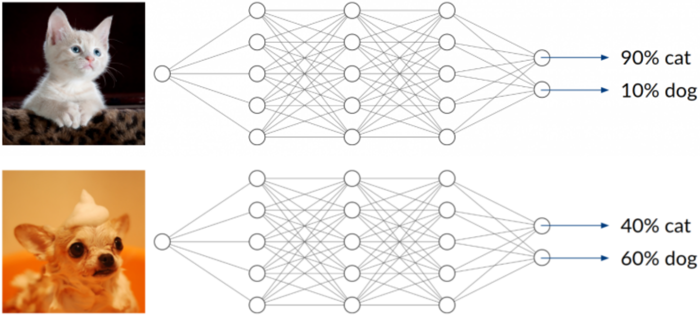
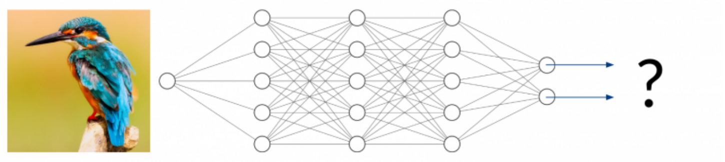
Variational Inference in RKHS for Uncertainty Estimation in Deep Learning

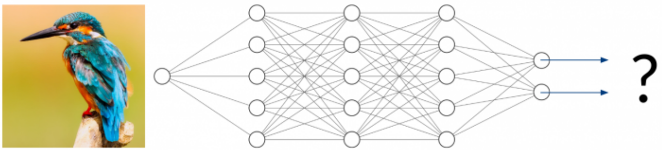
Luis Antonio Ortega Andrés
Simón Rodríguez Santana
Daniel Hernández Lobato

April 17, 2024

Autonomous University of Madrid

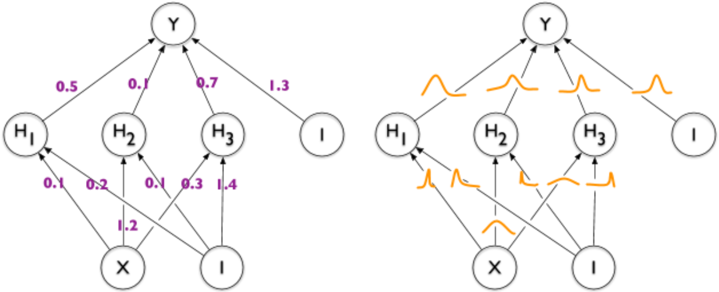

90% cat
10% dog
40% cat
60% dog





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 $q(\boldsymbol{\theta})$.

$$\Downarrow$$

Poor performance in many cases.

Approach

1. Learn a DL **deterministic** model h.

High Performance - No Uncertainty

Approach

1. Learn a DL **deterministic** model h.

   High Performance - No Uncertainty

2. Variational Sparse Gaussian Processes with **posterior mean** h.

   High Performance - Uncertainty Estimation

Approach

1. Learn a DL **deterministic** model h.

   High Performance - No Uncertainty

2. Variational Sparse Gaussian Processes with **posterior mean** h.

   High Performance - Uncertainty Estimation

3. Optimize parameters using function-space VI.

Uncertainty Estimation in function-space

Given a mean $m(\cdot)$ and covariance function $\kappa(\cdot, \cdot)$, defines a **Gaussian prior over function evaluations**:

$$p(f(\mathbf{x})) = \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))\,.$$

$$f \sim \mathcal{GP}(m, K)\,.$$

Uncertainty Estimation in function-space

Given a mean $m(\cdot)$ and covariance function $\kappa(\cdot, \cdot)$, defines a Gaussian prior over function evaluations:

$$p(f(\mathbf{x})) = \mathcal{N}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}))\,.$$

$$f \sim \mathcal{GP}(m, \kappa)\,.$$

Set of observations $(\mathbf{X}, \mathbf{y})$, the **predictive distribution** is Gaussian

$$p(y^\star|\mathbf{x}^\star, \mathbf{X}, \mathbf{y}) = \mathcal{N}(m^\star(\mathbf{x}^\star), K(\mathbf{x}^\star, \mathbf{x}^\star)) \,.$$

Set of observations $(\mathbf{X}, \mathbf{y})$, the **predictive distribution** is Gaussian

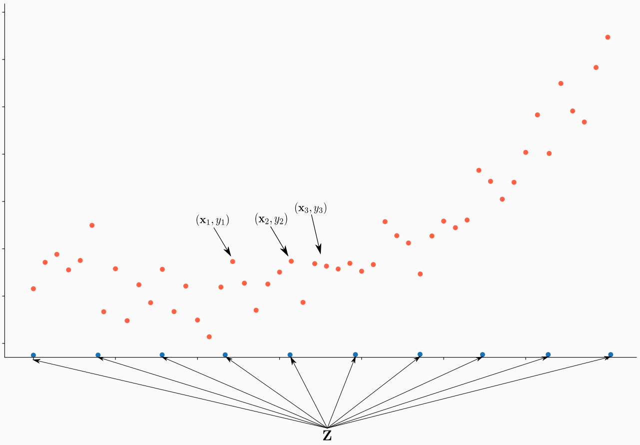
$$p(y^\star|\mathbf{x}^\star, \mathbf{X}, \mathbf{y}) = \mathcal{N}(m^\star(\mathbf{x}^\star), K(\mathbf{x}^\star, \mathbf{x}^\star))\,.$$

$$m^\star(\mathbf{x}^\star) = \kappa(\mathbf{x}^\star, \mathbf{X})(\kappa(\mathbf{X}, \mathbf{X}) + \sigma^2 \boldsymbol{I})^{-1}(\mathbf{y} - m(\mathbf{x}^\star))\,,$$
$$K(\mathbf{x}^\star, \mathbf{x}^\star) = \kappa(\mathbf{x}^\star, \mathbf{x}^\star) - \kappa(\mathbf{x}^\star, \mathbf{X})(\kappa(\mathbf{X}, \mathbf{X}) + \sigma^2 \boldsymbol{I})^{-1}\kappa(\mathbf{X}, \mathbf{x}^\star)\,.$$

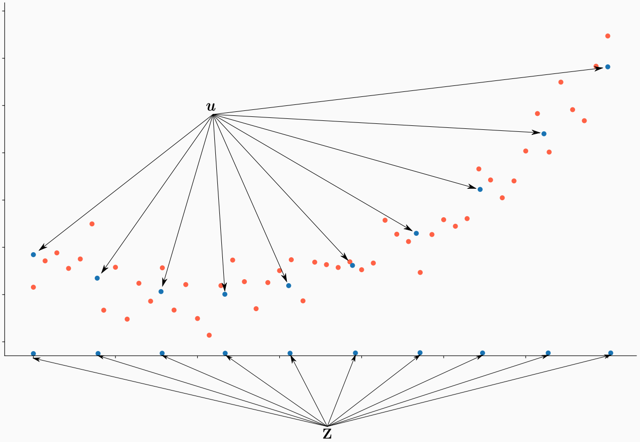
Gaussian noise with variance σ^2 is considered for the targets

Sparse Variational Gaussian Processes

Define a set of *inducing locations* $\mathbf{Z} \subset \mathbb{R}^D$ that "summarize" the training inputs $\mathbf{X}$.

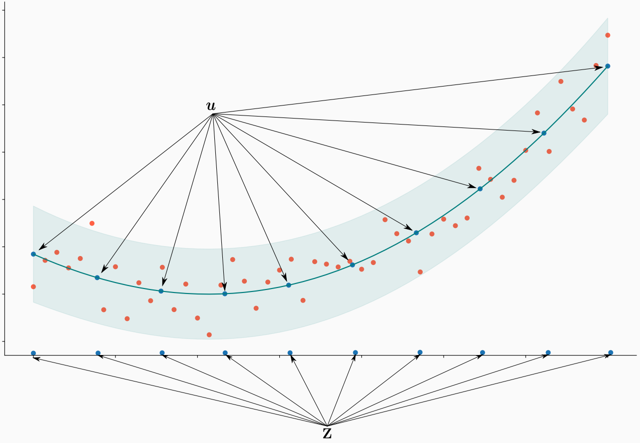


With $\mathbf{u} = f(\mathbf{Z})$, the posterior $p(\mathbf{u}|\mathbf{X}, \mathbf{y})$ is approximated with variational distribution $q(\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.



Sparse Variational Gaussian Processes

The inducing points can be marginalized in closed form to make predictions.



A **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 $\forall g \in \mathcal{H}, g(\mathbf{x}) = \langle \phi_{\mathbf{x}}, g \rangle$.

Dual representation of Gaussian Processes

A **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 $\forall g \in \mathcal{H}, g(\mathbf{x}) = \langle \phi_{\mathbf{x}}, g \rangle$.

A **Gaussian process** $f \sim \mathcal{GP}(m, K)$ 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

A **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 $\forall g \in \mathcal{H}, g(\mathbf{x}) = \langle \phi_{\mathbf{x}}, g \rangle$.

A **Gaussian process** $f \sim \mathcal{GP}(m, K)$ 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

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

We write $f \sim \mathcal{N}(\mu, \Sigma)$, which is a **Gaussian measure** in the RKHS.

This characterization in the RKHS allows the **rethink Gaussian Processes as Gaussian Measures** in the Hilbert space:

$$p(f) = \mathcal{N}(\mu, \Sigma)$$

This characterization in the RKHS allows the **rethink Gaussian Processes as Gaussian Measures** in the Hilbert space:

$$p(f) = \mathcal{N}(\mu, \Sigma)$$

The **posterior measure** can be specified as a **Gaussian**:

$$p(f|\mathbf{y}) = \mathcal{N}(\mu^\star, \Sigma^\star)$$

$$\mu^\star = \kappa(\cdot, \mathbf{X})(\kappa(\mathbf{X}, \mathbf{X}) + \sigma^2 \boldsymbol{I})^{-1}(\mathbf{y} - m(\cdot))$$
$$\Sigma^\star = I - \phi_{\mathbf{X}}^T(\kappa(\mathbf{X}, \mathbf{X}) + \sigma^2 \boldsymbol{I})^{-1}\phi_{\mathbf{X}}$$

Theorem. A SVGP is equivalent to restricting the mean and covariance functions 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

$$\Phi_{\mathbf{Z}}(\boldsymbol{a}) = \sum_{m=1}^{M} a_m \phi_{\mathbf{z}_m} \,, \quad \text{and} \quad \Phi_{\mathbf{Z}} \boldsymbol{A} \Phi_{\mathbf{Z}}^T = \sum_{i=1}^{M} \sum_{j=1}^{M} \phi_{\mathbf{z}_i} A_{i,j} \phi_{\mathbf{z}_j}^T$$

where $\boldsymbol{A} \in \mathbb{R}^{M \times M}$ such that $\tilde{\Sigma} \geq 0$.

Cheng, C.A. and Boots, B., 2016. Incremental variational sparse Gaussian process regression. Advances in Neural Information Processing Systems, 29.

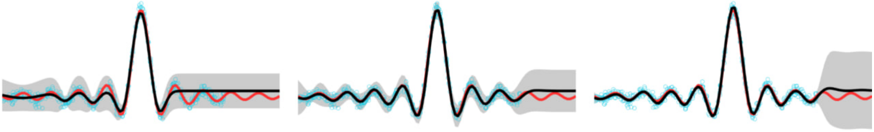
A SVGP can be generalized with mean and covariance functions of the dual representation in the RKHS to

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

where $\mathbf{Z}_\alpha$ and $\mathbf{Z}_\beta$ are two sets of inducing locations.

Cheng, C.A. and Boots, B., 2017. Variational inference for Gaussian process models with linear complexity. Advances in Neural Information Processing Systems, 30.

Comparison between models with shared and decoupled basis



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

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

Figure from Cheng, C.A. and Boots, B., 2017. Variational inference for Gaussian process models with linear complexity. Advances in Neural Information Processing Systems, 30.

Variational Optimization

SVGPs are optimized using the Evidence Lower Bound:

$$\mathsf{KL}\big(p(\mathbf{f},\mathbf{u}|\mathbf{y})|q(\mathbf{f},\mathbf{u})\big) = \log p(\mathbf{y}) \underbrace{- \mathbb{E}_{q(\mathbf{f})}[\log p(\mathbf{y}|\mathbf{f})] + \mathsf{KL}\big(q(\mathbf{u})|p(\mathbf{u})\big)}_{-ELBO}$$

with $\mathbf{f} = f(\mathbf{X})$ and $\mathbf{u} = f(\mathbf{Z})$.

Cheng, C.A. and Boots, B., 2016. Incremental variational sparse Gaussian process regression. Advances in Neural Information Processing Systems, 29.

Variational Optimization

SVGPs are optimized using the Evidence Lower Bound:

$$\mathsf{KL}\big(p(\mathbf{f},\mathbf{u}|\mathbf{y})|q(\mathbf{f},\mathbf{u})\big) = \log p(\mathbf{y}) \underbrace{- \mathbb{E}_{q(\mathbf{f})}[\log p(\mathbf{y}|\mathbf{f})] + \mathsf{KL}\big(q(\mathbf{u})|p(\mathbf{u})\big)}_{-ELBO}$$

with $\mathbf{f} = f(\mathbf{X})$ and $\mathbf{u} = f(\mathbf{Z})$.

Which is **equivalent** to optimize the ELBO in function-space

$$\mathsf{KL}\big(p(f|\mathbf{y})|q(f)\big) = \log p(\mathbf{y}) \underbrace{- \mathbb{E}_{q(f)}[\log p(\mathbf{y}|f)] + \mathsf{KL}\big(q(f)|p(f)\big)}_{-ELBO}$$

Cheng, C.A. and Boots, B., 2016. Incremental variational sparse Gaussian process regression. Advances in Neural Information Processing Systems, 29.

Variational Optimization

Optimizing the ELBO in the Hilbert space:

$$\max_{q(f)} \mathcal{L}(q(f)) = \max_{q(f)} \mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right) \,.$$

Where

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

and $\mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right]$ can be computed in regression and estimated in classification.

Fixing the Mean Function

If the kernel $\kappa(\cdot, \cdot)$ is universal, then, $\forall \epsilon > 0$, there exists a set of M_α points $\mathbf{Z}_\alpha \subset \mathbb{R}^D$ and coefficients $\boldsymbol{a} \in \mathbb{R}^{M_\alpha}$, such that

$$d_{\mathcal{H}}(h, \Phi_{\mathbf{Z}_\alpha}(\boldsymbol{a})) \leq \epsilon, \quad \text{with} \quad \Phi_{\mathbf{Z}_\alpha}(\boldsymbol{a}) := \sum_{m=1}^{M_\alpha} a_m \phi_{\mathbf{z}_m} .$$

For any $\epsilon > 0$, there exists a set of M_α inducing points $\mathbf{Z}_\alpha \subset \mathbb{R}^D$ and $\boldsymbol{a} \in \mathbb{R}^{M_\alpha}$

For any $\epsilon > 0$, there exists a set of M_α inducing points $\mathbf{Z}_\alpha \subset \mathbb{R}^D$ and $\boldsymbol{a} \in \mathbb{R}^{M_\alpha}$ such that the corresponding decoupled sparse Gaussian process corresponds to $\mathcal{GP}(m^\star, K^\star)$:

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

For any $\epsilon > 0$, there exists a set of M_α inducing points $\mathbf{Z}_\alpha \subset \mathbb{R}^D$ and $\boldsymbol{a} \in \mathbb{R}^{M_\alpha}$ such that the corresponding decoupled sparse Gaussian process corresponds to $\mathcal{GP}(m^\star, K^\star)$:

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

where $\mathbf{Z}_\beta \subset \mathbb{R}^D$ 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 it verifies

$$
d_{\mathcal{H}}(h, m^\star) \leq \epsilon
$$

Distributions over function-space with fixed mean to h.

Distributions over function-space with fixed mean to h.

Parameters: $\mathbf{Z}_\beta \subset \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{M_\beta \times M_\beta}$ (such that $\tilde{\Sigma} \geq 0$).

Distributions over function-space with fixed mean to h.

Parameters: $\mathbf{Z}_\beta \subset \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{M_\beta \times M_\beta}$ (such that $\tilde{\Sigma} \geq 0$).

Gaussian process posterior approximation $\mathcal{GP}(m^\star, K^\star)$:

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

Diagram - Distribution over function-space

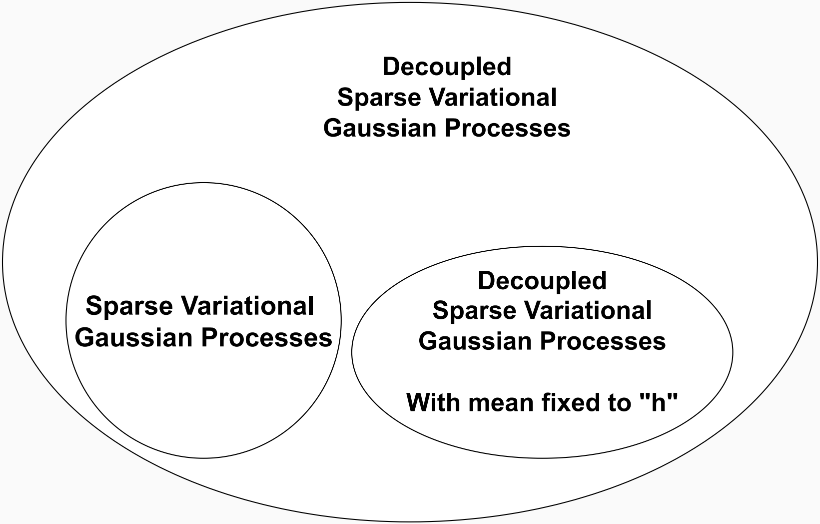
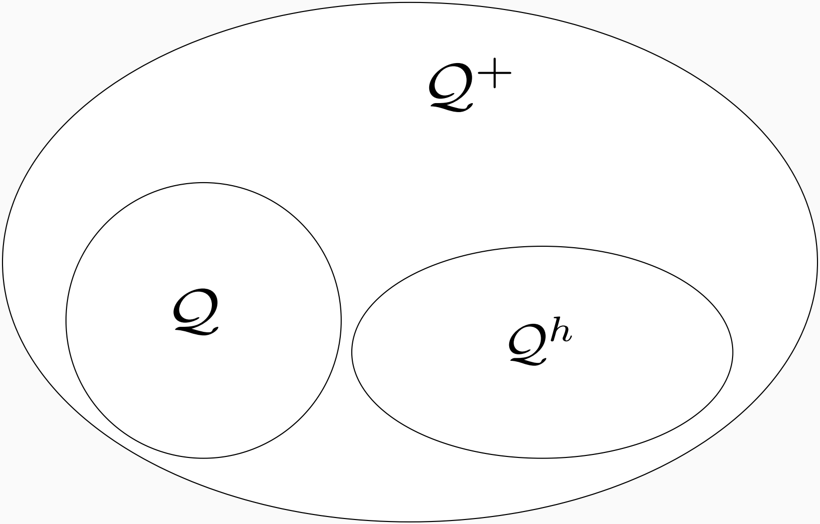


Diagram - Distribution over function-space



Sparse Variational Gaussian Processes

$$q^\star = \underset{q \in \mathcal{Q}}{\arg\max}\ \mathbb{E}_{q(f)}[\log p(\mathbf{y}|f)] - \mathsf{KL}\big(q|p\big)$$

Sparse Variational Gaussian Processes

$$q^\star = \underset{q \in \mathcal{Q}}{\arg\max}\ \mathbb{E}_{q(f)}[\log p(\mathbf{y}|f)] - \mathsf{KL}\big(q|p\big)$$

Decoupled Sparse Variational Gaussian Processes

$$q^\star = \underset{q \in \mathcal{Q}^+}{\arg\max}\ \mathbb{E}_{q(f)}[\log p(\mathbf{y}|f)] - \mathsf{KL}\big(q|p\big)$$

Variational Inference in Different Families

Sparse Variational Gaussian Processes

$$q^\star = \underset{q \in \mathcal{Q}}{\arg\max}\ \mathbb{E}_{q(f)}[\log p(\mathbf{y}|f)] - \mathsf{KL}\big(q|p\big)$$

Decoupled Sparse Variational Gaussian Processes

$$q^\star = \underset{q \in \mathcal{Q}^+}{\arg\max}\ \mathbb{E}_{q(f)}[\log p(\mathbf{y}|f)] - \mathsf{KL}\big(q|p\big)$$

Fixed Mean Sparse Variational Gaussian Processes

$$q^\star = \underset{q \in \mathcal{Q}^h}{\arg\max}\ \mathbb{E}_{q(f)}[\log p(\mathbf{y}|f)] - \mathsf{KL}\big(q|p\big)$$

Optimizing the ELBO in the Hilbert space:

$$q^\star = \underset{q \in \mathcal{Q}^+}{\arg\max} \; \mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right) \; .$$

where

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

and $\mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right]$ can be computed in regression and estimated in classification.

Variational Optimization

Optimizing the ELBO in the Hilbert space:

$$q^\star = \underset{q \in \mathcal{Q}^h}{\arg\max}\ \mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right) .$$

where

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

and $\mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right]$ can be computed in regression and estimated in classification.

$$q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right)$$

$$\boxed{q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \mathbb{E}_{q(f)} \left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right)}$$

The kernel $\kappa_\theta(\cdot, \cdot)$ may have **hyper-parameters** θ that affect q.

$$q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \mathbb{E}_{q(f)} \left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right)$$

The kernel $\kappa_\theta(\cdot, \cdot)$ may have **hyper-parameters** θ that affect q.

In regression: $\kappa_\theta(\cdot, \cdot) \to 0$ is a **local optima** in $\mathcal{Q}^h$.

$$\boxed{q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right)}$$

The kernel $\kappa_\theta(\cdot,\cdot)$ may have **hyper-parameters** θ that affect q.

In regression: $\kappa_\theta(\cdot,\cdot) \to 0$ is a **local optima** in $\mathcal{Q}^h$.

Solution: α-divergences.

$$\boxed{q^{\star} = \arg\max_{q \in \mathcal{Q}^h} \; \mathbb{E}_{q(f)}\left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right)}$$

$$\boxed{q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \mathbb{E}_{q(f)} \left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right)}$$

$$\Downarrow$$

$$\boxed{q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \frac{1}{\alpha} \log \; \mathbb{E}_{q(f)} \left[p(\mathbf{y}|f)^\alpha\right] - \mathsf{KL}\left(q|p\right)}$$

$$q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \mathbb{E}_{q(f)} \left[\log p(\mathbf{y}|f)\right] - \mathsf{KL}\left(q|p\right)$$

$$\Downarrow$$

$$q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \frac{1}{\alpha} \log \; \mathbb{E}_{q(f)} \left[p(\mathbf{y}|f)^\alpha\right] - \mathsf{KL}\left(q|p\right)$$

$$\Downarrow$$

$$q^\star = \arg\max_{q \in \mathcal{Q}^h} \; \text{Train LL} - \mathsf{KL}\left(q|p\right)$$

Intuitive Recap

1. Learn a **optimal deterministic** model h.

Intuitive Recap

1. Learn a **optimal deterministic** model h.
2. Define a **Sparse Variational GP**.

Intuitive Recap

1. Learn a **optimal deterministic** model h.
2. Define a **Sparse Variational GP**.
3. **Decouple the inducing locations** from the mean and covariance.

Intuitive Recap

1. Learn a **optimal deterministic** model h.
2. Define a **Sparse Variational GP**.
3. **Decouple the inducing locations** from the mean and covariance.
4. Consider the **subspace with fixed mean** h.

Intuitive Recap

1. Learn a **optimal deterministic** model h.
2. Define a **Sparse Variational GP**.
3. **Decouple the inducing locations** from the mean and covariance.
4. Consider the **subspace with fixed mean** h.
5. Train the (non-fixed) parameters using **function-space VI** and mini-batch optimization.

Intuitive Recap

1. Learn a **optimal deterministic** model h.
2. Define a **Sparse Variational GP**.
3. **Decouple the inducing locations** from the mean and covariance.
4. Consider the **subspace with fixed mean** h.
5. Train the (non-fixed) parameters using **function-space VI** and mini-batch optimization.
6. The resulting method **provides uncertainty estimation** for the deterministic model.

Results in Regression Problems

	Airline		Year		Taxi	
Model	NLL	CRPS	NLL	CRPS	NLL	CRPS
Deterministic	5.087	18.436	3.674	5.056	3.763	3.753
LLA Diag	5.096	**18.317**	3.650	4.957	3.714	3.979
LLA KFAC	5.097	**18.317**	3.650	4.955	3.705	3.977
LLA*	5.097	18.319	3.650	**4.954**	3.718	**3.975**
LLA* KFAC	5.097	**18.317**	3.650	**4.954**	3.718	3.976
ELLA	5.086	18.437	3.674	5.056	3.753	3.754
VaLLA	**4.923**	18.610	**3.527**	5.071	**3.287**	3.968
This Method	**4.903**	**17.552**	**3.485**	**4.721**	**3.208**	**3.493**

Limitations

$$q^\star = \arg\max_{q \in \mathcal{Q}^h} \text{ Train LL} - \text{KL}\left(q|p\right)$$

Limitations

$$q^\star = \arg\max_{q \in \mathcal{Q}^h} \text{ Train LL} - \text{KL}\left(q|p\right)$$

Regression

Zero Uncertainty $\implies$ Train LL $= -\infty$.

Limitations

$$q^\star = \arg\max_{q \in \mathcal{Q}^h} \text{ Train LL} - \text{KL}\left(q|p\right)$$

Regression

Zero Uncertainty $\implies$ Train LL $= -\infty$.

Multi-class Classification

Zero Uncertainty $\implies$ Train LL $= h$ Train LL.

Conclusions

1. Sparse GPs can be **characterized** in the RKHS.

Conclusions

1. Sparse GPs can be **characterized** in the RKHS.
2. Sparse GPs can be generalized to **decouple the inducing points**.

Conclusions

1. Sparse GPs can be **characterized** in the RKHS.
2. Sparse GPs can be generalized to **decouple the inducing points**.
3. There exists a **subspace** with posterior mean h.

Conclusions

1. Sparse GPs can be **characterized** in the RKHS.
2. Sparse GPs can be generalized to **decouple the inducing points**.
3. There exists a **subspace** with posterior mean h.
4. Preliminary results on **regression** are promising.

Conclusions

1. Sparse GPs can be **characterized** in the RKHS.
2. Sparse GPs can be generalized to **decouple the inducing points**.
3. There exists a **subspace** with posterior mean h.
4. Preliminary results on **regression** are promising.
5. Limited on **multi-class classification**.

Thank you for your attention!