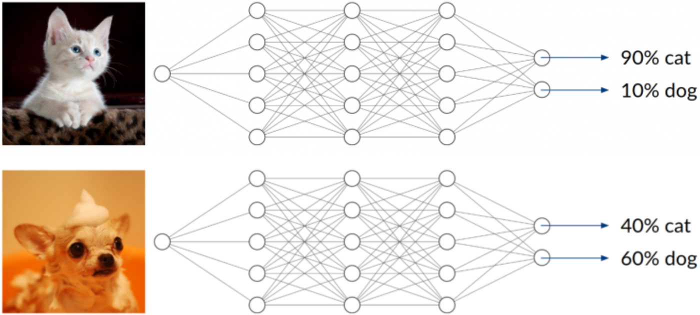
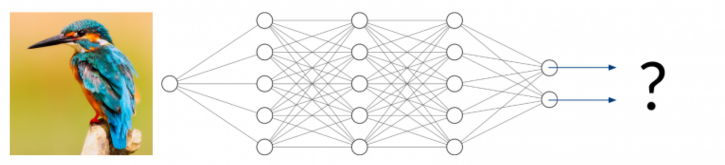
Fixed-Mean Gaussian Processes for post-hoc Bayesian Deep Learning

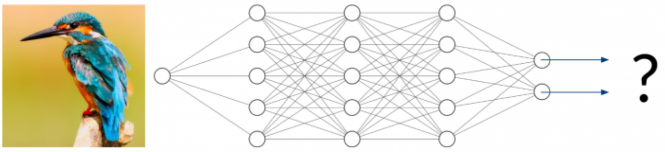
Luis Antonio Ortega Andrés
Simón Rodríguez Santana
Daniel Hernández Lobato

October 21, 2024

Autonomous University of Madrid

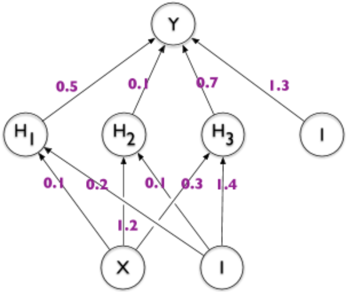

90% cat
10% dog
40% cat
60% dog

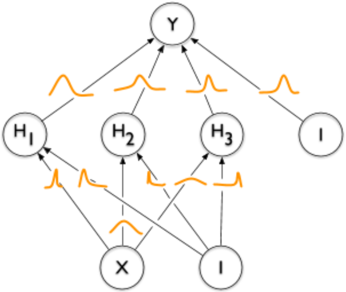




Deep learning methods are unable to quantify the uncertainty of their predictions!

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. Fixed-Mean Gaussian Processes with **posterior mean** h.

   Same Performance - Uncertainty Estimation

Approach

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

   High Performance - No Uncertainty

2. Fixed-Mean Gaussian Processes with **posterior mean** h.

   Same Performance - Uncertainty Estimation

3. Optimize parameters using function-space VI.

Uncertainty Estimation in function-space

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

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

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

Gaussian Processes

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

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

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

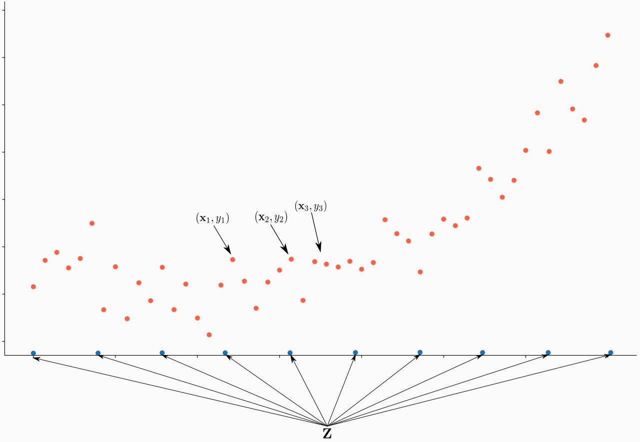
$$p(y^\star|\mathbf{x}^\star, \mathbf{X}, \mathbf{y}) = \mathcal{N}(m^\star(\mathbf{x}^\star), K^\star(\mathbf{x}^\star, \mathbf{x}^\star))\,.$$

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

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

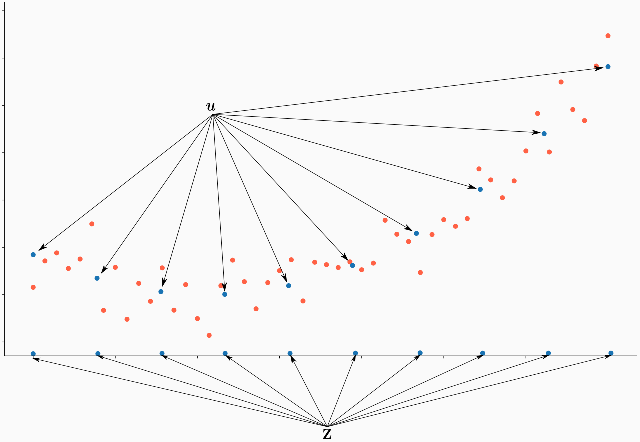
Gaussian noise with variance σ^2 is considered for the targets

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

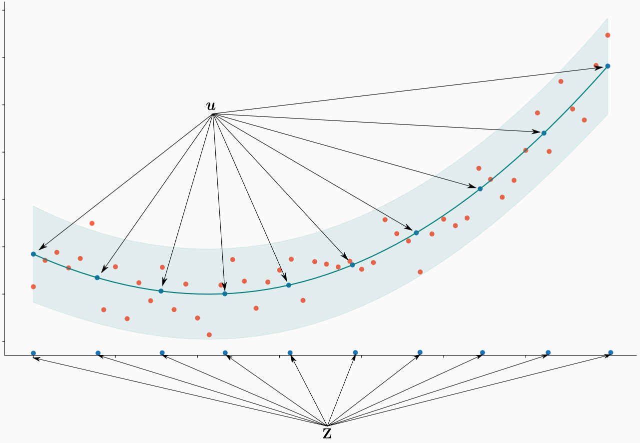


Sparse Variational Gaussian Processes

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



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



Hilbert Spaces and RKHS

An **RKHS** $\mathcal{H}$ is a Hilbert space of functions satisfying the **reproducing property**:

$$\forall \mathbf{x} \in \mathcal{X},\ \exists \phi_{\mathbf{x}} \in \mathcal{H}, \quad \text{such that} \quad \forall g \in \mathcal{H},\ g(\mathbf{x}) = \langle \phi_{\mathbf{x}}, g \rangle_{\mathcal{H}}\ .$$

An **RKHS** $\mathcal{H}$ is a Hilbert space of functions satisfying the **reproducing property**:

$$\forall \mathbf{x} \in \mathcal{X},\ \exists \phi_{\mathbf{x}} \in \mathcal{H}, \quad \text{such that} \quad \forall g \in \mathcal{H},\ g(\mathbf{x}) = \langle \phi_{\mathbf{x}}, g \rangle_{\mathcal{H}}\ .$$

Let $\mathcal{H}_0(\mathcal{X})$ be the linear span of K on $\mathcal{X}$ defined as

$$\mathcal{H}_0(\mathcal{X}) = \left\{ \sum_{i=1}^{n} a_i K(\cdot, \mathbf{x}_i)\ :\ n \in \mathbb{N},\ a_i \in \mathbb{R},\ \mathbf{x}_i \in \mathcal{X} \right\},$$

by Moore–Aronszajn theorem $\mathcal{H} := \overline{\mathcal{H}_0(\mathcal{X})}$ is the only Hilbert space verifying the reproducing property as $\phi_{\mathbf{x}} = K(\cdot, \mathbf{x}),\ \forall \mathbf{x} \in \mathcal{X}$.

Dual representation of Gaussian Processes

Given a **Gaussian process** $f \sim \mathcal{GP}(m, K)$, and the RKHS $\mathcal{H}$ defined by its kernel K. If $m \in \mathcal{H}$, the GP is equivalent to a Gaussian measure on a Banach space $\mathcal{B}$ which contains the RKHS $\mathcal{H}$.

There exists $\mu \in \mathcal{H}$ and a linear semi-definite positive operator $\Sigma : \mathcal{H} \to \mathcal{H}$ such that, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $\exists \phi_{\mathbf{x}}, \phi_{\mathbf{x}'} \in \mathcal{H}$, verifying

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

$\mathcal{N}(\mu, \Sigma)$ is a Gaussian measure in $\mathcal{B}$.

CA. Cheng and B. Boots. "Variational inference for Gaussian process models with linear complexity"

A **GP prior** is recovered with $\mu = 0$ and $\Sigma = I$:

$$m(\mathbf{x}) = \langle \phi_{\mathbf{x}}, \mu \rangle = 0, \quad K(\mathbf{x}, \mathbf{x}') = \langle \phi_{\mathbf{x}}, \Sigma(\phi_{\mathbf{x}'}) \rangle \ .$$

A **GP prior** is recovered with $\mu = 0$ and $\Sigma = I$:

$$m(\mathbf{x}) = \langle \phi_{\mathbf{x}}, \mu \rangle = 0, \quad K(\mathbf{x}, \mathbf{x}') = \langle \phi_{\mathbf{x}}, \Sigma(\phi_{\mathbf{x}'}) \rangle \ .$$

A **regression GP posterior** is recovered with

$$\mu = \sum_{i=1}^{N} \alpha_i \phi_{\mathbf{x}_i} \quad \text{and} \quad \Sigma(\phi) = \phi - \sum_{i=1}^{N} \sum_{j=1}^{N} \phi_{\mathbf{x}_i} \Lambda_{i,j} \left\langle \phi_{\mathbf{x}_j}, \phi \right\rangle \ ,$$

where $\boldsymbol{\Lambda} = (K(\mathbf{X}, \mathbf{X}) + \sigma^2 \boldsymbol{I})^{-1} \in \mathbb{R}^{N \times N}$ and $\boldsymbol{\alpha} = \boldsymbol{\Lambda} \boldsymbol{y} \in \mathbb{R}^N$.

A SVGP is equivalent to restricting the mean and covariance functions in the RKHS to

$$\tilde{\mu}_{\boldsymbol{a}} = \sum_{m=1}^{M} a_m \phi_{\mathbf{z}_m}\,, \quad \tilde{\Sigma}_{\boldsymbol{A}}(\phi) = \phi + \sum_{i=1}^{M}\sum_{j=1}^{M} \phi_{\mathbf{z}_i} A_{i,j} \left\langle \phi_{\mathbf{z}_j}, \phi \right\rangle\,,$$

where $\boldsymbol{a} = (a_1, \ldots, a_M)^T \in \mathbb{R}^M$, $\boldsymbol{A} = (A_{ij}) \in \mathbb{R}^{M \times M}$ such that $\tilde{\Sigma} \geq 0$ and $\phi_{\mathbf{z}} \in \mathcal{H}$, $\forall \mathbf{z} \in \mathbf{Z}$.

A SVGP with $q(\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ is *built* with

$$\boldsymbol{a} = K(\mathbf{Z}, \mathbf{Z})^{-1} \boldsymbol{\mu}\,, \quad \boldsymbol{A} = K(\mathbf{Z}, \mathbf{Z})^{-1} \boldsymbol{S} K(\mathbf{Z}, \mathbf{Z})^{-1} - K(\mathbf{Z}, \mathbf{Z})^{-1}$$

CA. Cheng and B. Boots. “Variational inference for Gaussian process models with linear complexity”

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

$$\tilde{\mu}_{\alpha,\boldsymbol{a}} = \sum_{m=1}^{M_\alpha} a_m \phi_{\mathbf{z}_{\alpha,m}}$$

$$\tilde{\Sigma}_{\beta,\boldsymbol{A}}(\phi) = \phi + \sum_{i=1}^{M_\beta} \sum_{j=1}^{M_\beta} \phi_{\mathbf{z}_{\beta,i}} A_{i,j} \left\langle \phi_{\mathbf{z}_{\beta,j}}, \phi \right\rangle \ .$$

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

CA. Cheng and B. Boots. “Variational inference for Gaussian process models with linear complexity”

Fixing the Mean Function

Let $\mathcal{Z} \subset \mathcal{X}$ any compact subset of the input space. If the kernel is **universal**, for any function $h \in C(\mathcal{Z})$ and $\epsilon > 0$, there exists $M_\alpha > 0$, a set of inducing locations $\{\mathbf{z}_1, \ldots, \mathbf{z}_{M_\alpha}\} \subset \mathcal{Z}$, and scalar values $a_1, \ldots, a_{M_\alpha}$ such that

$$m(\mathbf{x}) = \langle \phi_{\mathbf{x}}, \tilde{\mu}_{\alpha,\boldsymbol{a}} \rangle = \sum_{m=1}^{M_\alpha} a_m K(\mathbf{x}, \mathbf{z}_m)$$

verifies

$$\left\| h(\mathbf{x}) - m(\mathbf{x}) \right\|_{\mathcal{Z}} \leq \epsilon \,.$$

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 $\tilde{\Sigma} \geq 0$).

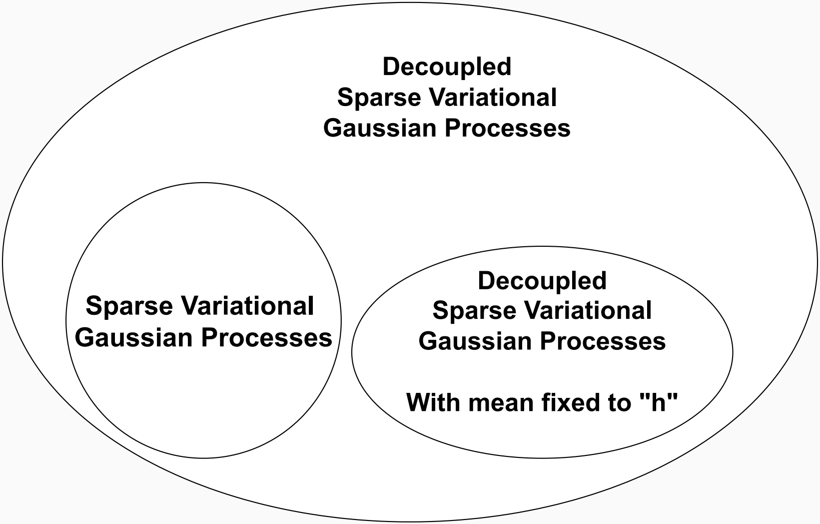
Distributions over function-space with fixed mean to h.

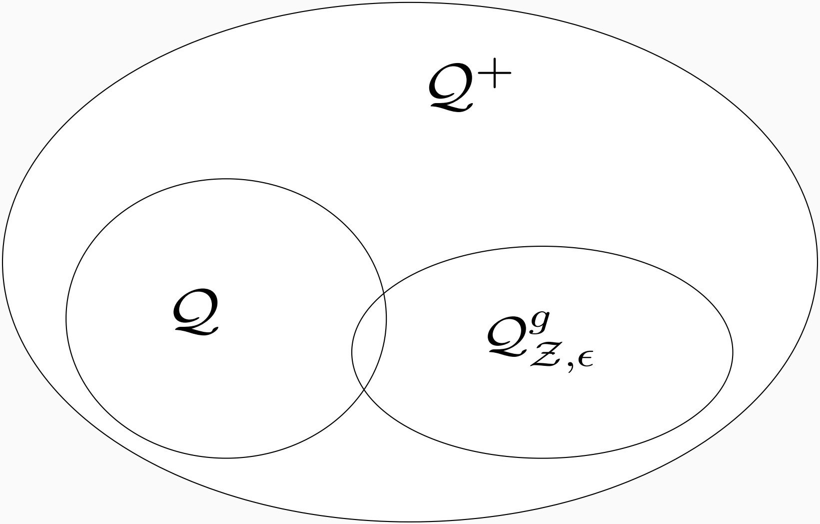
Parameters: $\mathbf{Z}_\beta \subset \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{M_\beta \times M_\beta}$ (such that $\tilde{\Sigma} \geq 0$).

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

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

Diagram - Distribution over function-space




$\mathcal{Q}^+$
$\mathcal{Q}$
$\mathcal{Q}^g_{\mathcal{Z},\epsilon}$

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

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

Decoupled Sparse Variational Gaussian Processes

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

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

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

Variational Optimization

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

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

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:

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

where

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

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

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.

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.

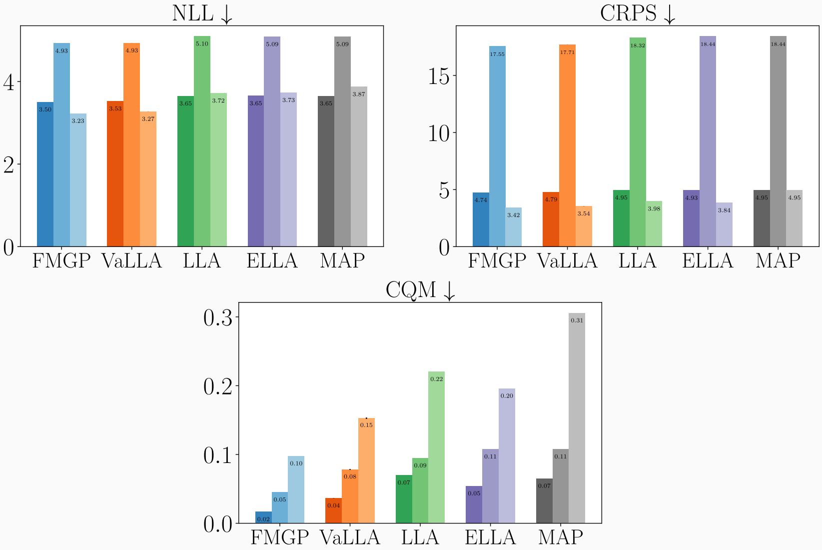
Intuitive Recap

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.

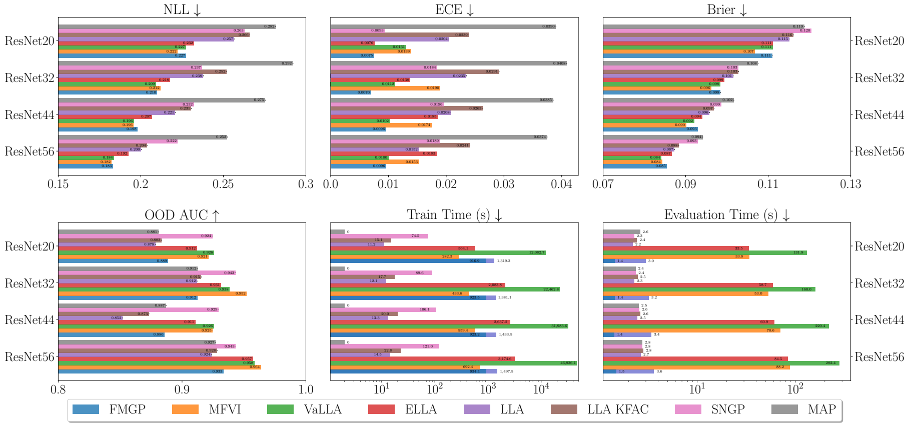
Intuitive Recap

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



Results in Cifar10 Problems



Results in Imagenet

Model	Method	NLL	ECE	Train Time	Test Time
ResNet18	MAP	$\mathbf{1.247 \pm 0.000}$	0.026 ± 0.000	$\mathbf{0.000 \pm 0.000}$	$\mathbf{5.058 \pm 0.029 \times 10^2}$
	ELLA	1.248 ± 0.000	$\mathbf{0.025 \pm 0.000}$	$\mathbf{7.890 \pm 0.275 \times 10^3}$	$8.060 \pm 0.010 \times 10^2$
	FMGP	1.248 ± 0.001	$\mathbf{0.015 \pm 0.001}$	$1.835 \pm 0.099 \times 10^4$	$\mathbf{7.324 \pm 0.001 \times 10^2}$
	MFVI	$\mathbf{1.242 \pm 0.001}$	0.040 ± 0.000	$7.602 \pm 0.032 \times 10^4$	$3.773 \pm 0.308 \times 10^4$
ResNet34	MAP	$\mathbf{1.081 \pm 0.000}$	0.035 ± 0.000	$\mathbf{0.000 \pm 0.000}$	$\mathbf{5.088 \pm 0.004 \times 10^2}$
	ELLA	1.082 ± 0.000	$\mathbf{0.034 \pm 0.000}$	$\mathbf{1.201 \pm 0.373 \times 10^4}$	$1.087 \pm 0.018 \times 10^3$
	FMGP	$\mathbf{1.077 \pm 0.000}$	$\mathbf{0.016 \pm 0.000}$	$1.942 \pm 0.103 \times 10^4$	$\mathbf{8.563 \pm 0.011 \times 10^2}$
ResNet50	MAP	$\mathbf{0.962 \pm 0.000}$	0.037 ± 0.000	$\mathbf{0.000 \pm 0.000}$	$\mathbf{4.954 \pm 0.010 \times 10^2}$
	ELLA	$\mathbf{0.962 \pm 0.000}$	$\mathbf{0.036 \pm 0.000}$	$2.997 \pm 1.215 \times 10^4$	$1.954 \pm 0.018 \times 10^3$
	FMGP	$\mathbf{0.958 \pm 0.001}$	$\mathbf{0.018 \pm 0.001}$	$\mathbf{2.543 \pm 0.046 \times 10^4}$	$\mathbf{1.100 \pm 0.010 \times 10^3}$
ResNet101	MAP	$\mathbf{0.912 \pm 0.000}$	0.049 ± 0.000	$\mathbf{0.000 \pm 0.000}$	$\mathbf{5.059 \pm 0.001 \times 10^2}$
	ELLA	0.913 ± 0.000	$\mathbf{0.048 \pm 0.000}$	$4.464 \pm 1.649 \times 10^4$	$2.808 \pm 0.001 \times 10^3$
	FMGP	$\mathbf{0.900 \pm 0.000}$	$\mathbf{0.030 \pm 0.001}$	$\mathbf{2.654 \pm 0.064 \times 10^4}$	$\mathbf{1.134 \pm 0.001 \times 10^3}$
ResNet152	MAP	$\mathbf{0.876 \pm 0.000}$	0.050 ± 0.000	$\mathbf{0.000 \pm 0.000}$	$\mathbf{6.324 \pm 0.004 \times 10^2}$
	ELLA	0.877 ± 0.000	$\mathbf{0.048 \pm 0.000}$	$6.820 \pm 0.526 \times 10^4$	$3.877 \pm 0.007 \times 10^3$
	FMGP	$\mathbf{0.865 \pm 0.001}$	$\mathbf{0.024 \pm 0.001}$	$\mathbf{2.973 \pm 0.069 \times 10^4}$	$\mathbf{1.267 \pm 0.002 \times 10^3}$

Results on Molecular Property Prediction

Method	NLL	CRPS
MAP	-1.76 ± 0.016	0.0221 ± 0.00
LLA	-1.78 ± 0.021	$\mathbf{0.0218 \pm 0.00}$
ELLA	$\mathbf{-1.80 \pm 0.013}$	0.0219 ± 0.00
FMGP	$\mathbf{-1.85 \pm 0.017}$	$\mathbf{0.0216 \pm 0.00}$

Table 1: Results on QM9 dipole moment prediction task.

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. Obtained results are promising.

Thank you for your attention!

Dual representation of Gaussian Processes

A **Gaussian process** $f \sim \mathcal{GP}(m, K)$ has a **dual representation** in a RKHS $\mathcal{H}$ *from a different unknown kernel* $\tilde{K}$.

There exists $\mu \in \mathcal{H}$ and a linear semi-definite positive operator $\Sigma : \mathcal{H} \to \mathcal{H}$ such that, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $\exists \phi_{\mathbf{x}}, \phi_{\mathbf{x}'} \in \mathcal{H}$, verifying

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

$\mathcal{N}(\mu, \Sigma)$ is a Gaussian measure in $\mathcal{H}$.

I. Holmes and A. N. Sengupta, "The Gaussian radon transform and machine learning"
CA. Cheng and B. Boots, "Incremental variational sparse Gaussian process regression"

Regularization

In standard sparse GPs, tuning hyper-parameters involves **balancing** the fit of the mean to training data against reducing the model's predictive variance.

We consider another Gaussian measure $q^\star \in \mathcal{Q}$ that shares q's parameters but also incorporates $\boldsymbol{a} \in \mathbb{R}^{M_\beta}$ and $\mathbf{Z} = \mathbf{Z}_\beta$ as additional parameters for its predictive mean.

$$\begin{aligned}\mathcal{L}(\boldsymbol{a}, \boldsymbol{A}, \mathbf{Z}, \theta) = & \underbrace{\mathbb{E}_{q(f)}[\log p(\mathbf{y}|f)] - \mathrm{KL}\big(q\big|p\big)}_{\mathrm{ELBO}(q)} \\ & + \underbrace{\mathbb{E}_{q^\star(f)}[\log p(\mathbf{y}|f)] - \mathrm{KL}\big(q^\star\big|p\big)}_{\mathrm{ELBO}(q^\star)}\end{aligned}$$