &= \max_{q(f),\boldsymbol{\theta}} \int q(f) \log \frac{p_{\boldsymbol{\theta}}(y|f)p(f)}{q(f)} df \\ &= \max_{q(f),\boldsymbol{\theta}} \mathbb{E}_q \left[\log p_{\boldsymbol{\theta}}(y|f)\right] - \mathsf{KL}\left(q \mid p\right) \ .\end{aligned}$$

where

$$\mathsf{KL}\left(q \mid p\right) = \frac{1}{2}\boldsymbol{a}^T \boldsymbol{K}_\alpha \boldsymbol{a} + \frac{1}{2}\log|\boldsymbol{I} + \boldsymbol{K}_\beta \boldsymbol{A}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_\beta (\boldsymbol{A}^{-1} + \boldsymbol{K}_\beta)^{-1}\right)$$

Intuitive Recap

1. The Linearized Laplace Approximation is equivalent to a GP.

Intuitive Recap

1. The Linearized Laplace Approximation is equivalent to a GP.
2. Sparse approach based on inducing points to approximate the posterior GP.

Intuitive Recap

1. The Linearized Laplace Approximation is equivalent to a GP.
2. Sparse approach based on inducing points to approximate the posterior GP.
3. Separate the inducing locations from the mean and covariance.

Intuitive Recap

1. The Linearized Laplace Approximation is equivalent to a GP.
2. Sparse approach based on inducing points to approximate the posterior GP.
3. Separate the inducing locations from the mean and covariance.
4. Use "*infinite*" inducing points for the mean. Fixing the mean to the pre-trained MAP solution.

Intuitive Recap

1. The Linearized Laplace Approximation is equivalent to a GP.
2. Sparse approach based on inducing points to approximate the posterior GP.
3. Separate the inducing locations from the mean and covariance.
4. Use "*infinite*" inducing points for the mean. Fixing the mean to the pre-trained MAP solution.
5. The resulting method does scale with the number of parameters and the dataset size (mini-batch optimization).

ELLA

Another existing sparse approximation is ELLA.

1. Approximated the kernel matrix of the full dataset using a random subset of it.

$$\hat{\mathbf{X}} \subset \mathbf{X} \implies \hat{\boldsymbol{K}} = \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\hat{\mathbf{X}})\mathcal{J}_{\hat{\boldsymbol{\theta}}}(\hat{\mathbf{X}})^T \approx \boldsymbol{K} = \mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{X})\mathcal{J}_{\hat{\boldsymbol{\theta}}}(\mathbf{X})^T$$

2. Finds eigen-decomposition of $\hat{\boldsymbol{K}}$ to create lower-dimensional features $\phi(\mathbf{x})$.
3. Use these features to approximate the true covariance matrix $\boldsymbol{K}$.
4. Requires an unique iteration over the whole training dataset.

Preliminary results

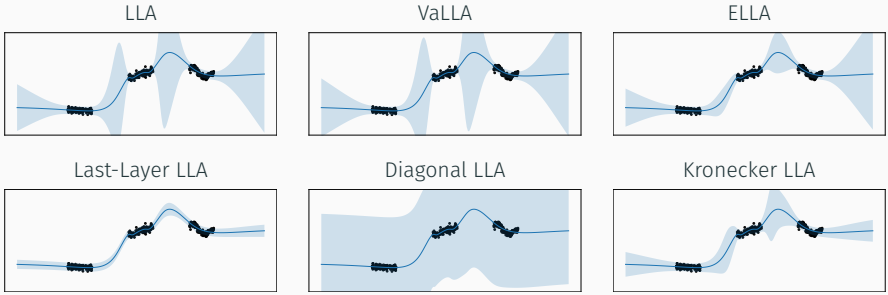
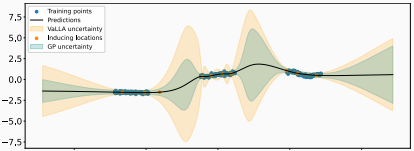
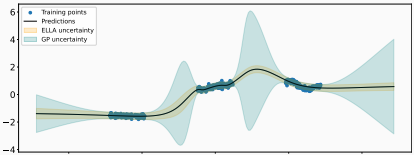


Figure 1: Predictive distribution (two times the standard deviation) on a toy 1D regression dataset with a 2 hidden layer MLP with 50 units.

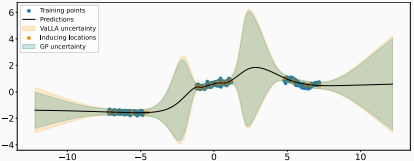
VaLLA $M = 10$



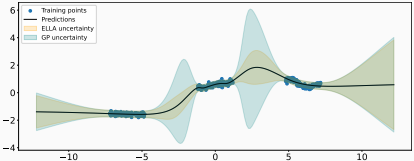
ELLA $M = 10$



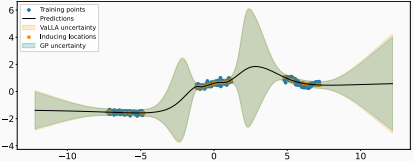
VaLLA $M = 20$



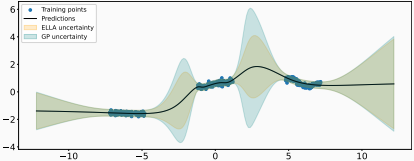
ELLA $M = 20$



VaLLA $M = 30$



ELLA $M = 30$



Conclusions

1. Using the dual definition in RKHS, sparse GPs can be generalized to decouple the inducing points.

Conclusions

1. Using the dual definition in RKHS, sparse GPs can be generalized to decouple the inducing points.
2. The decoupled definition can be used to fix the mean of the GP to the pre-trained MAP solution.

Conclusions

1. Using the dual definition in RKHS, sparse GPs can be generalized to decouple the inducing points.
2. The decoupled definition can be used to fix the mean of the GP to the pre-trained MAP solution.
3. Stochastic optimization can be employed to train the model.

Conclusions

1. Using the dual definition in RKHS, sparse GPs can be generalized to decouple the inducing points.
2. The decoupled definition can be used to fix the mean of the GP to the pre-trained MAP solution.
3. Stochastic optimization can be employed to train the model.
4. As a result, VaLLA is scalable in both parameters and dataset size.

Conclusions

1. Using the dual definition in RKHS, sparse GPs can be generalized to decouple the inducing points.
2. The decoupled definition can be used to fix the mean of the GP to the pre-trained MAP solution.
3. Stochastic optimization can be employed to train the model.
4. As a result, VaLLA is scalable in both parameters and dataset size.
5. The obtained preliminary results show competitive uncertainty estimation compared to other LLA approximations.

Thank you for your attention!

- Cheng, Ching-An and Byron Boot (2017). “Variational inference for Gaussian process models with linear complexity”. In: *Advances in Neural Information Processing Systems* 30.
- Cheng, Ching-An and Byron Boots (2016). “Incremental variational sparse Gaussian process regression”. In: *Advances in Neural Information Processing Systems* 29.
- Deng, Zhijie, Feng Zhou, and Jun Zhu (2022). “Accelerated Linearized Laplace Approximation for Bayesian Deep Learning”. In: *arXiv preprint arXiv:2210.12642*.
- Immer, Alexander, Maciej Korzepa, and Matthias Bauer (2021). “Improving predictions of Bayesian neural nets via local linearization”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 703–711.