Deep Gaussian Processes

Maurizio Filippone

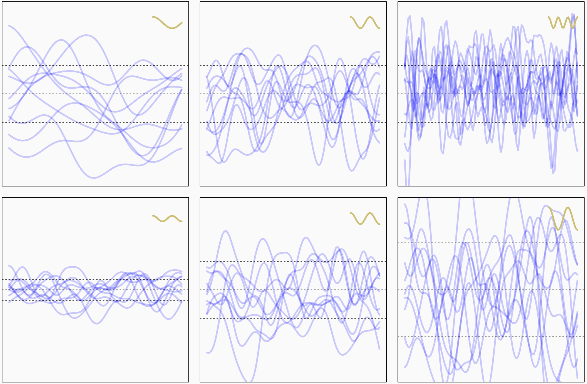
EURECOM, Sophia Antipolis, France

August 31st, 2018

Introduction

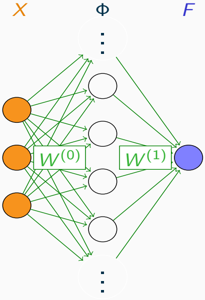
Gaussian Processes - Priors over Functions

- Infinite Gaussian random variables with parametric and input-dependent covariance



Rasmussen and Williams, 2006

Gaussian Processes as Infinitely-Wide Shallow Neural Nets



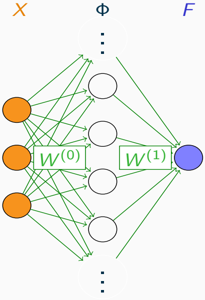
- Take $W^{(i)} \sim \mathcal{N}(\mathbf{0}, \alpha_i I)$
- Central Limit Theorem implies that F is Gaussian

- F has zero-mean
- $\mathrm{cov}(F) = \mathrm{E}_{p(W^{(0)}, W^{(1)})}[\Phi(XW^{(0)})W^{(1)}W^{(1)\top}\Phi(XW^{(0)})^\top]$

Neal, *LNS*, 1996

Gaussian Processes as Infinitely-Wide Shallow Neural Nets

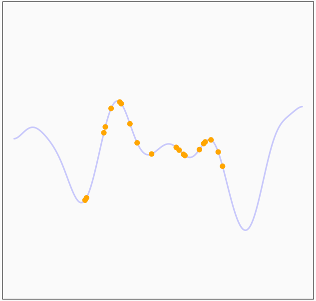
- Take $W^{(i)} \sim \mathcal{N}(\mathbf{0}, \alpha_i I)$
- Central Limit Theorem implies that F is Gaussian



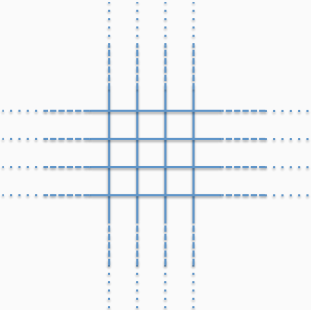
- F has zero-mean
- $\mathrm{cov}(F) = \alpha_1 \mathrm{E}_{p(W^{(0)})}[\Phi(XW^{(0)})\Phi(XW^{(0)})^\top]$
- Some choices of Φ lead to analytic expression of known kernels (RBF, Matérn, arc-cosine, Brownian motion, ...)

Neal, *LNS*, 1996

Gaussian Processes - Priors over Functions

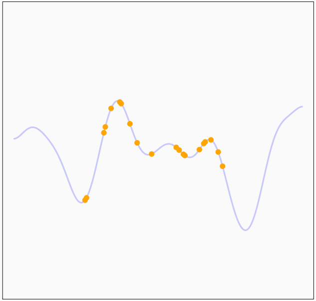


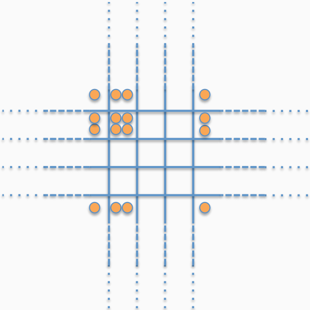
$K =$



Rasmussen and Williams, 2006

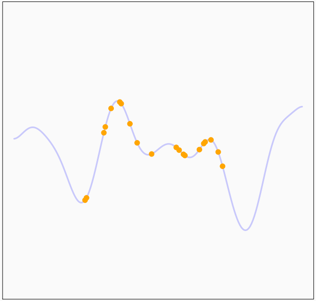
Gaussian Processes - Priors over Functions

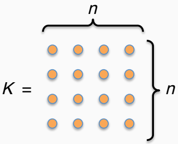


$K =$ 

Rasmussen and Williams, 2006

Gaussian Processes - Priors over Functions

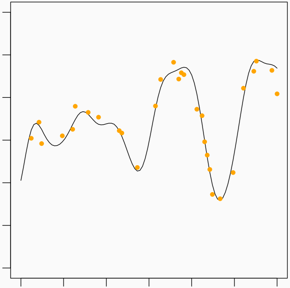




Rasmussen and Williams, 2006

Gaussian Processes - Regression example

- Inputs $= X$ Labels $= Y$
- Introduce latent variables F with covariance $K = K(X, \theta)$
- Introduce Gaussian likelihood $p(Y|F)$



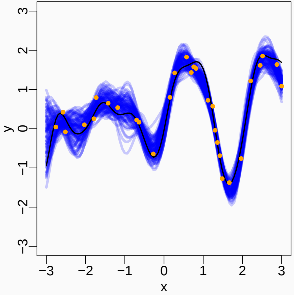
- Posterior $p(F|Y, X, \theta) \propto \frac{p(Y|F)p(F|X,\theta)}{\int p(Y|F)p(F|X,\theta)dF}$

Rasmussen and Williams, 2006

Gaussian Processes - Regression example

- Predictive distribution

$$p(F_*|Y, X, \theta) = \int p(F_*|F, \theta)p(F|Y, X, \theta)dF$$

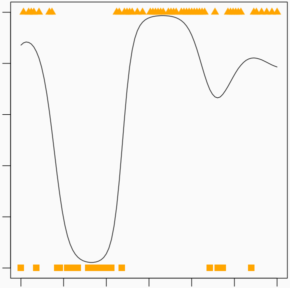


- Posterior $p(F|Y, X, \theta) \propto \dfrac{p(Y|F)p(F|X, \theta)}{\int p(Y|F)p(F|X, \theta)dF}$

Rasmussen and Williams, 2006

Gaussian Processes - Classification example

- Inputs $= X$ Labels $= Y$
- Introduce latent variables F with covariance $K = K(X, \theta)$
- Introduce Bernoulli likelihood $p(Y|F)$



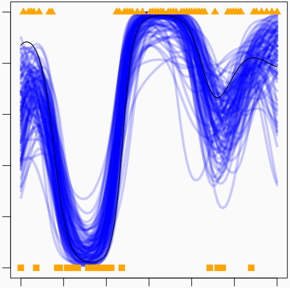
- Posterior $p(F|Y, X, \theta) \propto \frac{p(Y|F)p(F|X,\theta)}{\int p(Y|F)p(F|X,\theta)dF}$

Rasmussen and Williams, 2006

Gaussian Processes - Classification example

- Predictive distribution - needs approximation to $p(F|Y, X, \theta)$!

$$p(F_*|Y, X, \theta) = \int p(F_*|F, \theta) p(F|Y, X, \theta) dF$$

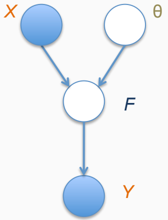


- Posterior $p(F|Y, X, \theta) \propto \dfrac{p(Y|F)p(F|X, \theta)}{\int p(Y|F)p(F|X, \theta)dF}$

Rasmussen and Williams, 2006

Challenges and Limitations

- Kernel design
- $p(Y|X,\theta)$ might be expensive to compute (factorize K)
- $p(Y|X,\theta)$ might not even be computable!

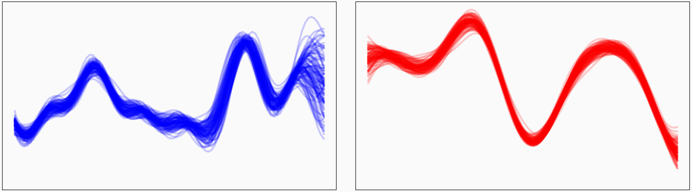


- Marginal likelihood

$$p(Y|X,\theta) = \int p(Y|F)p(F|X,\theta)dF$$

Deep Gaussian Processes for Large Representational Power

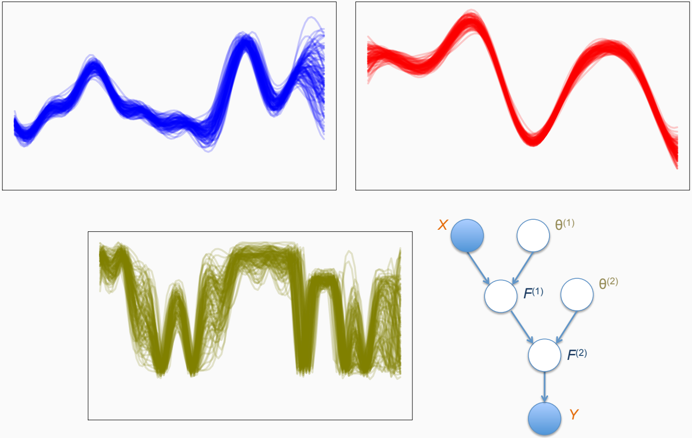
- Bypassing kernel design through **composition** of processes



$$(f \circ g)(x)??$$

Deep Gaussian Processes for Large Representational Power

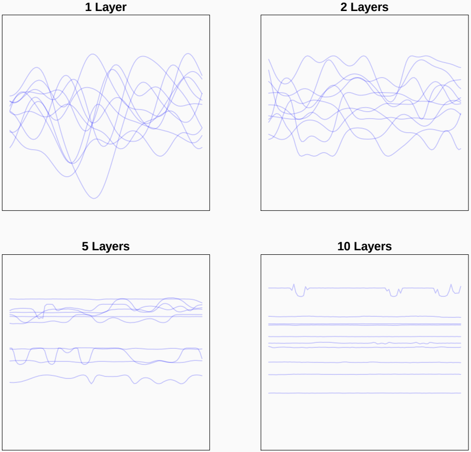
- Composition of stationary processes yields something very complex

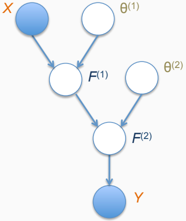


Neal, *LNS*, 1996 – Damianou and Lawrence, *AISTATS*, 2013 – Duvenaud et al., *AISTATS*, 2014

Pathologies of Deep Gaussian Processes

- Deep is not necessarily good!
- Example

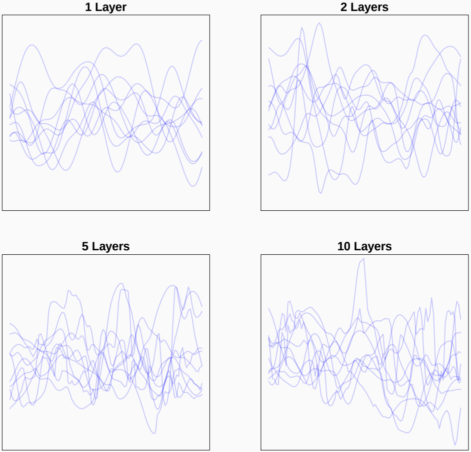


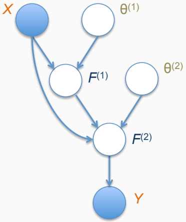


Neal, *LNS*, 1996 – Duvenaud et al., *AISTATS*, 2014 – Matthews et al., *arXiv*, 2018

Pathologies of Deep Gaussian Processes

- Deep is not necessarily good!
- Feeding input to each layer helps...





Neal, *LNS*, 1996 – Duvenaud et al., *AISTATS*, 2014 – Matthews et al., *arXiv*, 2018

Learning Deep Gaussian Processes

- Inference requires calculating integrals of this kind:

$$
\begin{aligned}
p(Y|X,\theta) \quad = \quad & \int p\left(Y|F^{(N_\mathrm{h})},\theta^{(N_\mathrm{h})}\right) \times \\
& p\left(F^{(N_\mathrm{h})}|F^{(N_\mathrm{h}-1)},\theta^{(N_\mathrm{h}-1)}\right) \times \ldots \times \\
& p\left(F^{(1)}|X,\theta^{(0)}\right) dF^{(N_\mathrm{h})} \ldots dF^{(1)}
\end{aligned}
$$

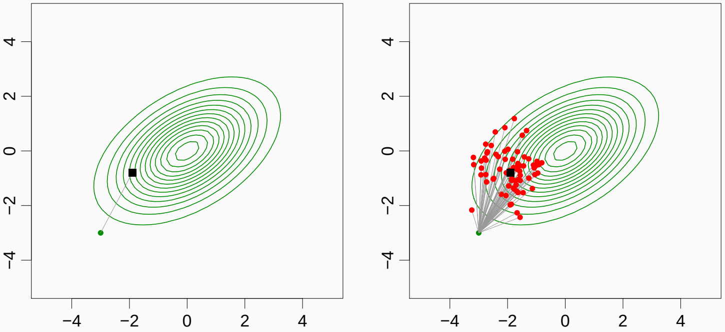
- Extremely challenging!

The Deep Learning Revolution

- Large representational power
- Mini-batch-based learning
- Exploit GPU and distributed computing
- Automatic differentiation
- Mature development of regularization (e.g., dropout)
- Application-specific representations (e.g., convolutional)

Stochastic Gradient Optimization

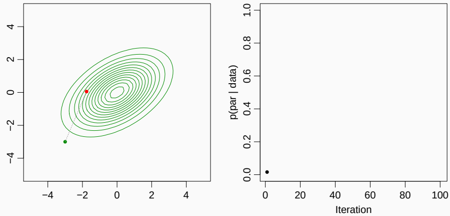
$$\mathrm{E}\left\{\widetilde{\nabla_{\text{par}}}\text{LowerBound}\right\} = \nabla_{\text{par}}\text{LowerBound}$$

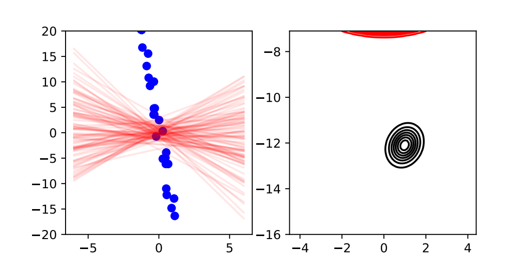


Robbins and Monro, *AoMS*, 1951

Stochastic Variational Inference - Illustration

$$\text{vpar}' = \text{vpar} + \frac{\alpha_t}{2}\widetilde{\nabla_{\text{vpar}}}(\text{LowerBound}) \qquad \alpha_t \to 0$$





Is There Any Hope for GPs and DGPs?

- Mini-batch training is straightforward when objective factorizes over training points

$$\text{objective} = \sum_{i} f(\mathbf{y}_i, \mathbf{x}_i, \text{par})$$

Is There Any Hope for GPs and DGPs?

- Mini-batch training is straightforward when objective factorizes over training points

$$\text{objective} = \sum_i f(\mathbf{y}_i, \mathbf{x}_i, \text{par})$$

- In GPs latent variables are fully correlated

$$p(F|X, \boldsymbol{\theta}) = \mathcal{N}(F|\mathbf{0}, K(X, \boldsymbol{\theta})) \propto \exp\left(-\frac{1}{2} F^\top K^{-1} F\right)$$

- Naïve mini-batch approaches would totally break this!

Can we exploit what made Deep Learning successful for practical and scalable learning of (Deep) Gaussian processes?

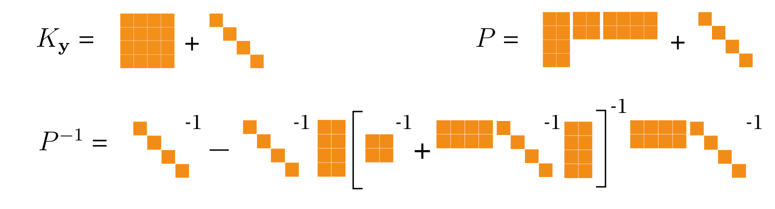
Inference for Deep Gaussian Processes

Inference for DGPs

- Inducing points-based approximations
  - VI+Titsias *AISTATS* 2009 Sparse GP
    - Damianou and Lawrence, *AISTATS*, 2013
    - Hensman and Lawrence, *arXiv*, 2014
    - Salimbeni and Deisenroth, *NIPS*, 2017
  - EP+FITC - Bui et al. *ICML*, 2016
  - MCMC+Titsias *AISTATS* 2009 Sparse GP
    - Havasi et al., *arXiv*, 2018
- Random feature-based approximations
  - Gal and Ghahramani, ICML 2016
  - Cutajar et al., ICML 2017

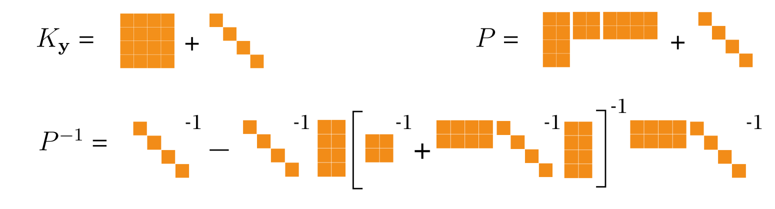
Inference for DGPs

- Low-Rank Approximation options - $\mathcal{O}(nm^2)$
- Call P as a low rank approximation to $\mathbf{K_y}$
- Woodbury identity exploits low rank structure of P



Inference for DGPs

- Low-Rank Approximation options - $\mathcal{O}(nm^2)$
- Call P as a low rank approximation to $\mathbf{K_y}$
- Woodbury identity exploits low rank structure of P

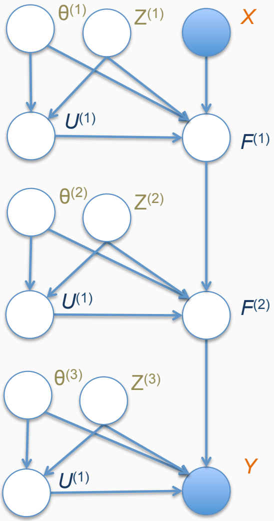


DGPs: Low-rank approximation of covariance at each layer

Scalable Expectation Propagation for DGPs

- Pseudo-inputs $Z^{(i)}$
- Inducing variables $U^{(i)}$
- VI targets

$$q\left(U^{(i)}\right)$$



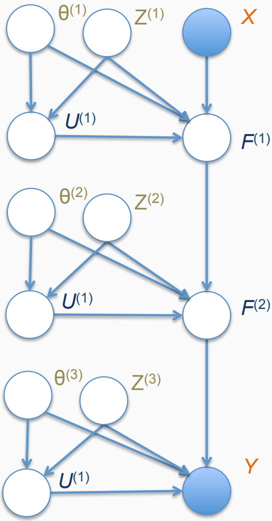
Bui et al., *ICML*, 2016

Scalable Expectation Propagation for DGPs

- Pseudo-inputs $Z^{(i)}$
- Inducing variables $U^{(i)}$
- VI targets

$$q\left(U^{(i)}\right)$$

- Assuming $q\left(U^{(i)}\right) \propto p\left(U^{(i)}\right) g\left(U^{(i)}\right)^N$ learn g as an average data factor
- Reduces memory and allows for factorization of the objective (output of each layer made Gaussian)

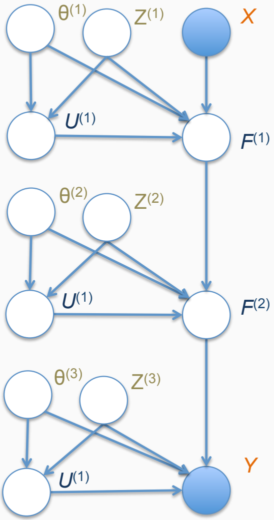


Bui et al., *ICML*, 2016

Inducing Points for DGPs extending Titsias, AISTATS, 2009

- Pseudo-inputs $Z^{(i)}$
- Inducing variables $U^{(i)}$
- VI targets $q\left(F^{(i)}, U^{(i)} | F^{(i-1)}\right)$

$$p\left(F^{(i)} | U^{(i)}, F^{(i-1)}\right) q\left(U^{(i)}\right)$$



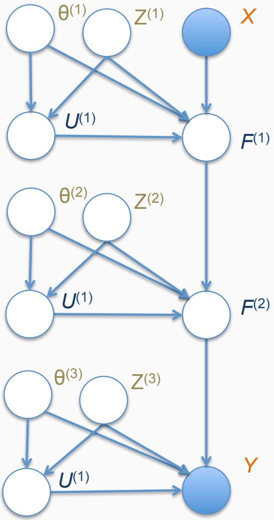
Hensman and Lawrence, *arXiv*, 2014 – Salimbeni and Deisenroth, *NIPS*, 2017

Inducing Points for DGPs extending Titsias, AISTATS, 2009

- Pseudo-inputs $Z^{(i)}$
- Inducing variables $U^{(i)}$
- VI targets $q\left(F^{(i)}, U^{(i)} | F^{(i-1)}\right)$

$$p\left(F^{(i)} | U^{(i)}, F^{(i-1)}\right) q\left(U^{(i)}\right)$$

- Lower bound factorizes across training points...
- ... and the ith marginal of the final layer depends only on the ith marginals of all layers



Hensman and Lawrence, *arXiv*, 2014 – Salimbeni and Deisenroth, *NIPS*, 2017

Random Feature Expansions for DGPs - Bochner's theorem

- Continuous shift-invariant covariance function

$$k(\mathbf{x}_i - \mathbf{x}_j | \boldsymbol{\theta}) = \sigma^2 \int p(\boldsymbol{\omega} | \boldsymbol{\theta}) \exp\left(\iota(\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\omega}\right) d\boldsymbol{\omega}$$

Random Feature Expansions for DGPs - Bochner's theorem

- Continuous shift-invariant covariance function

$$k(\mathbf{x}_i - \mathbf{x}_j | \boldsymbol{\theta}) = \sigma^2 \int p(\boldsymbol{\omega}|\boldsymbol{\theta}) \exp\left(\iota(\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\omega}\right) d\boldsymbol{\omega}$$

- Monte Carlo estimate

$$k(\mathbf{x}_i - \mathbf{x}_j | \boldsymbol{\theta}) \approx \frac{\sigma^2}{N_{\mathrm{RF}}} \sum_{r=1}^{N_{\mathrm{RF}}} \mathbf{z}(\mathbf{x}_i | \tilde{\boldsymbol{\omega}}_r)^\top \mathbf{z}(\mathbf{x}_j | \tilde{\boldsymbol{\omega}}_r)$$

with

$$\tilde{\boldsymbol{\omega}}_r \sim p(\boldsymbol{\omega}|\boldsymbol{\theta})$$

$$\mathbf{z}(\mathbf{x}|\boldsymbol{\omega}) = [\cos(\mathbf{x}^\top \boldsymbol{\omega}), \sin(\mathbf{x}^\top \boldsymbol{\omega})]^\top$$

Rahimi and Recht, *NIPS*, 2008 - Lázaro-Gredilla et al., *JMLR*, 2010

Random Feature Expansions for DGPs

- Define

$$\Phi^{(l)} = \sqrt{\frac{\sigma^2}{N_{\mathrm{RF}}^{(l)}}} \left[\cos\left(F^{(l)}\Omega^{(l)}\right), \sin\left(F^{(l)}\Omega^{(l)}\right)\right]$$

and

$$F^{(l+1)} = \Phi^{(l)} W^{(l)}$$

- We are stacking Bayesian linear models with

$$p\left(W_{\cdot i}^{(l)}\right) = \mathcal{N}\left(\mathbf{0}, I\right)$$

Random Feature Expansions for DGPs

- Define

$$\Phi^{(l)} = \sqrt{\frac{\sigma^2}{N_{\mathrm{RF}}^{(l)}}} \left[\cos\left(F^{(l)}\Omega^{(l)}\right), \sin\left(F^{(l)}\Omega^{(l)}\right)\right]$$

and

$$F^{(l+1)} = \Phi^{(l)} W^{(l)}$$

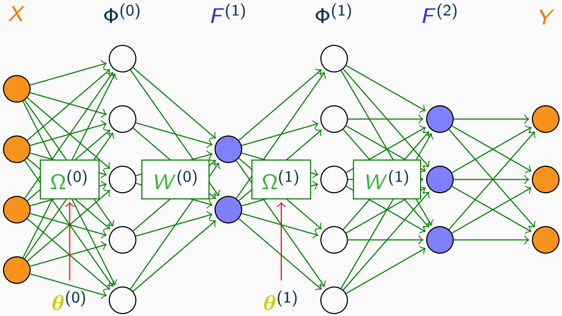
- We are stacking Bayesian linear models with

$$p\left(W_{\cdot i}^{(l)}\right) = \mathcal{N}\left(\mathbf{0}, I\right)$$

- Expansion of arc-cosine kernel yields ReLU activations!

Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017

DGPs with random features become DNNs



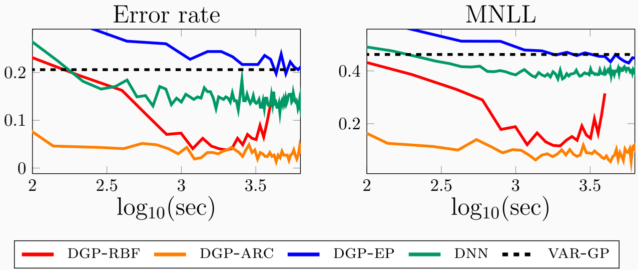
We can learn the model using Stochastic Variational Inference for Bayesian DNNs!

Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017

Results - Classification

EEG dataset

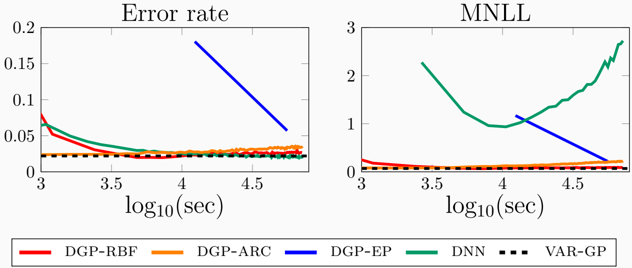
$(n = 14979,\ d = 14)$



Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017

MNIST dataset

$(n = 60000,\ d = 784)$



Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017

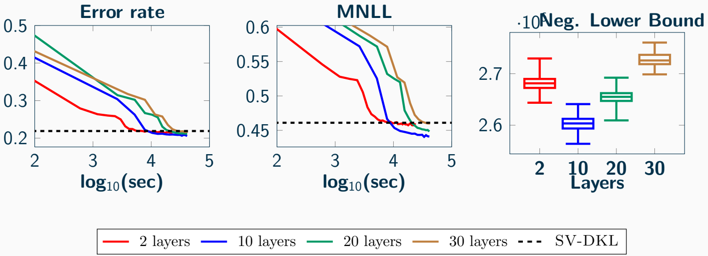
Results - MNIST-8M

- Variant of MNIST with $8.1\mathrm{M}$ images
- 99+% accuracy!
- Also, check out Krauth et al., UAI 2017

Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017 – Krauth, Cutajar, Bonilla, Filippone, *UAI*, 2017

Airline dataset

$(n = 5\mathrm{M}+,\ d = 8)$

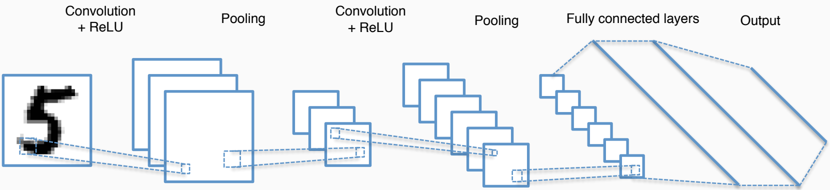


Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017

Convolutional Deep Gaussian Processes

Convolutional Nets

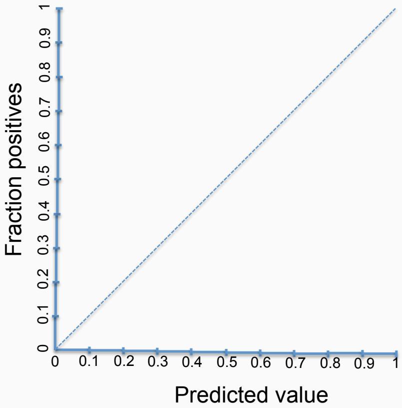
- Convolutional nets are widely used...
- ...but they are known to be overconfident!



Guo et al., *ICML*, 2017

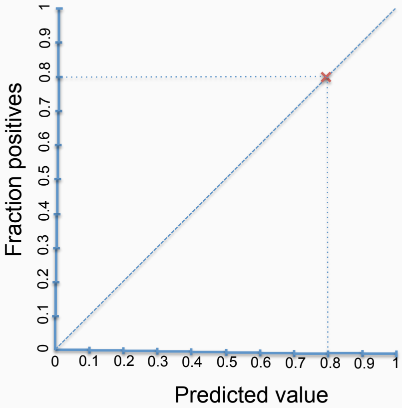
Calibration as a Measure of Quantification of Uncertainty

- Reliability diagrams



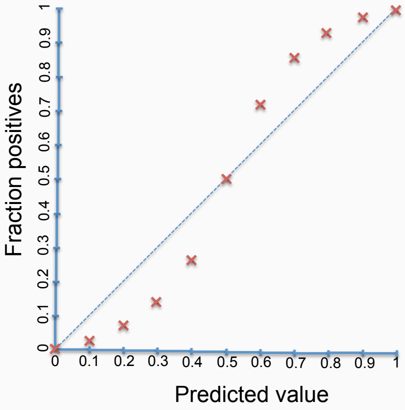
Calibration as a Measure of Quantification of Uncertainty

- Reliability diagrams



Calibration as a Measure of Quantification of Uncertainty

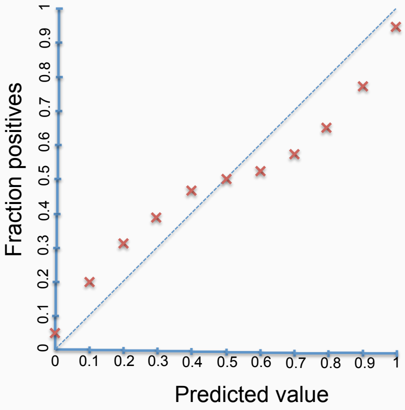
- Reliability diagrams - Under-confident predictions



- We can extract the Expected Calibration Error (ECE) score
- The BRIER score is another measure of calibration

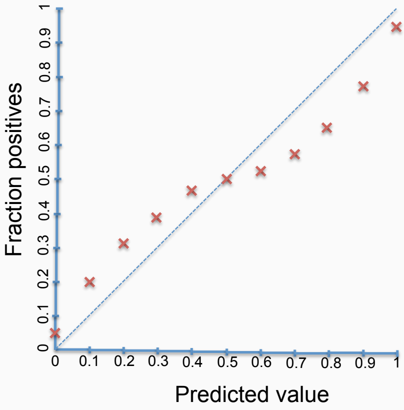
Calibration as a Measure of Quantification of Uncertainty

- Reliability diagrams - Overconfident predictions



Calibration as a Measure of Quantification of Uncertainty

- Reliability diagrams - Overconfident predictions

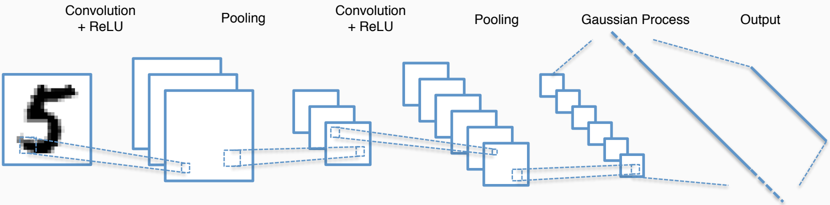


Reliability diagrams of modern Deep CNNs look like this!

Post-calibration fixes it

Combining Convolutional Nets with GPs

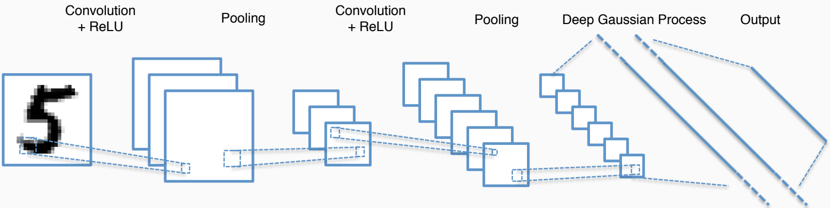
- There have been attempts to combine CNNs with GPs
- Most popular ones replace fully connected layers with GPs



Wilson et al., *NIPS*, 2016 – Bradshaw et al., *arXiv*, 2017 – Tran et al., *arXiv*, 2018

Combining Convolutional Nets with GPs

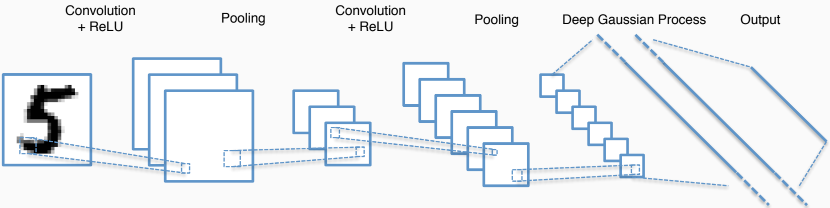
- There have been attempts to combine CNNs with GPs
- Most popular ones replace fully connected layers with GPs



- Better quantification of uncertainty??

Combining Convolutional Nets with GPs

- There have been attempts to combine CNNs with GPs
- Most popular ones replace fully connected layers with GPs

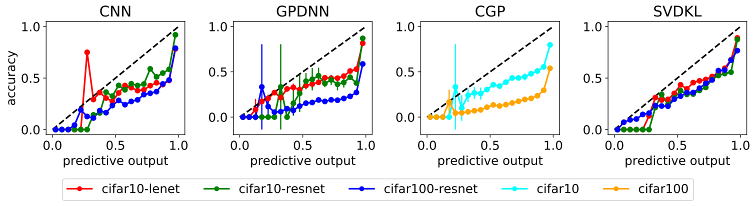


- Better quantification of uncertainty?? **NO!**

Tran et al., *arXiv*, 2018

Existing Combinations of CNNs and GPs

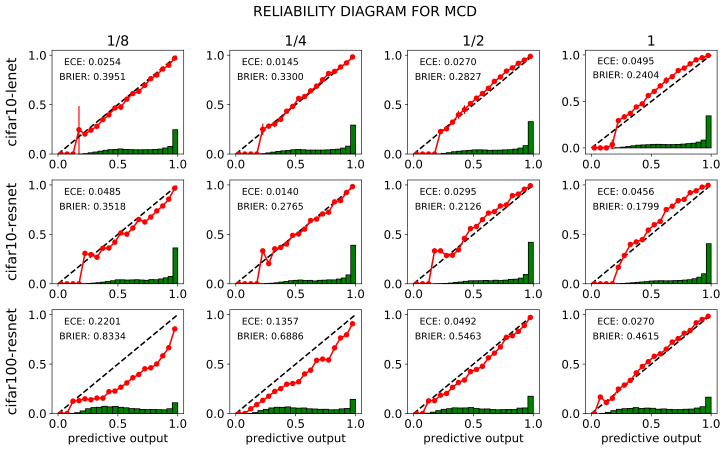
- Convolutional Neural Nets - CNN
- Hybrid GPs and DNNs - GPDNN
- Stochastic Variational Deep Kernel Learning - SVDKL
- Convolutional GP - CGP



Bradshaw et al., *arXiv*, 2017 – Wilson et al., *NIPS*, 2016 – van der Wilk et al., *NIPS*, 2017

Bayesian CNNs are calibrated

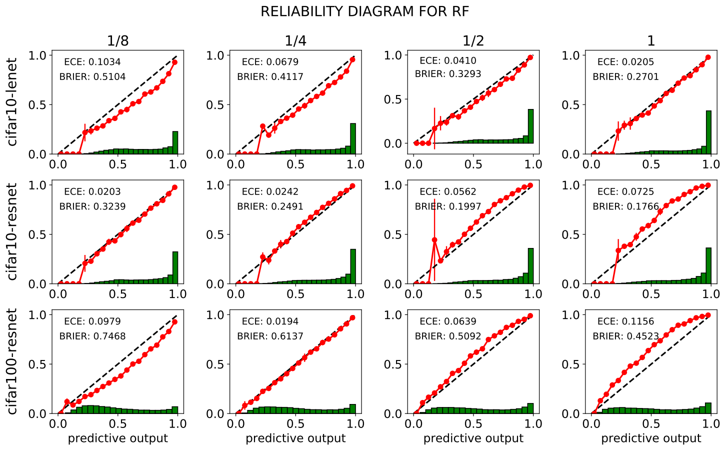
- Inferring parameters of convolutional filter recovers calibration
- Example with Monte Carlo Dropout



Tran et al., *arXiv*, 2018

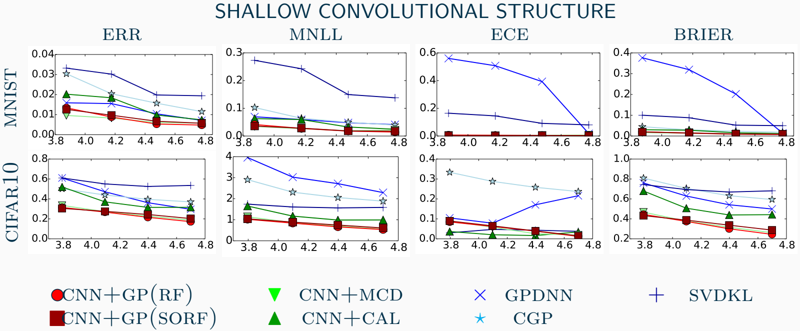
Bayesian CNNs with DGPs with Random Features

- We extended our work on Random Feature Expansions for DGPs to replace fully connected layers



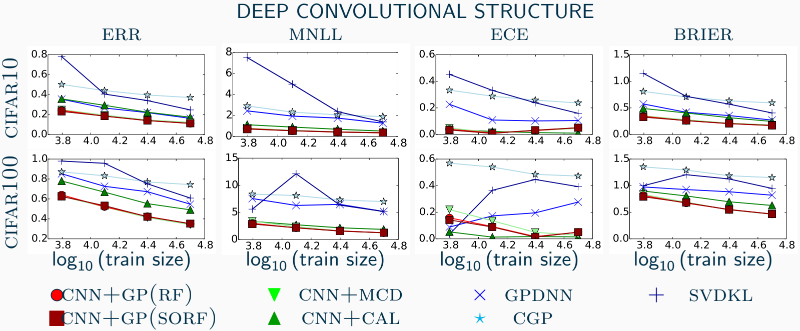
Tran et al., *arXiv*, 2018

Comparison with competitors



Tran et al., *arXiv*, 2018

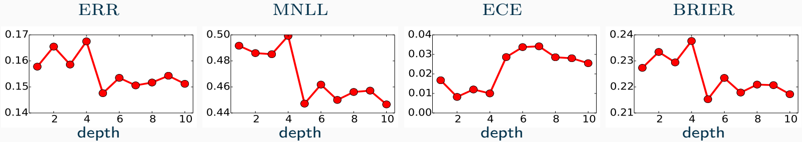
Comparison with competitors



Tran et al., *arXiv*, 2018

Analysis of Depth of DGP

- Increasing depth of DGP slightly improves error rate...
- ... and slightly worsen calibration



Tran et al., *arXiv*, 2018

Other Interesting DGP-based models

- Autoencoders Dai et al. *ICLR*, 2015 – Domingues et al., *Mach. Learn.*, 2018
- DGPs with constrained dynamics Lorenzi and Filippone, *ICML*, 2018

Conclusions

Conclusions

- DGPs offer probabilistic deep learning with sensible priors

Conclusions

- DGPs offer probabilistic deep learning with sensible priors
- Inference for DGPs is hard
  - Model approximations
  - Approximate inference
- Difficult to assess the impact of these approximations

Conclusions

- We are borrowing ideas from GPs and deep learning
  - Stochastic-based approximate inference
  - Low-rank process decompositions
  - Algebraic/computational tricks

Conclusions

- Combinations of GPs with CNNs slightly disappointing
    - Quantification of uncertainty not for free…
    - … regularization of filters is necessary
    - Performance gains are small compared to plain CNNs

We are hiring PhDs, Post-docs and Assistant Professors

Acknowledgments and References

Thank you!

References

Bayesian Deep Nets and Deep Gaussian Processes

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. **Random feature expansions for deep Gaussian processes.** In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893, International Convention Centre, Sydney, Australia, Aug. 2017. PMLR.

D. K. Duvenaud, O. Rippel, R. P. Adams, and Z. Ghahramani. **Avoiding pathologies in very deep networks.** In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 202–210. JMLR.org, 2014.

Y. Gal and Z. Ghahramani. **Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.** In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1050–1059. JMLR.org, 2016.

J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri. **Deep Neural Networks as Gaussian Processes.** In *International Conference on Learning Representations*, 2018.

A. G. De G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. **Gaussian Process Behaviour in Wide Deep Neural Networks.** In *International Conference on Learning Representations*, 2018.

R. M. Neal. **Bayesian Learning for Neural Networks (Lecture Notes in Statistics).** Springer, 1 edition, Aug. 1996.

References

Inference for Deep Gaussian Processes

T. D. Bui, D. Hernández-Lobato, J. M. Hernández-Lobato, Y. Li, and R. E. Turner. **Deep Gaussian Processes for Regression using Approximate Expectation Propagation.** In M.-F. Balcan and K. Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48, pages 1472–1481. JMLR.org, 2016.

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. **Random feature expansions for deep Gaussian processes.** In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893, International Convention Centre, Sydney, Australia, Aug. 2017. PMLR.

A. C. Damianou and N. D. Lawrence. **Deep Gaussian Processes.** In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, volume 31 of *JMLR Proceedings*, pages 207–215. JMLR.org, 2013.

J. Hensman and N. D. Lawrence. **Nested Variational Compression in Deep Gaussian Processes, Dec. 2014.**

M. Havasi, J. M. Hernández-Lobato, and J. J. Murillo-Fuentes. **Inference in Deep Gaussian Processes using Stochastic Gradient Hamiltonian Monte Carlo, June 2018.** arXiv:1806.05490.

M. D. Hoffman. **Learning deep latent Gaussian models with Markov chain Monte Carlo.** In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1510–1519, International Convention Centre, Sydney, Australia, Aug. 2017. PMLR.

H. Salimbeni and M. Deisenroth. **Doubly Stochastic Variational Inference for Deep Gaussian Processes.** In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4588–4599. Curran Associates, Inc., 2017.

References

Convolutional Nets and Gaussian Processes

J. Bradshaw, A. G. De G. Matthews, and Z. Ghahramani. **Adversarial Examples, Uncertainty, and Transfer Testing Robustness in Gaussian Process Hybrid Deep Networks, July 2017.** arXiv:1707.02476.

R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth. **Manifold Gaussian Processes for regression.** In *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*, pages 3338–3345, 2016.

V. Kumar, V. Singh, P. K. Srijith, and A. Damianou. **Deep Gaussian Processes with Convolutional Kernels, June 2018.** arXiv:1806.01655.

G.-L. Tran, E. V. Bonilla, J. P. Cunningham, P. Michiardi, and M. Filippone. **Calibrating Deep Convolutional Gaussian Processes, May 2018.** arXiv:1805.10522.

M. van der Wilk, C. E. Rasmussen, and J. Hensman. **Convolutional Gaussian Processes.** In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2849–2858. Curran Associates, Inc., 2017.

A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing. **Stochastic Variational Deep Kernel Learning.** In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2586–2594. Curran Associates, Inc., 2016.

References

Bayesian Convolutional Nets

Y. Gal and Z. Ghahramani. **Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference, Jan. 2016.** arXiv:1506.02158.

A. Garriga-Alonso, L. Aitchison, and C. E. Rasmussen. **Deep Convolutional Networks as shallow Gaussian Processes, Aug. 2018.** arXiv:1808.05587.

F. Laumann, K. Shridhar, and A. L. Maurin. **Bayesian Convolutional Neural Networks, June 2018.** arXiv:1806.05978.

Calibration of (Bayesian) Convolutional Nets

A. Niculescu-Mizil and R. Caruana. **Predicting Good Probabilities with Supervised Learning.** In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 625–632, New York, NY, USA, 2005. ACM.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. **On Calibration of Modern Neural Networks.** In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, International Convention Centre, Sydney, Australia, Aug. 2017. PMLR.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. **Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.** In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017.

G.-L. Tran, E. V. Bonilla, J. P. Cunningham, P. Michiardi, and M. Filippone. **Calibrating Deep Convolutional Gaussian Processes, May 2018.** arXiv:1805.10522.

References

Random Feature Expansions for Shallow Gaussian Processes

Q. Le, T. Sarlos, and A. Smola. **Fastfood - Approximating Kernel Expansions in Loglinear Time.** In *30th International Conference on Machine Learning (ICML)*, 2013.

A. Rahimi and B. Recht. **Random Features for Large-Scale Kernel Machines.** In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

F. X. Yu, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar. **Orthogonal Random Features.** In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1975–1983. Curran Associates, Inc., 2016.

Random Feature Expansions for Deep Gaussian Processes

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. **Random feature expansions for deep Gaussian processes.** In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893, International Convention Centre, Sydney, Australia, Aug. 2017. PMLR.

Y. Gal and Z. Ghahramani. **Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.** In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1050–1059. JMLR.org, 2016.

References

Variational Inference

A. Graves. **Practical Variational Inference for Neural Networks.** In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2348–2356. Curran Associates, Inc., 2011.

D. P. Kingma and M. Welling. **Auto-Encoding Variational Bayes.** In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, Apr. 2014.

References

Unsupervised learning with Deep Gaussian Processes

Z. Dai, A. Damianou, J. González, and N. Lawrence. **Variational Auto-encoded Deep Gaussian Processes, Feb. 2016.**

R. Domingues, P. Michiardi, J. Zouaoui, and M. Filippone. **Deep Gaussian process autoencoders for novelty detection.** *Machine Learning*, 107(8-10):1363–1383, 2018.