Highlights from NeurIPS 2019

Vancouver, Canada

Function-Space Distributions over Kernels

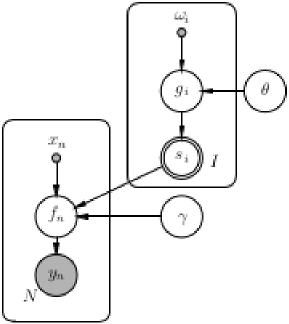
Gregory W. Benton[*1] Wesley J. Maddox[*2] Jayson P. Salkey[*1]
Júlio Albinati[‡3] Andrew Gordon Wilson[1,2]

[1]Courant Institute of Mathematical Sciences, New York University
[2]Center for Data Science, New York University
[3]Microsoft

[1]

[1]Gregory Benton et al. “Function-Space Distributions over Kernels”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14939–14950.

Spectral Representation



Bochner's Theorem: Can represent a (positive definite) kernel as

$$k(\tau) = \int_{\mathbb{R}} e^{2\pi i \omega \tau} S(\omega) d\omega,$$

with $\tau = |x - x'|$, the distance between any two inputs, and some positive finite function $S(\omega)$. Only need to learn the (unnormalized) spectral density $S(\omega)$ to learn $k(\tau)$.

Model Specification

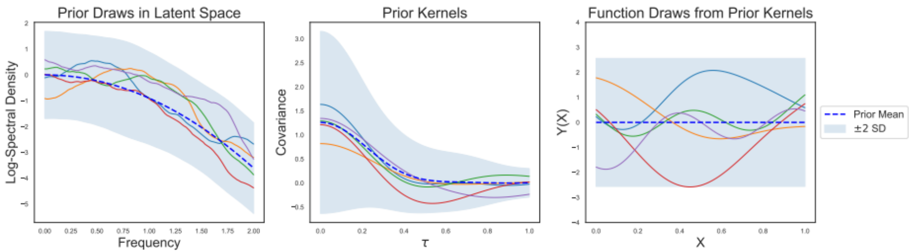
{Hyperprior} $\quad p(\boldsymbol{\phi}) = p(\theta, \gamma)$

{Latent GP} $\quad g(\omega)|\theta \sim \mathcal{GP}\left(\mu(\omega;\theta), k_g(\omega, \omega';\theta)\right)$

{Spectral Density} $\quad S(\omega) = \exp\{g(\omega)\}$

{Data GP} $\quad f(x_n)|S(\omega), \gamma \sim \mathcal{GP}(\gamma_0, k(\tau; S(\omega))).$

Forward sampling



[2]

[2]Gregory Benton et al. “Function-Space Distributions over Kernels”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14939–14950.

Inference & Prediction

Updating Hyper-Parameters: Considering the model specification in Eq. 4, we can define a loss as a function of $\phi = \{\theta, \gamma\}$ for an observation of the density, $\tilde{g}(\omega)$, and data observations $y(x)$. This loss corresponds to the entropy, marginal log-likelihood of the latent GP with fixed data GP, and the marginal log-likelihood of the data GP.

$$\mathcal{L}(\phi) = -\left(\log p(\phi) + \log p(\tilde{g}(\omega)|\theta, \omega) + \log p(y(x)|\tilde{g}(\omega), \gamma, x)\right). \tag{7}$$

Updating Latent Gaussian Process: With fixed hyper-parameters ϕ, the posterior of the latent GP is

$$p(g(\omega)|\phi, x, y(x), f(x)) \propto \mathcal{N}(\mu(\omega; \theta), k_g(\omega; \theta))p(f(x)|g(\omega), \gamma). \tag{9}$$

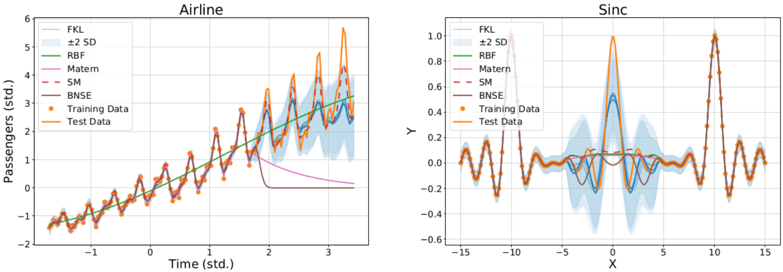
Prediction: The predictive distribution for any test input x^* is given by

$$p(f^*|x^*, x, y, \phi) = \int p(f^*|x^*, x, y, \phi, k)p(k|x^*, x, y, \phi)dk$$

$$p(f^*|x^*, x, y, \phi) \approx \frac{1}{J}\sum_{j=1}^{J} p(f^*|x^*, x, y, \phi, k_j), \quad k_j \sim p(k|x^*, x, y, \phi).$$ [3]

[3] Gregory Benton et al. “Function-Space Distributions over Kernels”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14939–14950.

Extrapolation and Interpolation



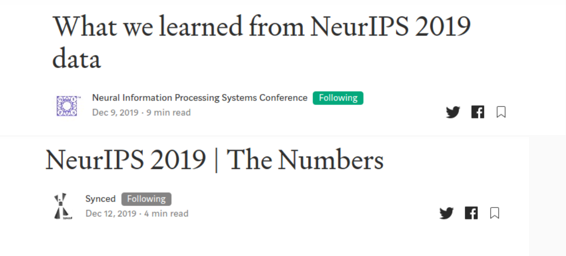
[4]

[4] Gregory Benton et al. “Function-Space Distributions over Kernels”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14939–14950.

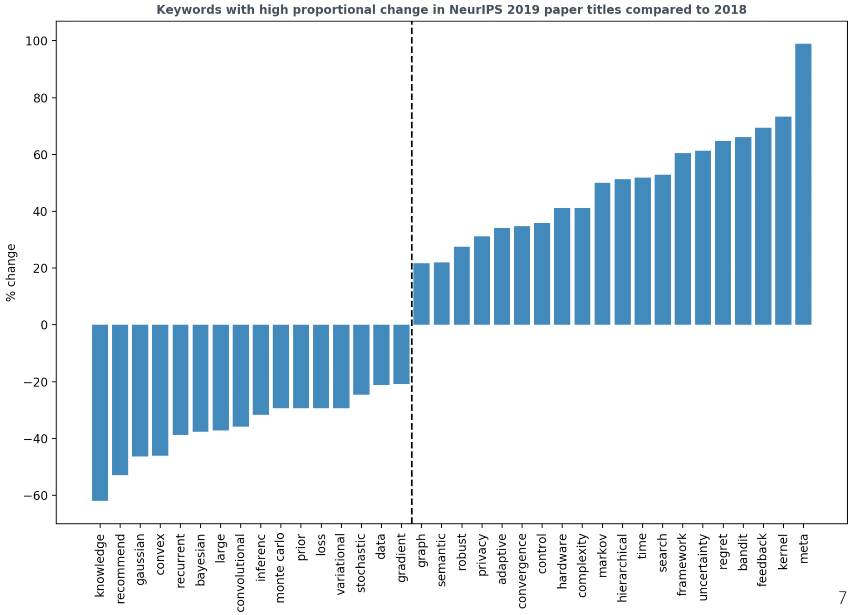
1. Matthew D Hoffman and Yian Ma. “Langevin Dynamics as Nonparametric Variational Inference”. In: *2nd Symposium on Advances in Approximate Bayesian Inference* (2019) ”Best Industry paper”

2. Yoshua Bengio Keynote on **”Going from System 1 to System 2 Deep learning”**

3. NeurIPS Summaries (on Medium):



Trivia



Keywords with high proportional change in NeurIPS 2019 paper titles compared to 2018
% change
100
80
60
40
20
0
−20
−40
−60
knowledge
recommend
gaussian
convex
recurrent
bayesian
large
convolutional
inferenc
monte carlo
prior
loss
variational
stochastic
data
gradient
graph
semantic
robust
privacy
adaptive
convergence
control
hardware
complexity
markov
hierarchical
time
search
framework
uncertainty
regret
bandit
feedback
kernel
meta