Scalable Gaussian Processes, with Guarantees: Kernel Approximations and Deep Feature Extractions

Petros Dellaportas

Joint work with Costis Daskalakis (MIT) and Aris Panos (UCL)

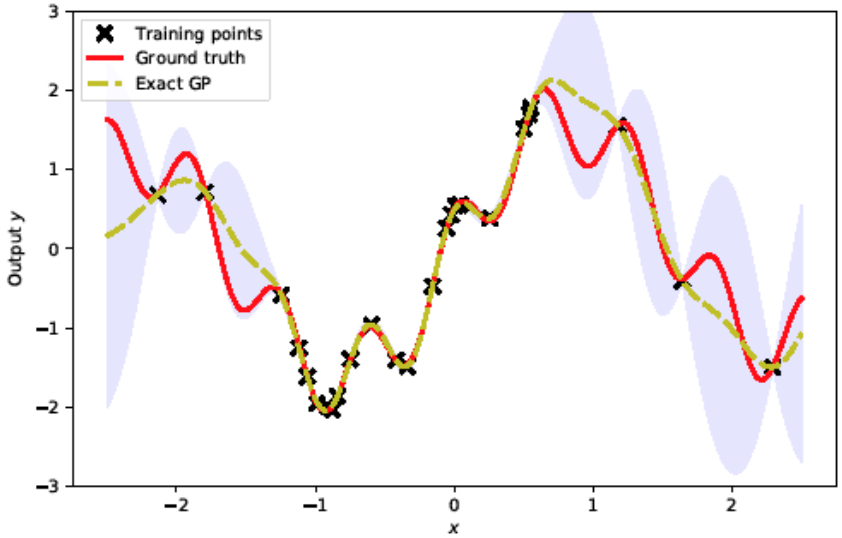
Gaussian processes

$$y = (y_i)_{i=1}^N \in \mathbb{R}^N$$

$$X = (\mathbf{x}_i)_{i=1}^N \in \mathbb{R}^{N \times D}$$

$$(f(\mathbf{x}_i))_{i=1}^N \sim MVN\left(0, K(k_\theta, X)\right)$$

$$K(k_\theta, X) := (k_\theta(\mathbf{x}_i, \mathbf{x}_j))_{ij}$$



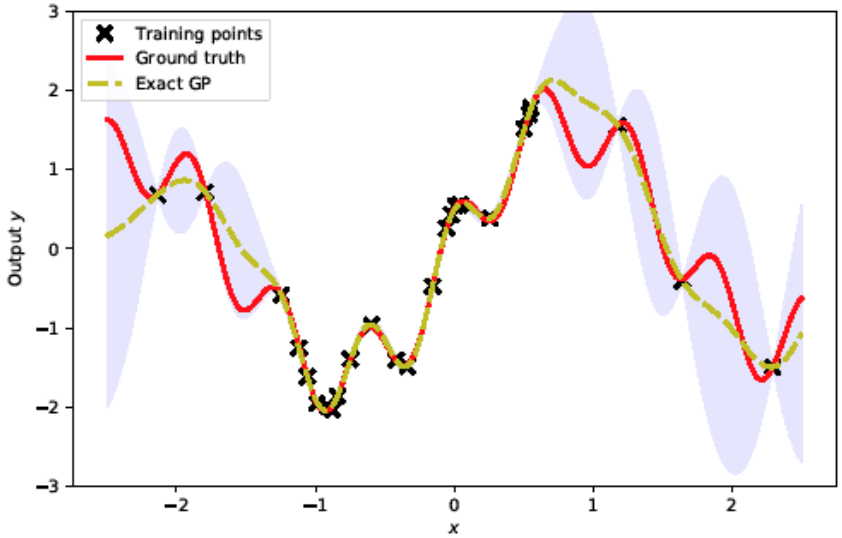
Gaussian processes

$$y = (y_i)_{i=1}^N \in \mathbb{R}^N$$

$$X = (\mathbf{x}_i)_{i=1}^N \in \mathbb{R}^{N \times D}$$

$$(f(\mathbf{x}_i))_{i=1}^N \sim MVN\left(0, K(k_\theta, X)\right)$$

$$K(k_\theta, X) := (k_\theta(\mathbf{x}_i, \mathbf{x}_j))_{ij}$$



Noisy observations:

$$A = K(k_\theta, X) + \sigma^2 I_N$$

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^\top A^{-1}\mathbf{y} - \frac{1}{2}\log|A| - \frac{N}{2}\log(2\pi)$$

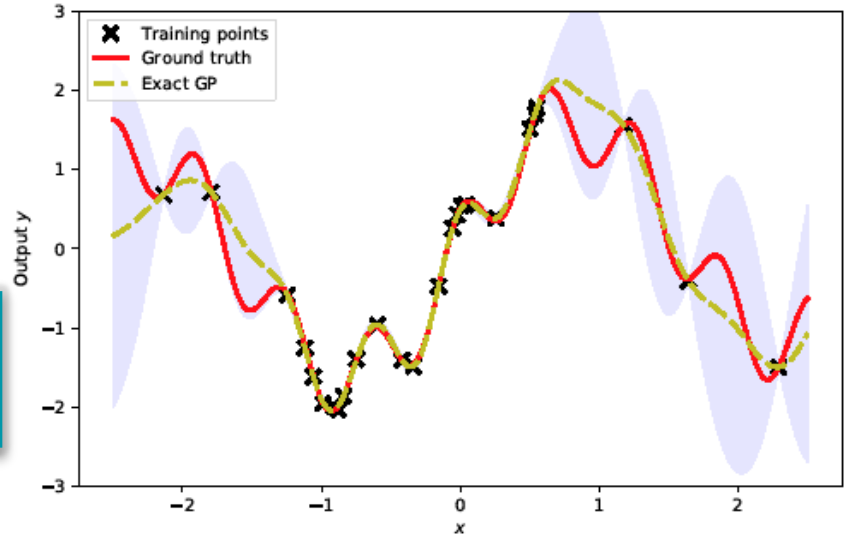
Gaussian processes

$$y = (y_i)_{i=1}^N \in \mathbb{R}^N$$

$$X = (\mathbf{x}_i)_{i=1}^N \in \mathbb{R}^{N \times D}$$

$$(f(\mathbf{x}_i))_{i=1}^N \sim MVN\,(0, K(k_\theta, X))$$

$$K(k_\theta, X) := (k_\theta(\mathbf{x}_i, \mathbf{x}_j))_{ij}$$



Noisy observations:

$$A = K(k_\theta, X) + \sigma^2 I_N$$

O(N^3)

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^\top A^{-1}\mathbf{y} - \frac{1}{2}\log|A| - \frac{N}{2}\log(2\pi)$$

Low rank approximations

$$y = (y_i)_{i=1}^N \in \mathbb{R}^N$$
$$X = (\mathbf{x}_i)_{i=1}^N \in \mathbb{R}^{N \times D}$$

Feature map

$$\phi : \mathbb{R}^D \to \mathbb{R}^r$$

$$k_\theta(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

$$K(k_\theta, X) = \Xi\Xi^\top$$

Ξ is an $N \times r$ matrix

$$A = \Xi\Xi^\top + \sigma^2 I_N$$

Low rank approximations

$$y = (y_i)_{i=1}^N \in \mathbb{R}^N$$
$$X = (\mathbf{x}_i)_{i=1}^N \in \mathbb{R}^{N \times D}$$

$$\phi : \mathbb{R}^D \to \mathbb{R}^r$$

$$k_\theta(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

$$K(k_\theta, X) = \Xi\Xi^\top$$

$$A = \Xi\Xi^\top + \sigma^2 I_N$$

$$A^{-1} = \sigma^{-2} I_N - \sigma^{-2}\Xi(\sigma^2 I_r + \Xi^\top \Xi)^{-1}\Xi^\top$$

$O(r^3 + r^3 N)$

Gaussian kernel

$$k_{\sigma_f^2,\Delta}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \Delta (\mathbf{x}_i - \mathbf{x}_j)),$$

where $\Delta = \mathrm{diag}(\epsilon_1^2, \ldots, \epsilon_D^2)$ contains the length scales along the D dimensions of the covariates, and σ_f^2 is the variance. The parameters of the kernel are $\theta = (\sigma_f^2, \Delta)$.

Low-rank approximations

(1) : Random Fourier features

(2) : Mercer expansion

Random Fourier expansion

Bochner's Theorem: Any continuous shift-invariant covariance function $k(\mathbf{x}_i,\mathbf{x}_j) = k(\mathbf{x}_i-\mathbf{x}_j)$ is positive definite iff can be written as the Fourier transform of a non-negative measure $p(\boldsymbol{\omega})$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \int p(\omega) \exp(i(\mathbf{x}_i - \mathbf{x}_j)^T \omega) d\omega$$

Random Fourier expansion

Bochner's Theorem: Any continuous shift-invariant covariance function $k(\mathbf{x}_i,\mathbf{x}_j) = k(\mathbf{x}_i-\mathbf{x}_j)$ is positive definite iff can be written as the Fourier transform of a non-negative measure $p(\boldsymbol{\omega})$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \int p(\omega) \exp(i(\mathbf{x}_i - \mathbf{x}_j)^T \omega) d\omega$$

Gaussian kernel

Gaussian density

Use only real part

Random Fourier expansion

- Sample $\frac{r}{2}$ spectral frequencies $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{\frac{r}{2}}$ from the spectral density $p(\boldsymbol{\eta})$ of the stationary kernel $k_\theta(\cdot, \cdot)$
- Create the feature map $\phi(\mathbf{z}) : \mathbb{R}^d \to \mathbb{R}^r$, defined by the vector

$$\sqrt{\frac{2}{r}}[\cos(\boldsymbol{\eta}_1^\top \mathbf{z}), \ldots, \cos(\boldsymbol{\eta}_{\frac{r}{2}}^\top \mathbf{z}), \sin(\boldsymbol{\eta}_1^\top \mathbf{z}), \ldots, \sin(\boldsymbol{\eta}_{\frac{r}{2}}^\top \mathbf{z})]^\top$$

Mercer expansion

$$k_\theta(\mathbf{x}, \mathbf{x}') = \sum_{t=1}^{\infty} \lambda_t e_t(\mathbf{x}) e_t(\mathbf{x}'),$$

Eigenvalues and eigenfunctions

$$K(k_\theta, X) \equiv \sum_{t=1}^{\infty} \lambda_t \boldsymbol{\omega}_t \boldsymbol{\omega}_t^\top,$$

$$\boldsymbol{\omega}_t = (e_t(\mathbf{x}_1), e_t(\mathbf{x}_2), \ldots, e_t(\mathbf{x}_N))$$

$$k_{\sigma_f^2, \Delta}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\mathbf{n} \in \mathbb{N}^D} \lambda_{\mathbf{n}} e_{\mathbf{n}}(\mathbf{x}_i) e_{\mathbf{n}}(\mathbf{x}_j)$$

Tensor product

Guarantees

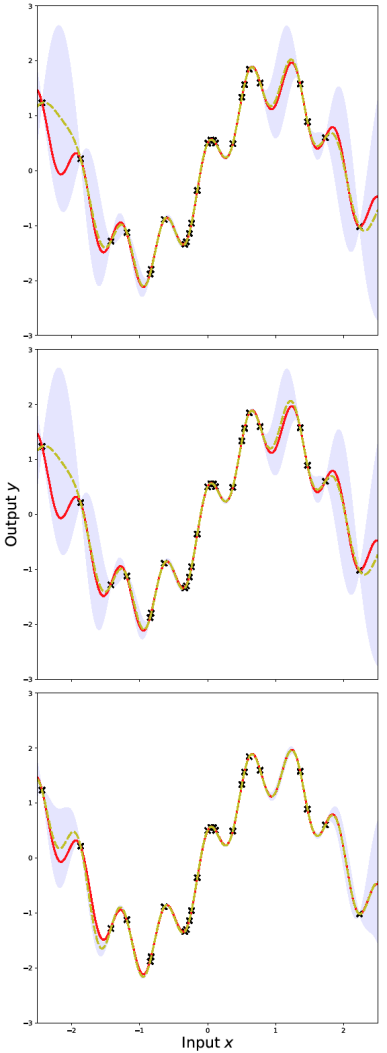
We provide bounds for the Kullback–Leibler divergence between the idealized $K(k_\theta, X)$ and low-rank approximations based on random Fourier features and Mercer expansions that can become smaller than a desired εN for moderate values of the rank of $K(k_\theta, X)$.

$$f(x) = \frac{1}{2}\left(3\sin(2x) + \cos(10x) + \frac{x}{4}\right)$$

Exact GP

Mercer GP, r=34

Random Fourier GP, r=64

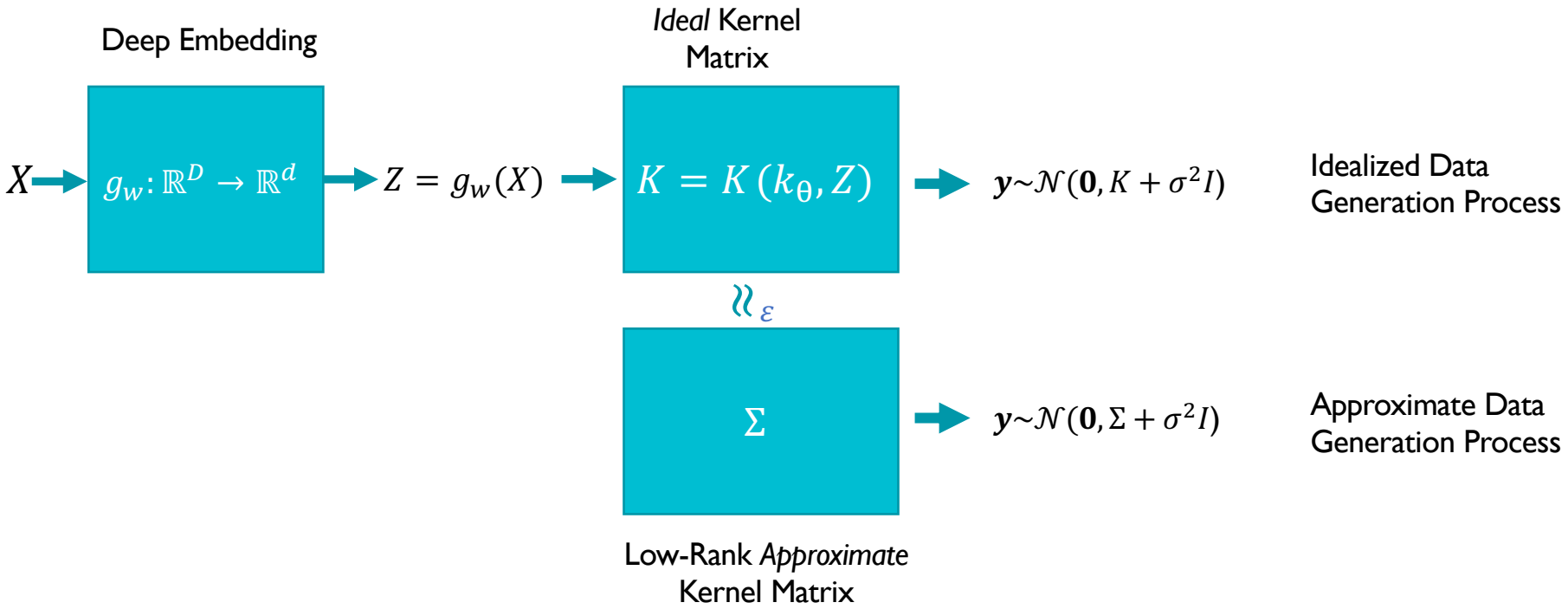


DNN for feature extraction

- $g_w : \mathbf{x} \mapsto \mathbf{z}$ embeds a feature vector $\mathbf{x}$ to a feature vector $\mathbf{z} \in \mathbb{R}^d$
- $\mathbf{y} \sim \mathcal{N}(0, K(k_\theta, Z) + \sigma^2 I_N), Z = (g_w(\mathbf{x}_i) \equiv \mathbf{z}_i)_{i=1}^N$
- Identify a feature map $\phi_{\theta,\varepsilon} : \mathbb{R}^d \to \mathbb{R}^r$, providing a guarantee of the form

$$K(k_\theta, Z) \approx_\varepsilon \Sigma(\phi_{\theta,\varepsilon}, Z) = (\phi_{\theta,\varepsilon}(\mathbf{z}_i)^\top \phi_{\theta,\varepsilon}(\mathbf{z}_j))_{ij}$$

DNN for feature extraction (end-to-end differentiable)



Methods

- Deep Mercer GP with d=1, r=15
- Deep Fourier GP with d=4, r=40
- Stochastic variational inference GP with 250 and 500 inducing points Hensman *etal. (2013)*
- Sparse GP regression with 250 and 500 inducing points Titsias (2009)
- Deep kernel learning with 5000 and 10000 inducing points Wilson *etal.* (2016)
- Deep GP with random Fourier features Cutajar *etal.* (2017)

Data sets

	N	N^*	D
ELEVATORS	14,939	1,660	18
PROTEIN	41,157	4,573	9
SARCOS	44,039	4,894	21
3DROAD	391,386	43,488	3
SONG	463,810	51,535	90
BUZZ	524,925	58,325	77
ELECTRIC	1,844,352	204,928	19

	Elevators	Protein	Sarcos	3DRoad	Song	Buzz	Electric
	Negative Log-Predictive Density						
N	14939	41157	44039	391386	463810	524925	1844352
N^*	1660	4573	4894	43488	51535	58325	204928
D	18	9	21	3	90	77	19
SVIGP	0.444(0.021)	1.041(0.007)	−0.422(0.006)	0.652(0.008)	1.208(0.005)	0.087(0.006)	0.804(0.003)
SVIGP+	0.435(0.018)	0.991(0.006)	−0.479(0.004)	0.541(0.008)	1.205(0.005)	0.078(0.005)	0.769(0.002)
SGPR	0.433(0.017)	0.997(0.007)	−0.370(0.007)	0.799(0.007)	1.202(0.006)	0.216(0.005)	0.871(0.002)
SGPR+	0.420(0.017)	0.944(0.005)	−0.468(0.009)	0.737(0.011)	1.198(0.006)	0.186(0.004)	0.810(0.001)
DKL	0.527(0.011)	0.958(0.020)	0.395(0.040)	0.744(0.129)	1.261(0.057)	0.460(0.003)	0.447(0.013)
DKL+	0.536(0.011)	0.961(0.037)	0.430(0.034)	0.687(0.047)	1.315(0.158)	0.438(0.017)	0.448(0.012)
RFEDGP	0.434(0.021)	1.028(0.006)	−0.303(0.061)	0.583(0.009)	1.207(0.006)	0.238(0.032)	0.616(0.004)
DMGP	0.371(0.036)	0.857(0.015)	**−0.777**(0.015)	0.140(0.010)	**1.185**(0.004)	−0.008(0.022)	0.078(0.002)
DFGP	**0.350**(0.029)	**0.853**(0.018)	**−0.777**(0.020)	**0.139**(0.012)	1.189(0.005)	**−0.016**(0.002)	**0.067**(0.004)
DNN+S	0.402(0.030)	0.904(0.013)	−0.559(0.021)	0.239(0.020)	1.211(0.001)	0.019(0.003)	0.165(0.001)
DNN+M	0.401(0.030)	0.893(0.016)	−0.585(0.029)	0.233(0.020)	1.208(0.001)	0.025(0.016)	0.164(0.001)
DNN+F	0.380(0.022)	0.895(0.022)	−0.628(0.044)	0.237(0.008)	1.210(0.002)	0.012(0.001)	0.155(0.001)
	Training Time (seconds)						
SVIGP	59(2)	182(24)	269(19)	2096(297)	2527(19)	2615(165)	8231(302)
SVIGP+	150(5)	425(3)	455(2)	3895(92)	4845(132)	5715(102)	18878(1513)
SGPR	**49**(1)	156(15)	227(11)	1697(53)	2012(13)	2114(109)	11190(68)
SGPR+	144(3)	381(11)	419(7)	3661(124)	4676(118)	5208(336)	30357(573)
DKL	285(27)	435(4)	455(5)	2531(31)	2916(180)	2854(408)	14455(596)
DKL+	774(80)	1317(227)	740(402)	2377(194)	2885(161)	3182(245)	14833(1608)
RFEDGP	184(9)	559(43)	629(49)	2862(296)	4627(66)	4276(232)	26256(1647)
DMGP	121(26)	375(23)	448(26)	3602(238)	3598(95)	3963(63)	14311(134)
DFGP	51(2)	**137**(13)	**146**(1)	**1363**(8)	**1898**(15)	**2092**(26)	**6785**(323)
DNN+S	28(2)	80(7)	78(4)	752(57)	224(8)	513(7)	2211(614)
DNN+M	41(4)	113(10)	113(7)	1061(79)	331(13)	711(12)	3069(841)
DNN+F	53(3)	150(17)	171(13)	1326(93)	245(13)	1890(133)	6504(294)

	RMSE						
	ELEVATORS	PROTEIN	SARCOS	3DROAD	SONG	BUZZ	ELECTRIC
N	14939	41157	44039	391386	463810	524925	1844352
N^*	1660	4573	4894	43488	51535	58325	204928
D	18	9	21	3	90	77	19
SVIGP	0.379(0.009)	0.683(0.005)	0.160(0.001)	0.462(0.004)	0.810(0.005)	0.271(0.003)	0.540(0.002)
SVIGP+	0.375(0.007)	0.649(0.005)	0.151(0.001)	0.413(0.004)	0.807(0.004)	0.270(0.003)	0.521(0.001)
SGPR	0.375(0.007)	0.653(0.005)	0.168(0.002)	0.537(0.004)	0.806(0.005)	0.315(0.003)	0.577(0.001)
SGPR+	0.370(0.007)	0.620(0.004)	0.153(0.002)	0.506(0.006)	0.802(0.005)	0.308(0.003)	0.542(0.001)
DKL	0.352(0.010)	0.630(0.012)	0.230(0.047)	0.499(0.074)	0.815(0.006)	0.274(0.014)	0.285(0.008)
DKL+	0.361(0.009)	0.632(0.022)	0.276(0.035)	0.474(0.024)	0.813(0.004)	0.268(0.014)	0.296(0.015)
RFEDGP	0.355(0.013)	0.678(0.004)	0.179(0.012)	0.434(0.004)	0.809(0.005)	0.307(0.009)	0.448(0.002)
DMGP	0.346(0.010)	0.564(0.007)	**0.111**(0.002)	**0.277**(0.003)	**0.791**(0.003)	**0.237**(0.000)	0.261(0.001)
DFGP	**0.341**(0.008)	**0.562**(0.008)	**0.111**(0.002)	0.278(0.003)	0.795(0.004)	0.238(0.000)	**0.259**(0.001)
DNN+S	0.359(0.007)	0.588(0.006)	0.144(0.001)	0.311(0.005)	0.806(0.001)	0.251(0.000)	0.288(0.001)
DNN+M	0.359(0.007)	0.581(0.006)	0.140(0.003)	0.310(0.005)	0.804(0.001)	0.250(0.001)	0.287(0.001)
DNN+F	0.354(0.005)	0.582(0.009)	0.135(0.004)	0.311(0.001)	0.805(0.002)	0.250(0.001)	0.285(0.001)

	Protein			Sarcos		
	Negative log-predictive density-DMGP					
$\sqrt[d]{r}$	$d=1$	$d=2$	$d=3$	$d=1$	$d=2$	$d=3$
2	0.883(0.014)	0.872(0.012)	0.872(0.015)	−0.778(0.012)	−0.762(0.019)	−0.754(0.029)
4	0.856(0.015)	0.863(0.014)	0.867(0.025)	−0.777(0.015)	−0.775(0.019)	−0.780(0.016)
8	0.857(0.015)	0.855(0.013)	0.848(0.014)	−0.778(0.015)	−0.773(0.021)	−0.780(0.019)
10	0.857(0.015)	0.855(0.013)	0.848(0.014)	−0.777(0.015)	−0.772(0.021)	−0.780(0.020)
16	0.857(0.015)	0.855(0.013)	0.848(0.015)	−0.778(0.015)	−0.772(0.021)	−0.770(0.020)
32	0.857(0.015)	0.855(0.013)	–	−0.777(0.015)	−0.772(0.021)	–
$\frac{r}{2}$	Negative log-predictive density-DFGP					
2	0.871(0.013)	0.873(0.014)	0.862(0.013)	−0.608(0.130)	−0.697(0.069)	−0.771(0.019)
4	0.856(0.013)	0.851(0.012)	0.847(0.014)	−0.784(0.014)	−0.778(0.021)	−0.783(0.019)
8	0.856(0.014)	0.854(0.012)	0.846(0.013)	−0.784(0.014)	−0.779(0.021)	−0.784(0.020)
10	0.856(0.014)	0.855(0.013)	0.846(0.014)	−0.784(0.014)	−0.779(0.020)	−0.786(0.020)
16	0.856(0.014)	0.854(0.012)	0.846(0.015)	−0.784(0.014)	−0.779(0.021)	−0.784(0.022)
32	0.856(0.014)	0.853(0.012)	0.847(0.015)	−0.785(0.014)	−0.781(0.021)	−0.785(0.019)
$\sqrt[d]{r}$	Training Time-DMGP					
2	115(4)	132(2)	154(2)	127(3)	144(3)	174(3)
4	112(1)	170(3)	374(12)	127(6)	187(2)	417(21)
8	114(1)	308(9)	874(26)	124(6)	325(11)	980(18)
10	116(5)	369(11)	2147(63)	128(5)	401(15)	2325(84)
16	117(1)	404(12)	88649(163)	130(4)	456(16)	94965(293)
32	122(1)	1864(21)	–	135(6)	2071(72)	–
$\frac{r}{2}$	Training Time-DFGP					
2	108(1)	108(1)	108(1)	124(3)	121(3)	124(3)
4	109(1)	111(2)	110(2)	123(7)	130(2)	126(4)
8	112(1)	112(1)	113(2)	125(2)	132(1)	133(2)
10	115(4)	113(6)	126(3)	127(4)	134(5)	136(3)
16	118(1)	118(2)	156(17)	129(8)	136(4)	188(9)
32	126(1)	126(2)	176(20)	141(5)	145(4)	199(5)

Elevators							
DMGP				**DFGP**			
$\sqrt[d]{r}$	$d=1$	$d=2$	$d=3$	$\frac{r}{2}$	$d=1$	$d=2$	$d=3$
NLPD							
2	0.381(0.037)	0.361(0.032)	0.377(0.044)	2	0.411(0.037)	0.381(0.040)	0.367(0.029)
4	0.371(0.036)	0.351(0.032)	0.353(0.028)	4	0.380(0.037)	0.357(0.032)	0.357(0.030)
8	0.371(0.036)	0.351(0.032)	0.352(0.029)	8	0.379(0.036)	0.356(0.032)	0.356(0.030)
10	0.371(0.037)	0.351(0.032)	0.352(0.029)	10	0.379(0.036)	0.357(0.031)	0.357(0.029)
16	0.371(0.036)	0.351(0.032)	0.352(0.029)	16	0.379(0.037)	0.357(0.032)	0.357(0.030)
32	0.371(0.036)	0.351(0.032)	–	32	0.379(0.036)	0.356(0.032)	0.357(0.030)
Training Time							
2	40(1)	49(1)	59(1)	2	39(1)	38(1)	40(0)
4	41(1)	58(1)	135(2)	4	39(1)	38(0)	40(0)
8	41(2)	100(1)	303(5)	8	40(1)	39(1)	41(0)
10	41(2)	117(2)	710(19)	10	41(2)	40(1)	42(1)
16	42(1)	140(3)	31593(151)	16	42(2)	41(0)	51(3)
32	44(2)	620(10)	–	32	46(2)	44(1)	55(3)

Spotlight of future research

- Classification / multilabel / multiclass
- Point processes
- Reinforcement learning