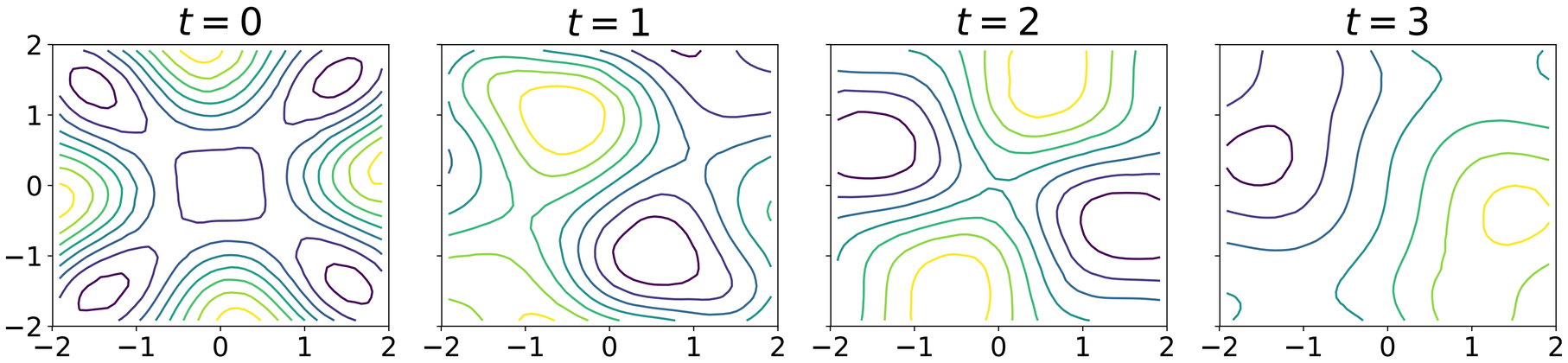
Scalable Variational Gaussian Processes via Harmonic Kernel Decomposition



Shengyang Sun, Jiaxin Shi, Andrew Gordon Wilson, Roger Grosse

Gaussian Processes

- *Given a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a Gaussian process $\mathcal{GP}(0, k)$ is a distribution of functions. For any finite set of inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ their function values satisfy a multivariate Gaussian distribution,*

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}),$$

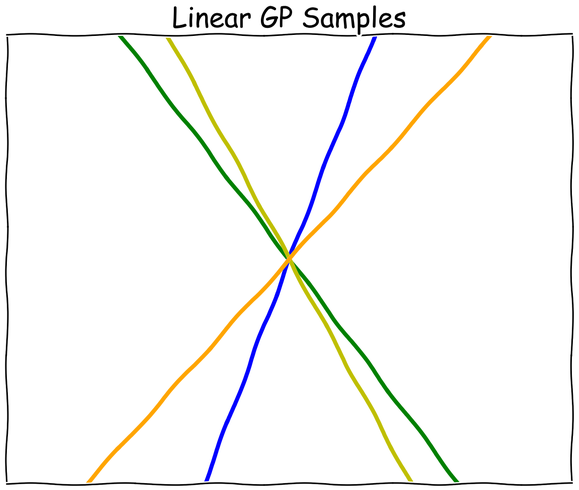
Where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$

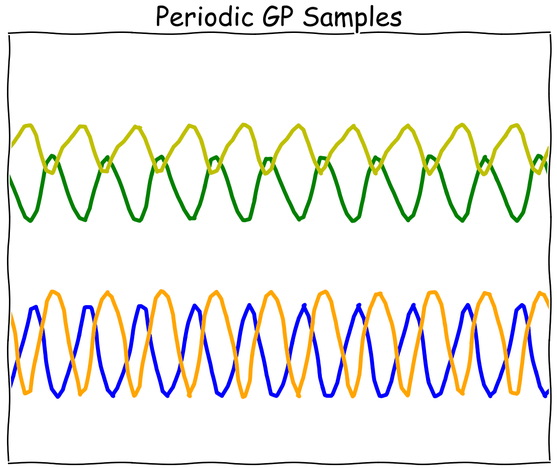
Gaussian Processes

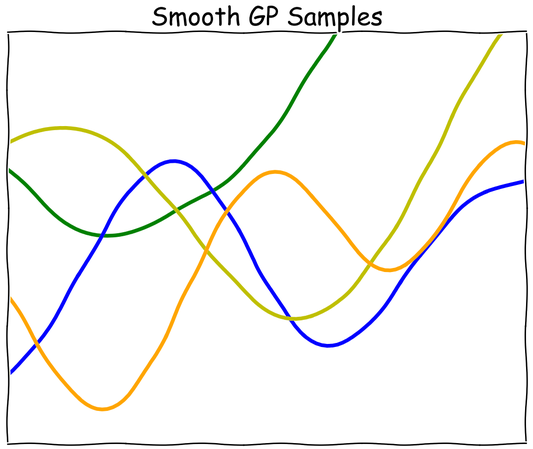
- *Given a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a Gaussian process $\mathcal{GP}(0, k)$ is a distribution of functions. For any finite set of inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ their function values satisfy a multivariate Gaussian distribution,*

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}),$$

Where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$







Gaussian Processes

- *Given a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a Gaussian process $\mathcal{GP}(0, k)$ is a distribution of functions. For any finite set of inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ their function values satisfy a multivariate Gaussian distribution,*

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}),$$

Where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$

- Under a Gaussian likelihood, the posterior is a multivariate Gaussian, for $\mathbf{x}_\star$

$$\mathbf{f}_\star | \mathbf{y} \sim \mathcal{N}(\mathbf{K}_{\star f}(\mathbf{K}_{ff} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \mathbf{K}_{\star\star} - \mathbf{K}_{\star f}(\mathbf{K}_{ff} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{f\star})$$

- Inferring the exact posterior requires cubic computations to the dataset size.

Variational Inference of GPs

- Consider $\mathbf{u} = f(\mathbf{z}_{1:m})$, $\mathbf{z}_{1:m}$ are inducing points that summarizes the dataset.
- Define a variational posterior over the augmented space,

$$q(f, \mathbf{u}) = p(f|\mathbf{u})q(\mathbf{u})$$

- Variational Inference (Titsias, 2009) optimizes the variational posterior by maximizing the Evidence Lower Bound (ELBO),

$$\begin{aligned}\log p(\mathcal{D}) \geq & \mathbb{E}_{q(f,\mathbf{u})}[\log p(\mathcal{D}|f,\mathbf{u})] - \mathrm{KL}[q(f,\mathbf{u})\|p(f,\mathbf{u})] \\ = & \mathbb{E}_{q(f,\mathbf{u})}[\log p(\mathcal{D}|f,\mathbf{u})] - \mathrm{KL}[q(\mathbf{u})\|p(\mathbf{u})]\end{aligned}$$

stochastic estimations cubic of m computations

Harmonic Kernel Decomposition

- Given a transformation $G : \mathcal{X} \to \mathcal{X}$, we consider the sequence of kernel values,

$$[k(\mathbf{x}, G^0(\mathbf{x}')), k(\mathbf{x}, G^1(\mathbf{x}'))..., k(\mathbf{x}, G^{T-1}(\mathbf{x}'))],$$

Discrete Fourier transform

$$[k_0(\mathbf{x}, \mathbf{x}'), k_1(\mathbf{x}, \mathbf{x}'), \ldots, k_{T-1}(\mathbf{x}, \mathbf{x}')]$$

Harmonic Kernel Decomposition

- Given a transformation $G : \mathcal{X} \to \mathcal{X}$, we consider the sequence of kernel values,

$$[k(\mathbf{x}, G^0(\mathbf{x}')), k(\mathbf{x}, G^1(\mathbf{x}'))..., k(\mathbf{x}, G^{T-1}(\mathbf{x}'))],$$

Discrete Fourier transform

$$[k_0(\mathbf{x}, \mathbf{x}'), k_1(\mathbf{x}, \mathbf{x}'), \ldots, k_{T-1}(\mathbf{x}, \mathbf{x}')]$$

G is T-cyclic

$$\forall \mathbf{x} \in \mathcal{X},\ G^T(\mathbf{x}) := \overbrace{G \circ \cdots \circ G}^{T}(\mathbf{x}) = \mathbf{x}.$$

The kernel is invariant to G

$$\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X},\ k(G(\mathbf{x}), G(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}').$$

Each k_t is a Hermitian kernel.

Harmonic Kernel Decomposition

- The HKD is an orthogonal decomposition of kernels and RKHSs,

$$k(\mathbf{x}, \mathbf{x}') = \sum_{t=0}^{T-1} k_t(\mathbf{x}, \mathbf{x}') \qquad \mathcal{H}_k = \bigoplus_{t=0}^{T-1} \mathcal{H}_{k_t}$$

Harmonic Kernel Decomposition

- The HKD is an orthogonal decomposition of kernels and RKHSs,

$$k(\mathbf{x}, \mathbf{x}') = \sum_{t=0}^{T-1} k_t(\mathbf{x}, \mathbf{x}') \qquad \mathcal{H}_k = \bigoplus_{t=0}^{T-1} \mathcal{H}_{k_t}$$

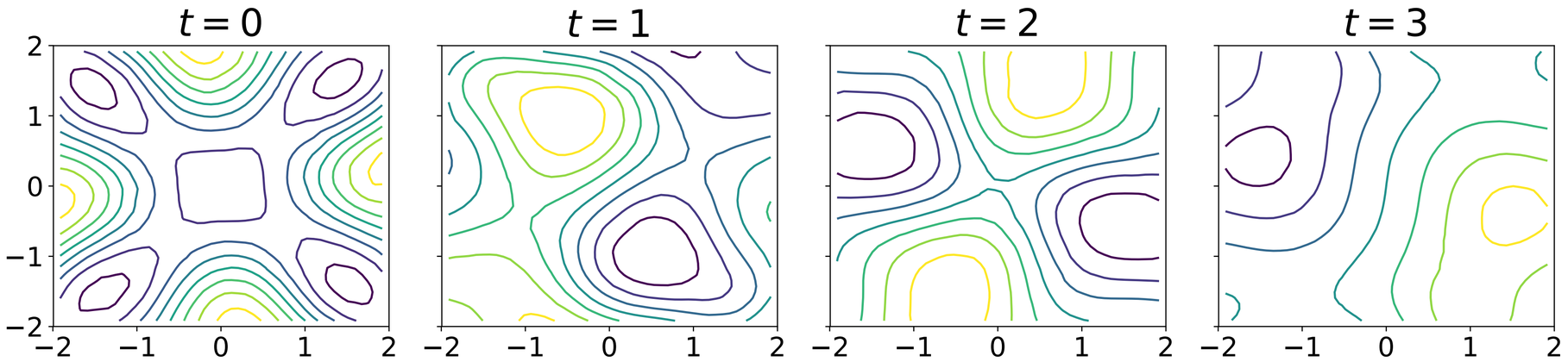
- The HKD is widely applicable to many kernels: RBF, Matérn, polynomial, periodic, …

Kernels k	Inner-Product	Stationary	Stationary
Input Space $\mathcal{X}$	Complex, Real	Real	Torus
Transformation G	Rotation, Reflection	Negation	Translation

Harmonic Variational Gaussian Processes

- From kernel decomposition to GP decomposition:

$$f = \sum_{t=0}^{T-1} f_t, \;\; f_t \sim \mathcal{GP}(0, k_t)$$



Harmonic Variational Gaussian Processes

- From kernel decomposition to GP decomposition:

$$f = \sum_{t=0}^{T-1} f_t,\ f_t \sim \mathcal{GP}(0, k_t)$$

- The HVGP introduces an independent variational posterior for each component GP,

$$f = \sum_{t=0}^{T-1} f_t,\ q_t(f_t, \mathbf{u}_t) = p_t(f_t|\mathbf{u}_t)q_t(\mathbf{u}_t)$$

- The variational posterior can be optimized by maximizing the ELBO,

$$\mathbb{E}_{q(f_0,\ldots,f_{T-1})}\left[\log p\left(\mathbf{y}|\sum_{t=0}^{T-1} f_t, \mathbf{X}\right)\right] - \sum_{t=0}^{T-1} \mathrm{KL}\left(q_t(\mathbf{u}_t)\|p_t(\mathbf{u}_t)\right)$$

Harmonic Variational Gaussian Processes

- HVGP: a scalable variational Gaussian process approximation

  Similar to SVGP:

  Large Datasets High Dimensions Trainable Inducing Points

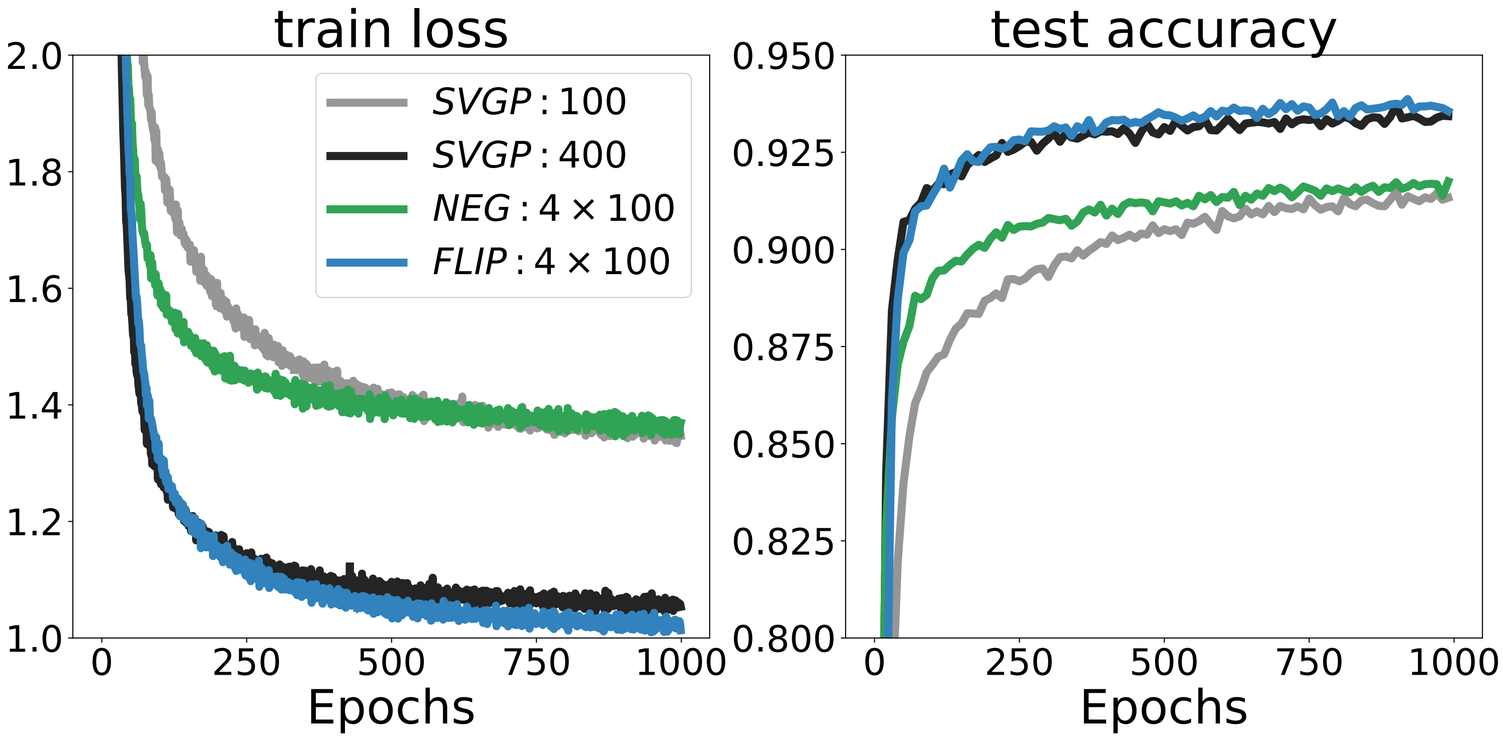
  Better than SVGP:

  More Inducing Points Lower Computational Costs Easier Parallelisms

Harmonic Variational Gaussian Processes

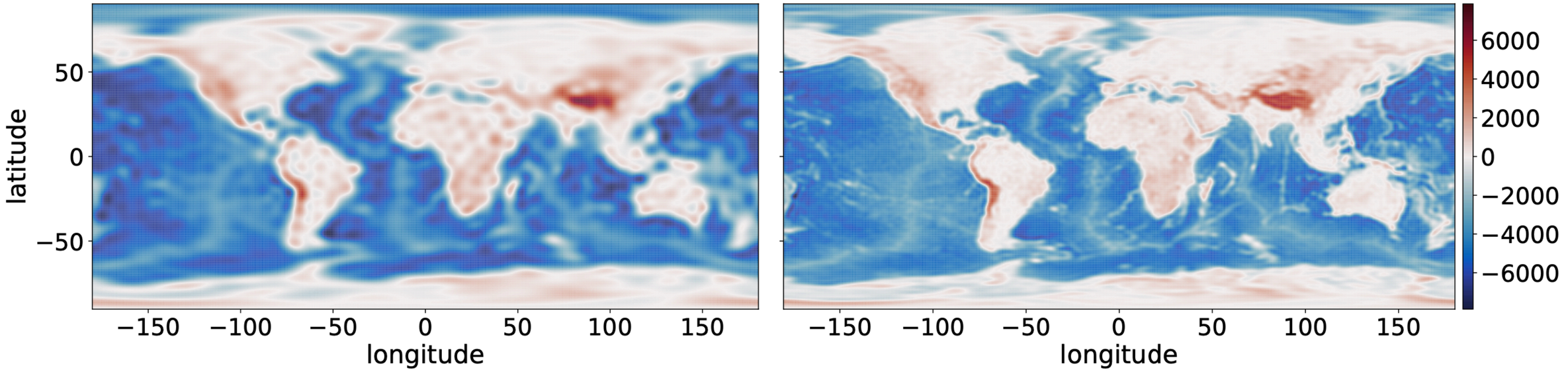
- Accurate Posterior Inference

- *Statement (Informal): HVGPs approximate the true posterior accurately when the input distribution has symmetry over the transformation* G



Harmonic Variational Gaussian Processes

HVGP enables large number of inducing points

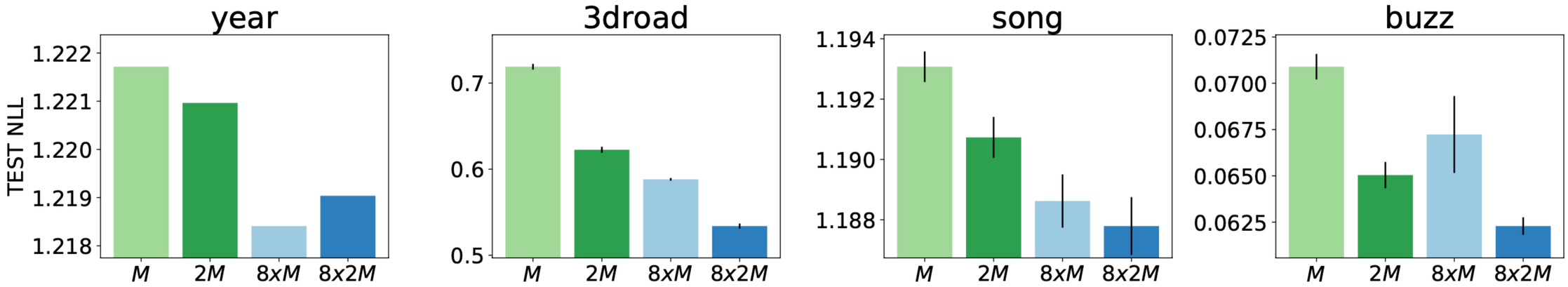


SVGP: 1000

HVGP: 13x1000

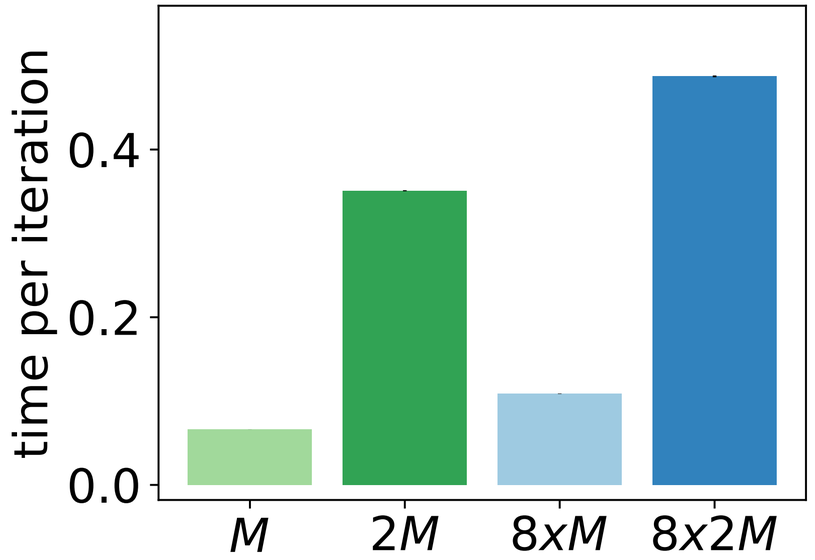
Harmonic Variational Gaussian Processes

Regression Benchmarks



Harmonic Variational Gaussian Processes

Regression Benchmarks



Harmonic Variational Gaussian Processes

HVGP achieves state-of-art performances on CIFAR10 among GPs

384x2, 1K	M	79.01±0.11	0.86±0.00	0.17
	2M	80.27±0.04	0.81±0.00	0.52
	M+M	79.98 ±0.21	0.80±0.01	0.46
	2xM	80.04±0.04	0.80±0.00	0.37
	4xM	**80.52±0.20**	**0.75±0.01**	0.37
384x3, 1K	M	82.41±0.08	0.73±0.01	0.40
	2M	-	-	-
	M+M	83.26±0.19	0.69±0.01	1.24
	2xM	**84.97±0.08**	0.60±0.00	0.90
	4xM	**84.85±0.11**	**0.58±0.00**	0.90

Thank you !

DFT meets kernel methods, leading to a scalable variational GP

Code: https://github.com/ssydasheng/Harmonic-Kernel-Decomposition