SAMPLING-FREE PREDICTIVE UNCERTAINTY USING GAUSSIAN PROCESSES WITH A NEURAL TANGENT KERNEL

Jongseok Lee and Rudolph Triebel

DLR

What do we mean by ‘uncertainty’?

Return a distribution over predictions rather than a single prediction.

Classification

- Output label along with its confidence.

Regression

- Output mean along with its variance.

Well-calibrated uncertainty estimates quantify **when we can trust** the predictions from machine learning models → **Trustworthy AI systems!**

Challenges – uncertainty in neural networks

Deterministic methods

Efficiency

Stochastic methods

Uncertainty Quality

Challenges – uncertainty in neural networks

Deterministic methods

Efficiency

Stochastic methods

Uncertainty Quality

Stochastic methods

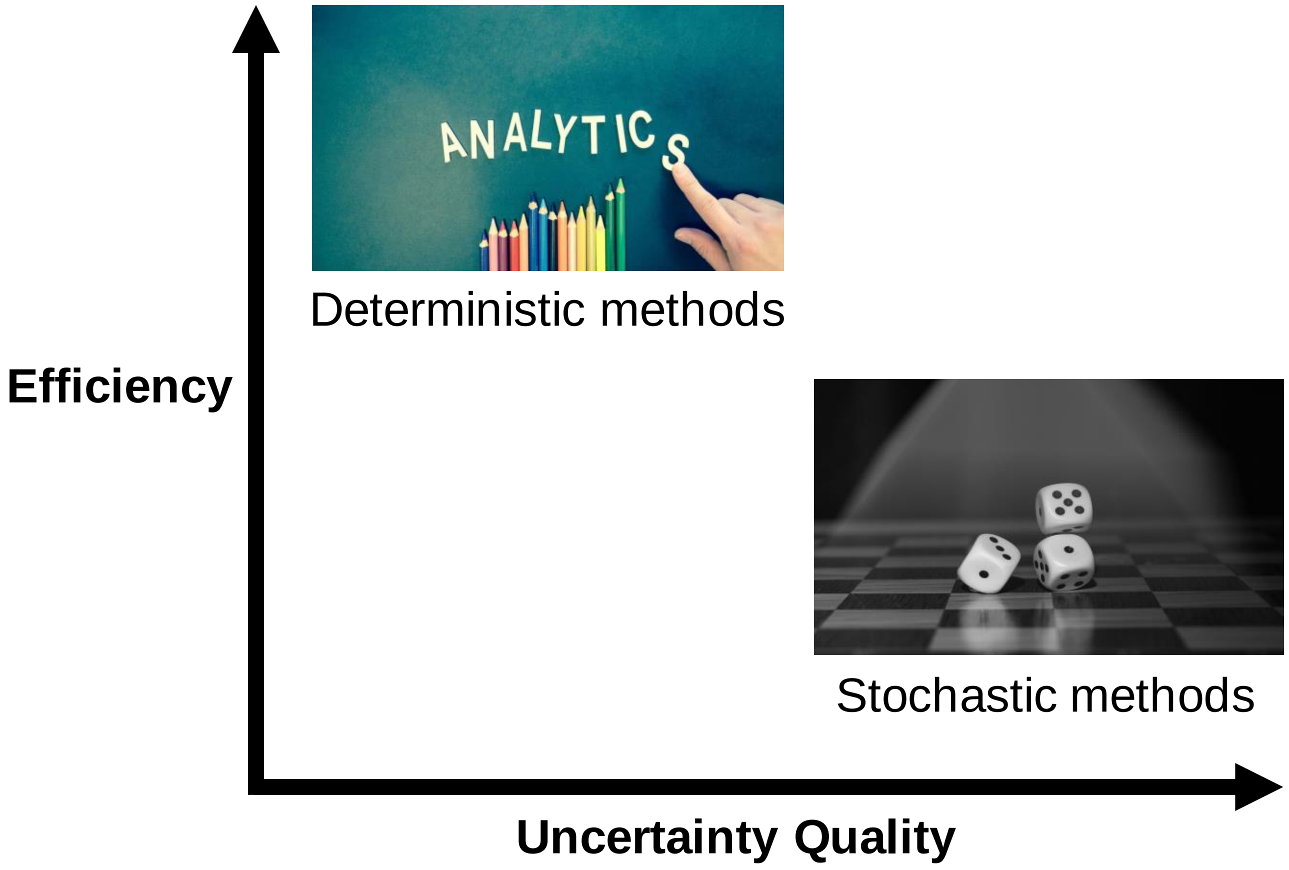
$$p(y^*|\mathcal{D}, x^*) = \int p(y^*|x^*, w)\, p(w|\mathcal{D})\, dw$$

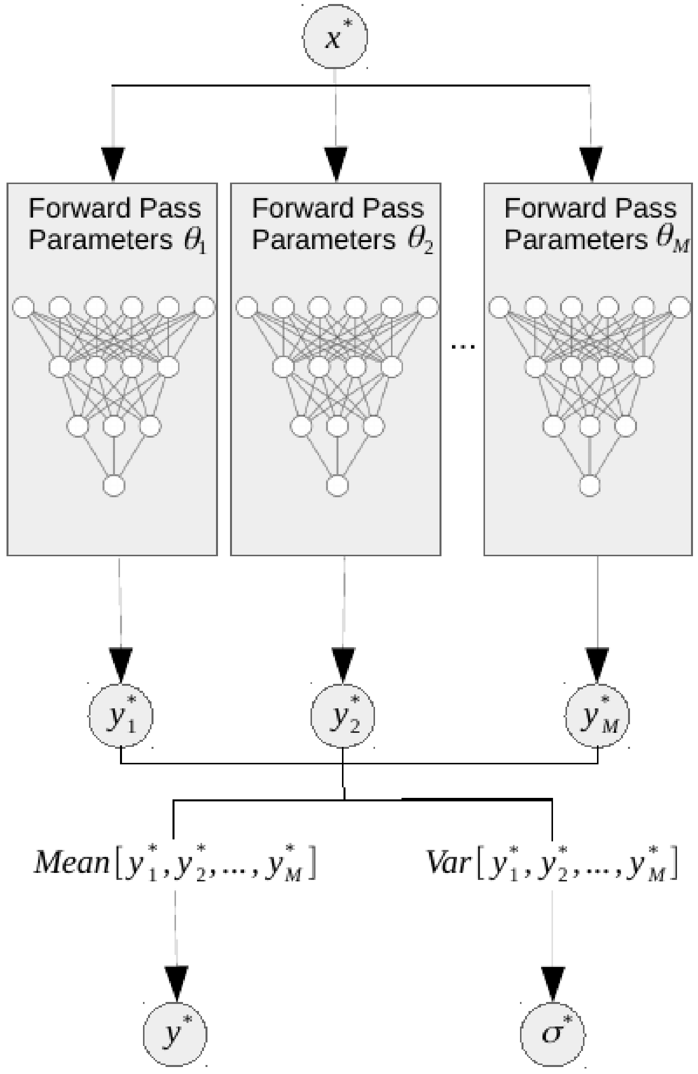
$$\approx \frac{1}{T}\sum_{t=1}^{T} y^*(x^*, w_t^s) \qquad w_t^s \sim p(w|X, Y).$$

Well-known examples

- MC-dropout (Gal et al 2015).
- Deep ensemble (Lakshminarayanan et a 2017).

Challenges – uncertainty in neural networks





Challenges – uncertainty in neural networks

Deterministic methods

Efficiency

Stochastic methods

Uncertainty Quality

Deterministic methods

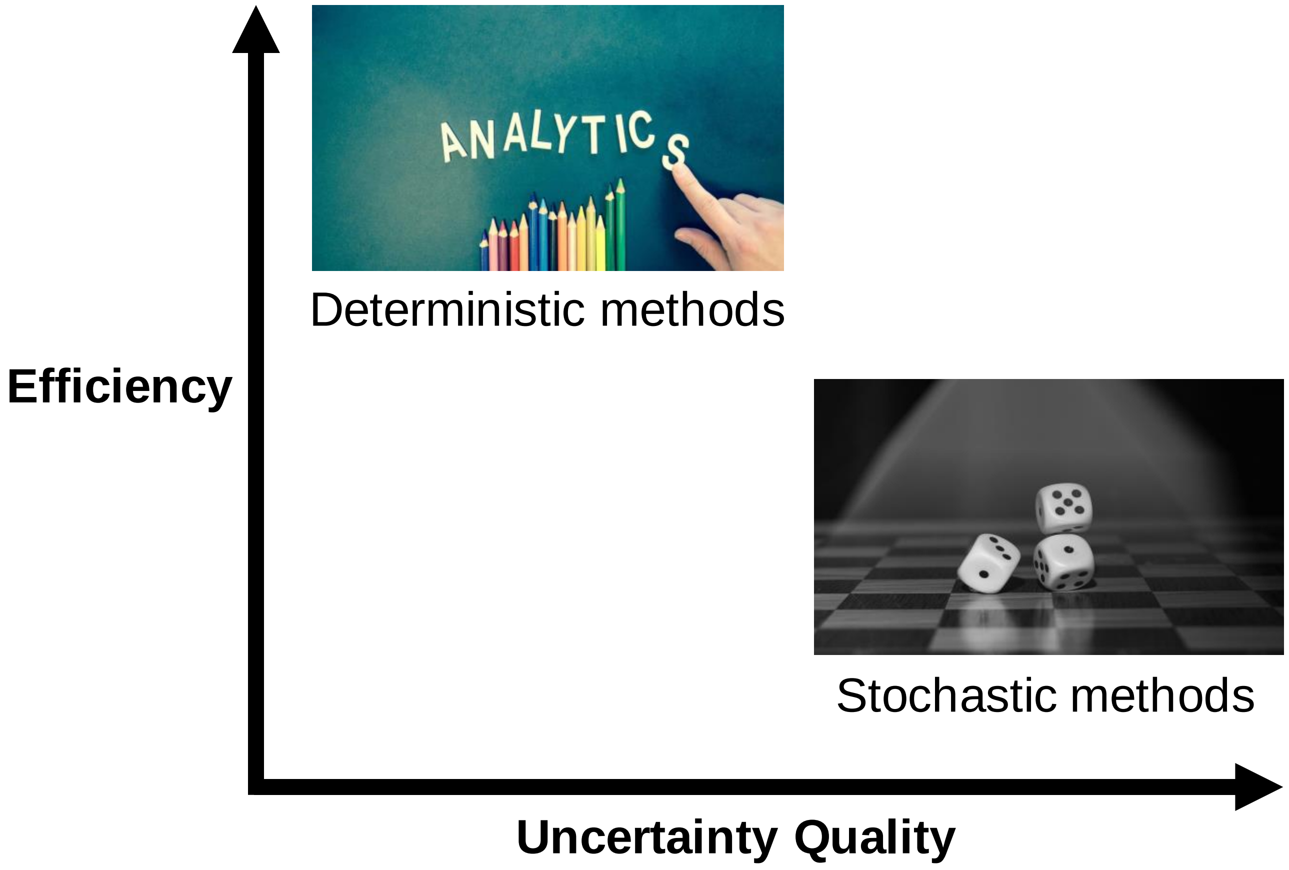
$$p(\boldsymbol{y}^*|\mathcal{D},\boldsymbol{x}^*) = \int \boldsymbol{p}(\boldsymbol{y}^*|\boldsymbol{x}^*,\boldsymbol{w})\boldsymbol{p}(\boldsymbol{w}|\mathcal{D})d\boldsymbol{w}$$

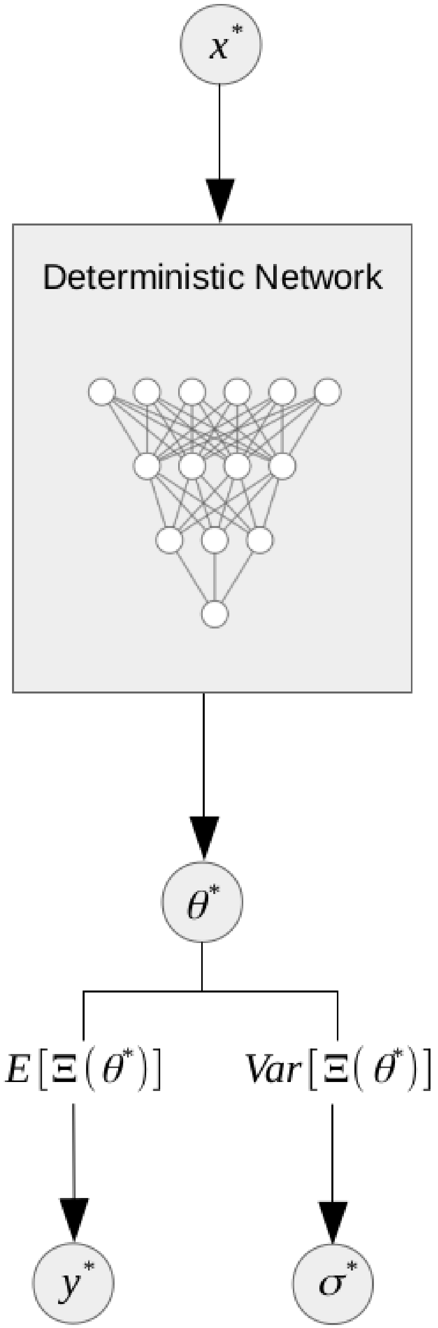
$$\approx \boldsymbol{f}(\mathbf{x}^*,\boldsymbol{w}).$$

Well-known examples

- Distillation (Korattikara et al 2015).
- Linear propagation (Postels et a 2019).

Challenges – uncertainty in neural networks





Challenges – uncertainty in neural networks

★ **How to keep the efficiency of deterministic methods while increasing the quality?**

Efficiency

Deterministic methods

Stochastic methods

Uncertainty Quality

Challenges – uncertainty in neural networks

Deterministic methods

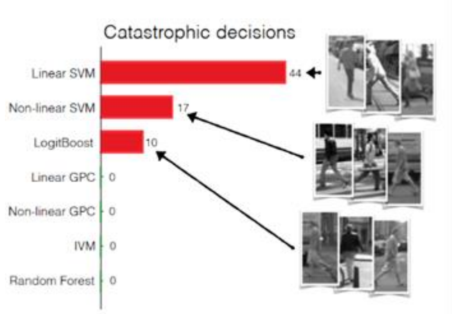
Efficiency

Stochastic methods

Uncertainty Quality

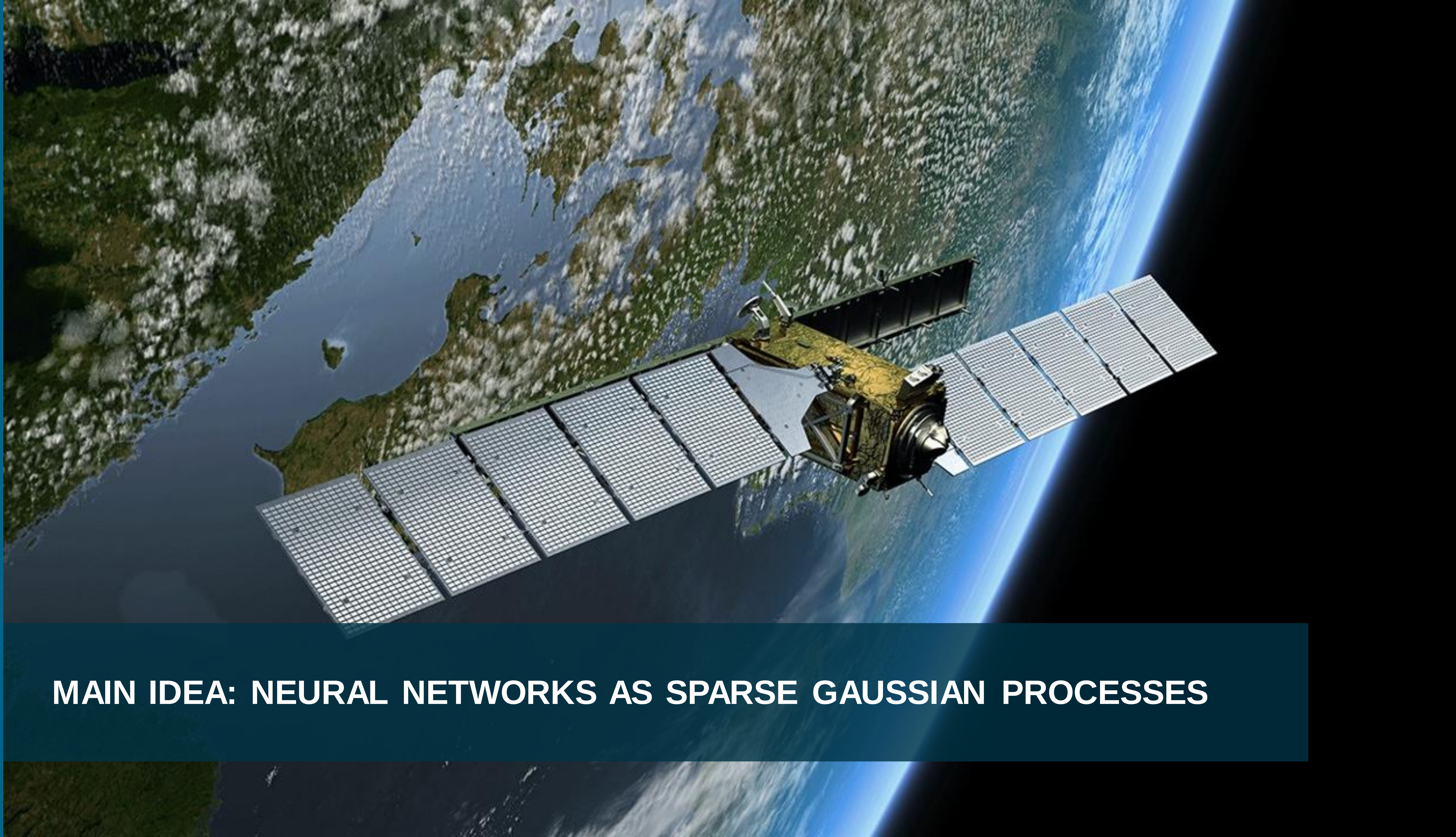
How to keep the efficiency of deterministic methods while increasing the quality?

- Gaussian Processes (GPs) as golden standards of probabilistic machine learning.
  - (Rasmussen and Williams 2006, MIT Press)



(Grimmett et al, 2016 IJRR)

- State-of-the-art GPs – efficient predictions, e.g., (Pleiss et al, 2018 ICML).



MAIN IDEA: NEURAL NETWORKS AS SPARSE GAUSSIAN PROCESSES

Neural Tangent Kernel theory – inspirations

Earlier works in 1990s:

Radford Neal, “Priors for Infinite Networks”, 1994.

- Pioneered the connections between GPs and neural networks.
- Assume increasing width, single hidden layer, independent priors on neural network weights.
- Then, neural networks converge to GPs with a specific kernel, known as “the Neural Tangent Kernel (NTK)”.

Recent breakthroughs

- Multiple hidden layers!
→ (J.Lee et al, ICLR 2018).
→ (Matthews et al, ICLR 2018).

- Convolution layers!
→ (Alonso et al, ICLR 2019).

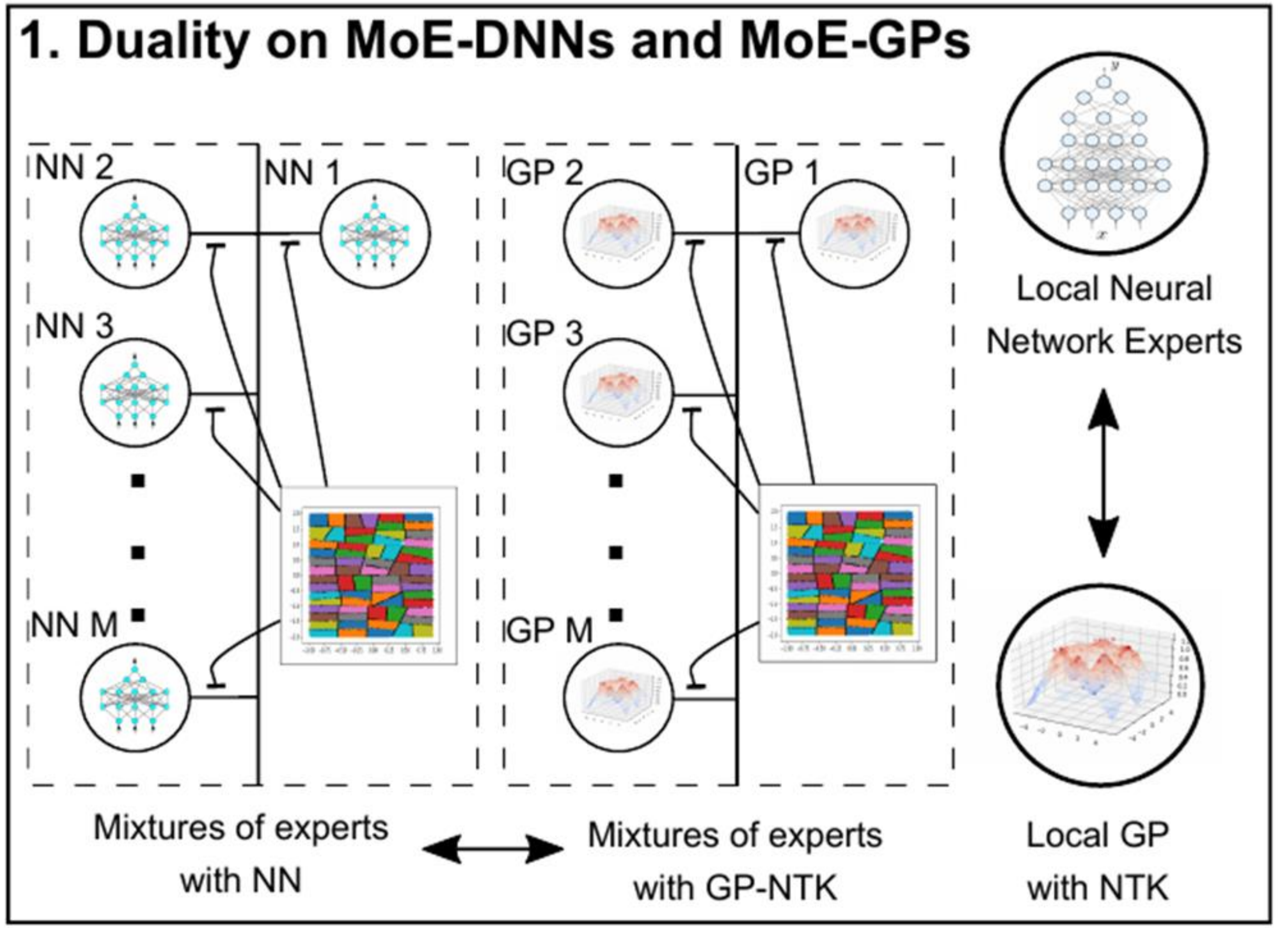
- Bayesian inference!
→ (Khan et al, NeurIPS 2019).

- Finite width!
→ (Novak et al, ICML 2022).

And many others!

These works greatly advance the state-of-the-art learning theory of deep learning!

The derived theory and proof paths



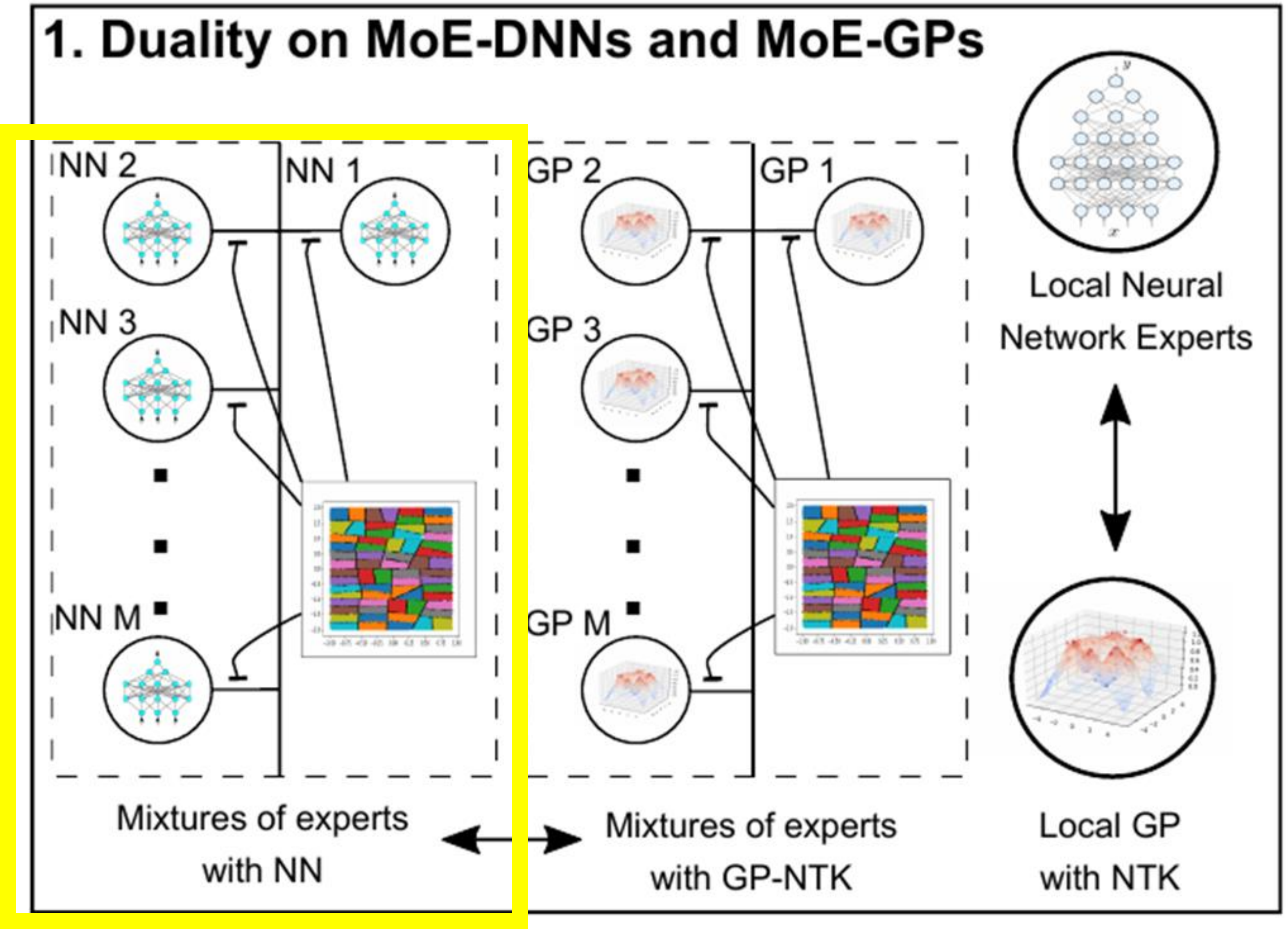
The derived theory and proof paths

Preliminaries

- Mixtures of experts (MoE) are an ensemble model with a gating function and many experts/models (Jacobs et al 1991).
- Assume a strict division of data.

$$y = \sum_{m=1}^{M} g_m(x) f_w(x)$$

with 1 pick at the time.



NN is any neural networks (MLP, convolution, etc.)

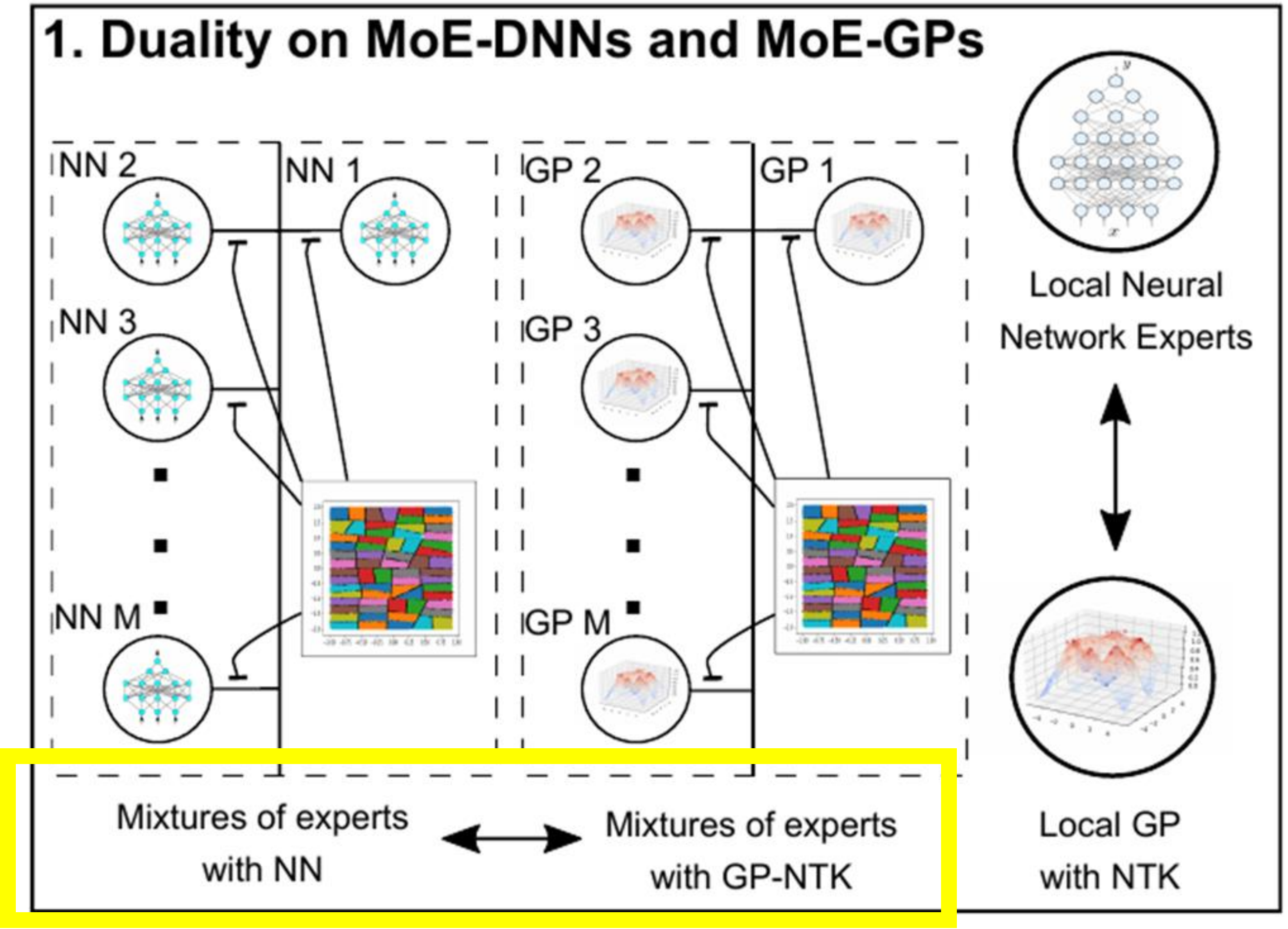
that have valid Jacobians!

The derived theory and proof paths

Preliminaries

- Mixtures of experts (MoE) are an ensemble model with a gating function and many experts/models (Jacobs et al 1991).
- Assume a strict division of data.

Bayesian Duality



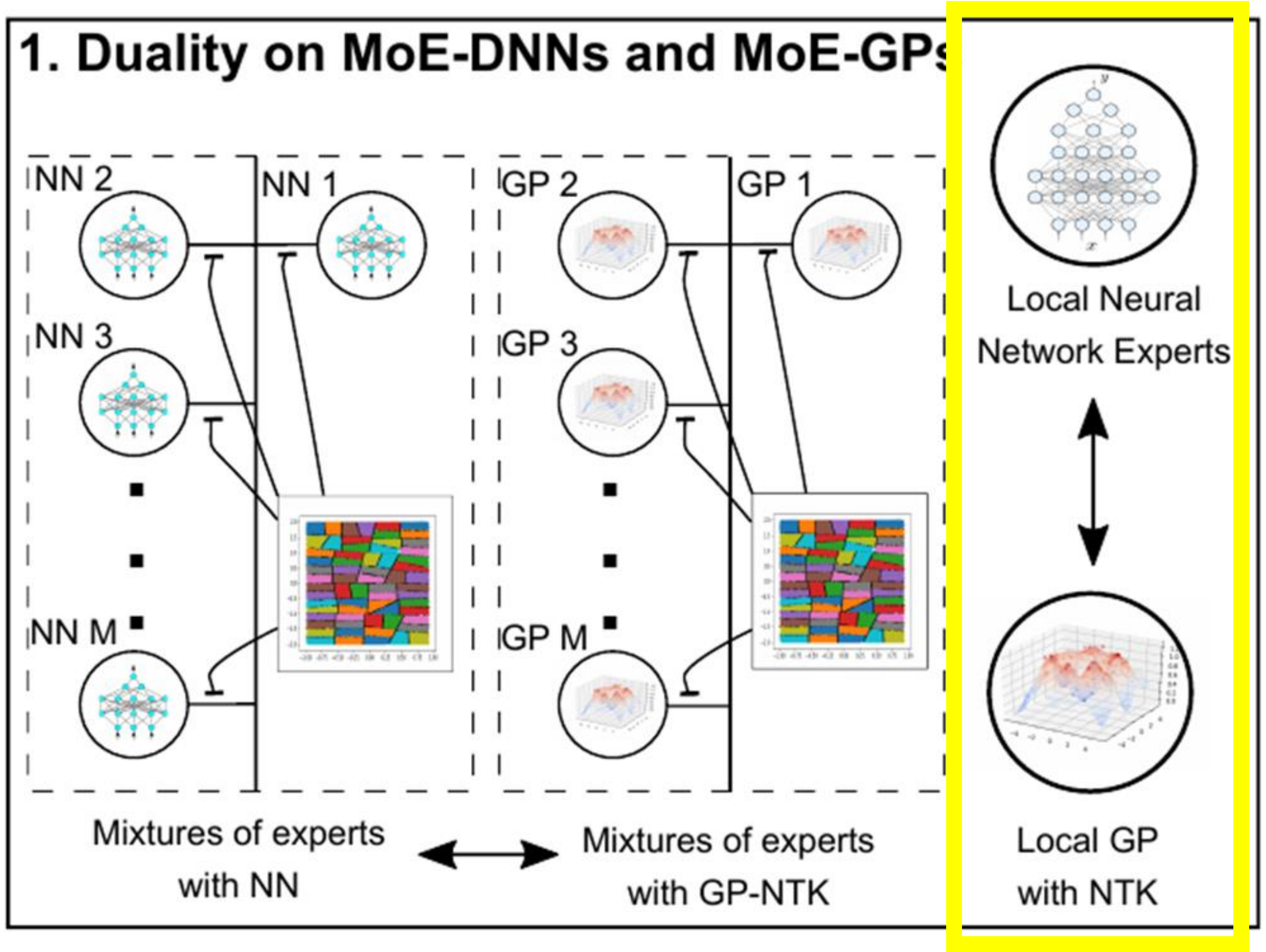
The derived theory and proof paths

Preliminaries

- Mixtures of experts (MoE) are an ensemble model with a gating function and many experts/models (Jacobs et al 1991).
- Assume a strict division of data.

Bayesian Duality

1. Direct application of (Khan et al 2019) on each neural network (NNs) experts.



The derived theory and proof paths

Preliminaries

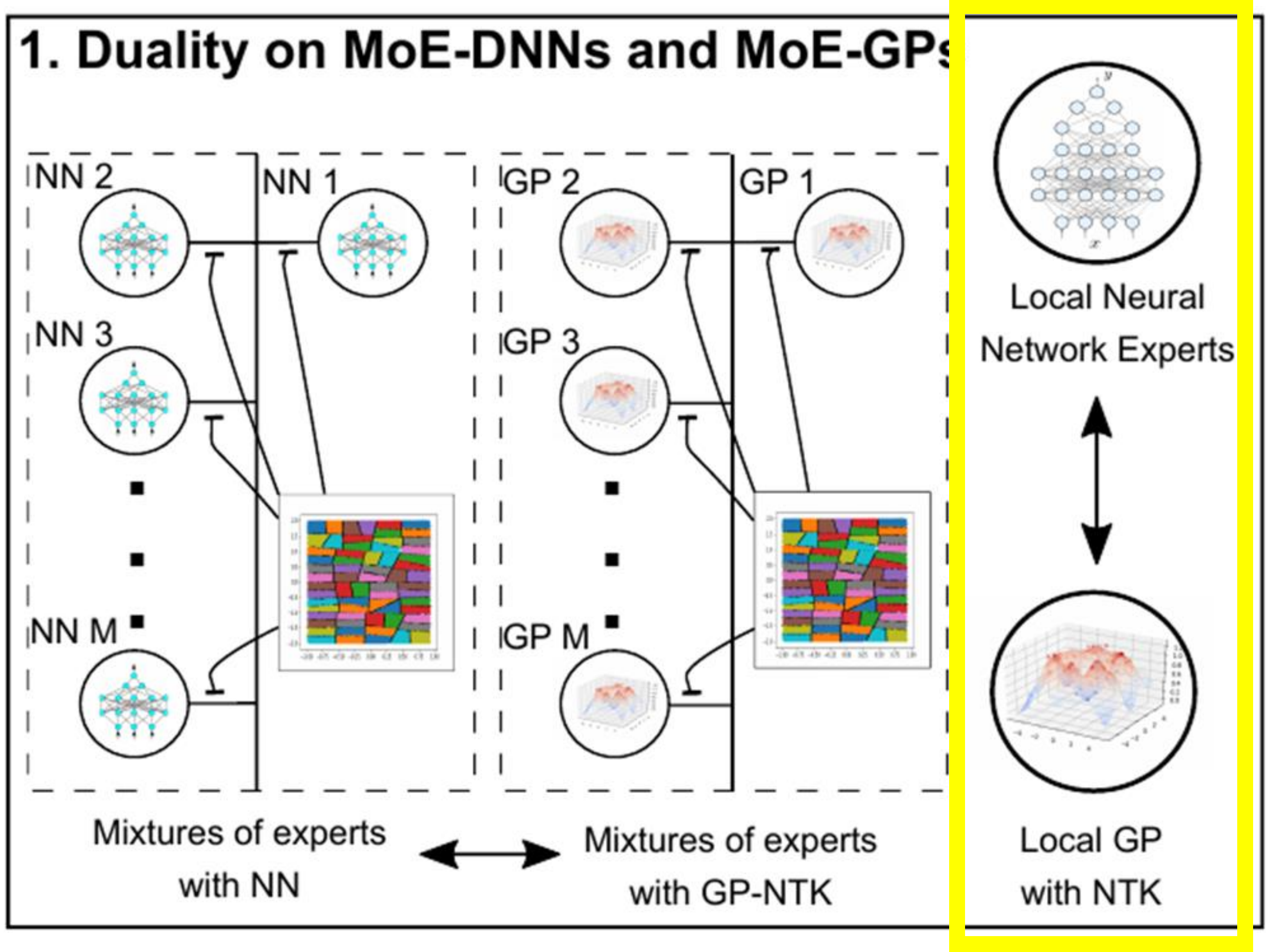
- Mixtures of experts (MoE) are an ensemble model with a gating function and many experts/models (Jacobs et al 1991).
- Assume a strict division of data.

Bayesian Duality

1. Direct application of (Khan et al 2019) on each neural network (NNs) experts.

2. Local DNN experts, cast as local GPs with the NTK (in a Bayesian sense).

$$p(w_m; D_m) \longrightarrow p\left(f_m; \widetilde{D_m}\right)$$



The derived theory and proof paths

Preliminaries

- Mixtures of experts (MoE) are an ensemble model with a gating function and many experts/models (Jacobs et al 1991).
- Assume a strict division of data.

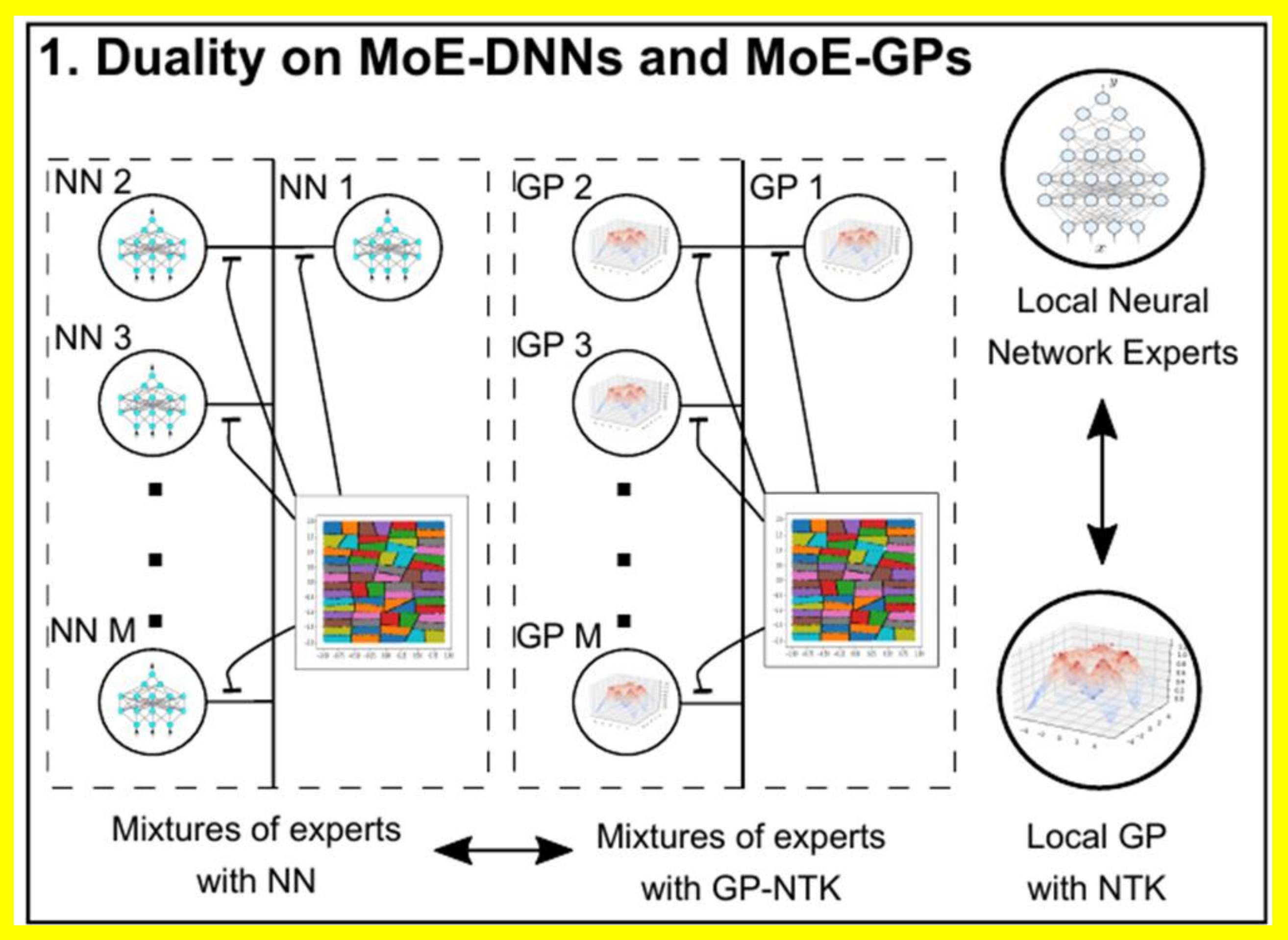
Bayesian Duality

1. Direct application of (Khan et al 2019) on each neural network (NNs) experts.

2. Local DNN experts, cast as local GPs with the NTK (in a Bayesian sense).

3. Then, probabilistic independence between each experts → a simple proof technique.

$$\prod p(w_m; D_m) \longrightarrow \prod p(f_m; \widetilde{D_m})$$



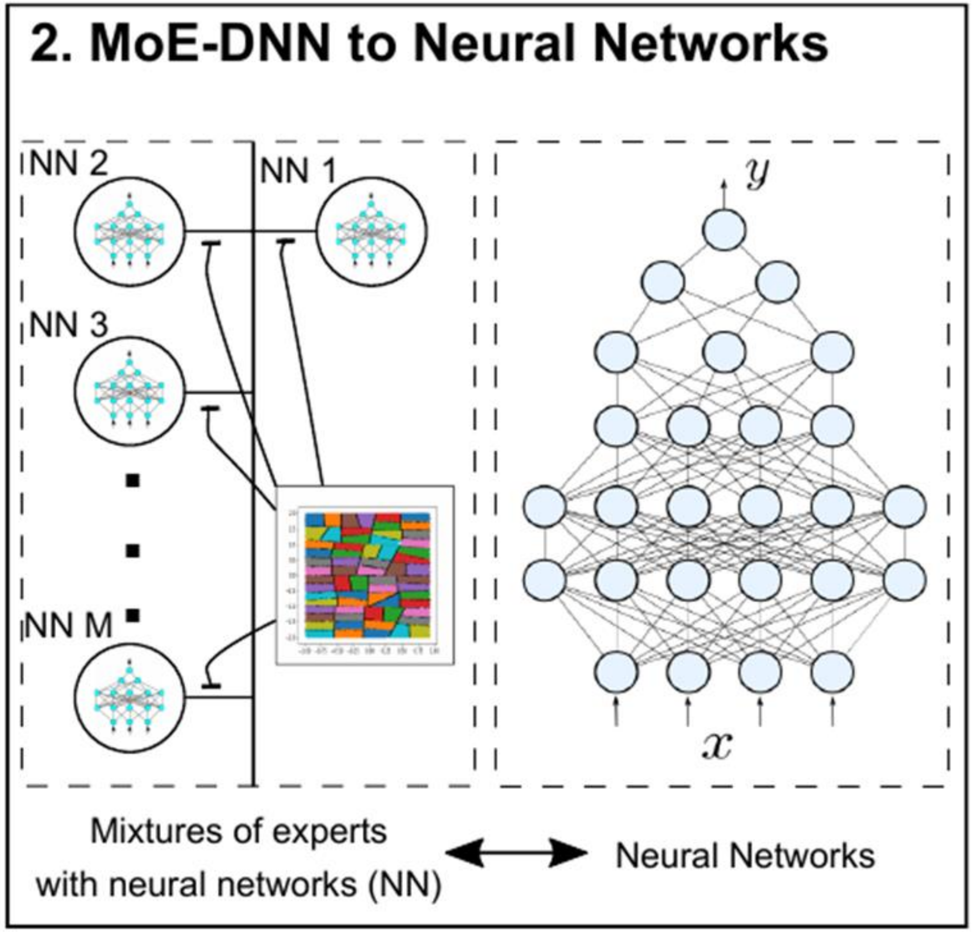
The derived theory and proof paths

Preliminaries

Bayesian Duality

Problem

- But we wanted to connect between a single DNN and MoE-GPs. Not MoE-NNs!



The derived theory and proof paths

Preliminaries

Bayesian Duality

Problem

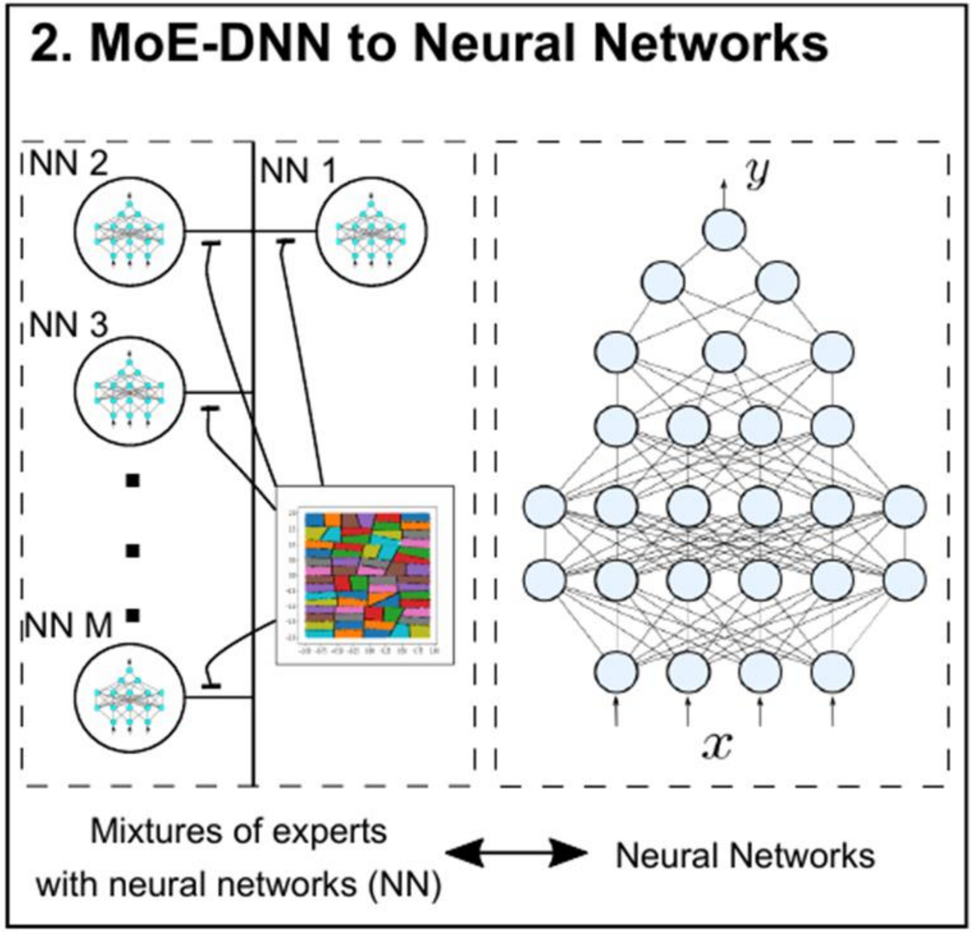
- But we wanted to connect between a single DNN and MoE-GPs. Not MoE-NNs!

Key insights

- Imagination that we don't try to train these models, but they are given pre-trained.

$$f_w(x) = f_{w_1}(x) = f_{w_2}(x) \ldots = f_{w_M}(x)$$

$$y = \sum_{m=1}^{M} g_m(x) f_w(x) = f_w(x)$$



The derived theory and proof paths

Preliminaries

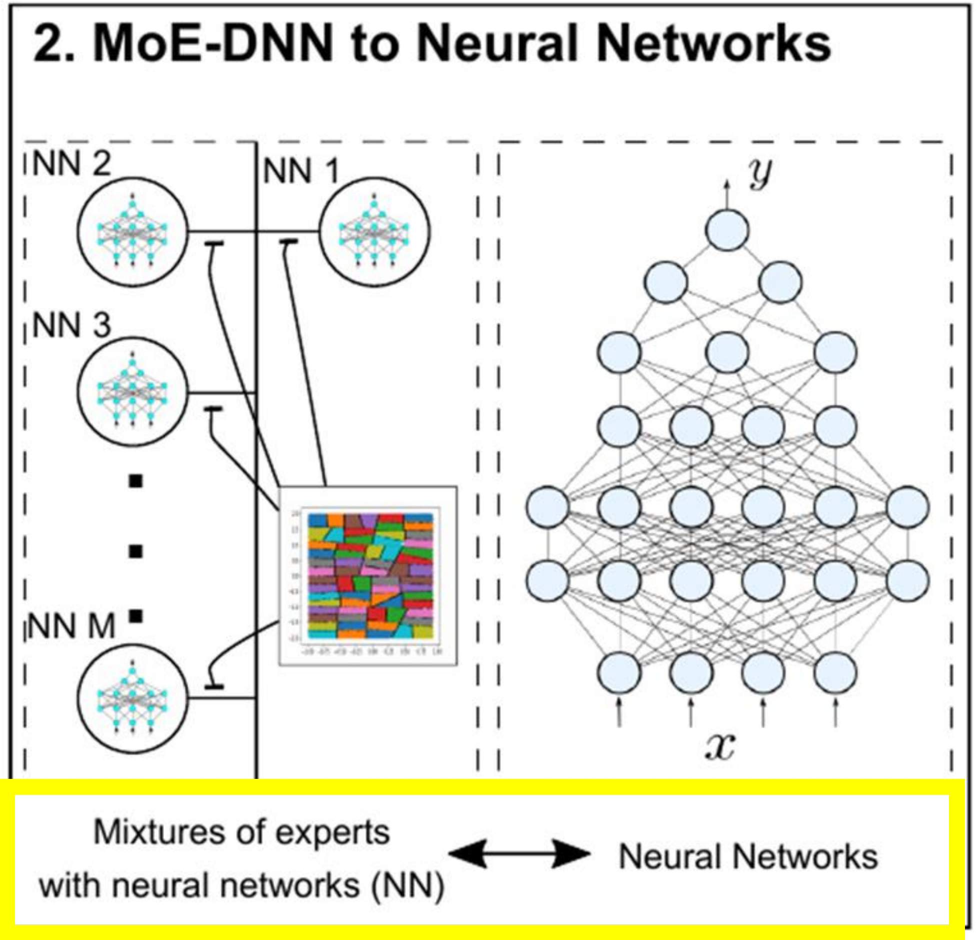
Bayesian Duality

Problem

- But we wanted to connect between a single DNN and MoE-GPs. Not MoE-NNs!

Key insights

- Imagination that we don't try to train these models, but they are given pre-trained.
- Due to hard portioning, we can prove that input-prediction relationships of a single DNN and a MoE-GP are equivalent, if all DNN experts are the same as a single NN.
- Single NN can be an already well trained NN with maximum likelihood principles.



The derived theory and proof paths

Preliminaries

Bayesian duality

Problem

Key insights

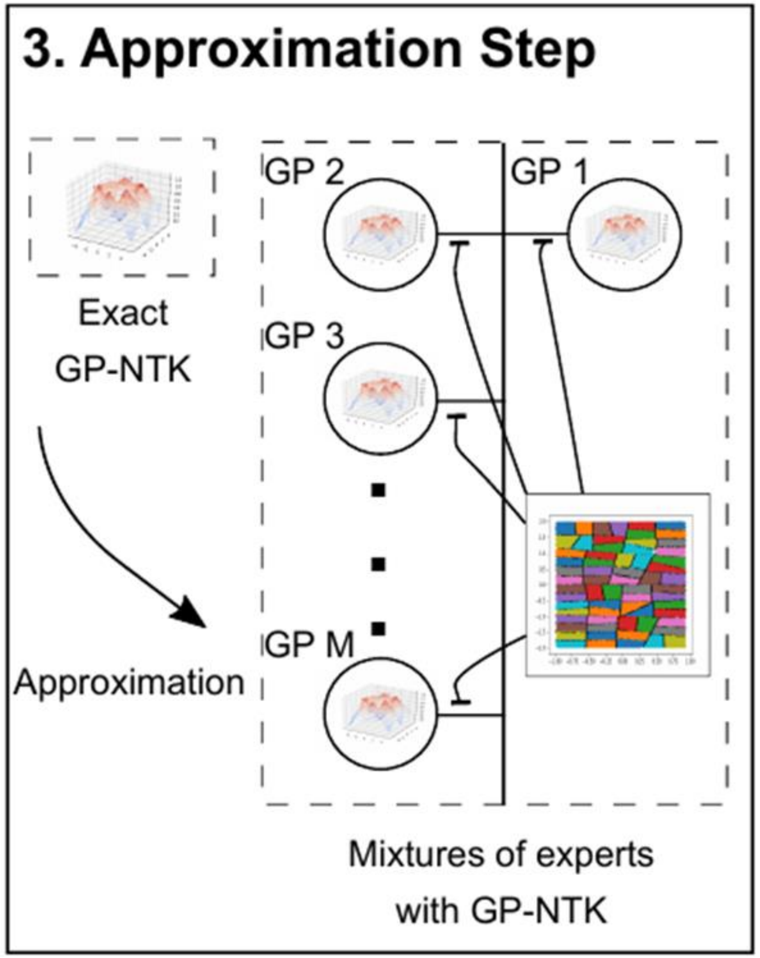
Final point

- This leads to an approximation step. A MoE-GP with the NTK approximates the true equivalent GP with NTK by:

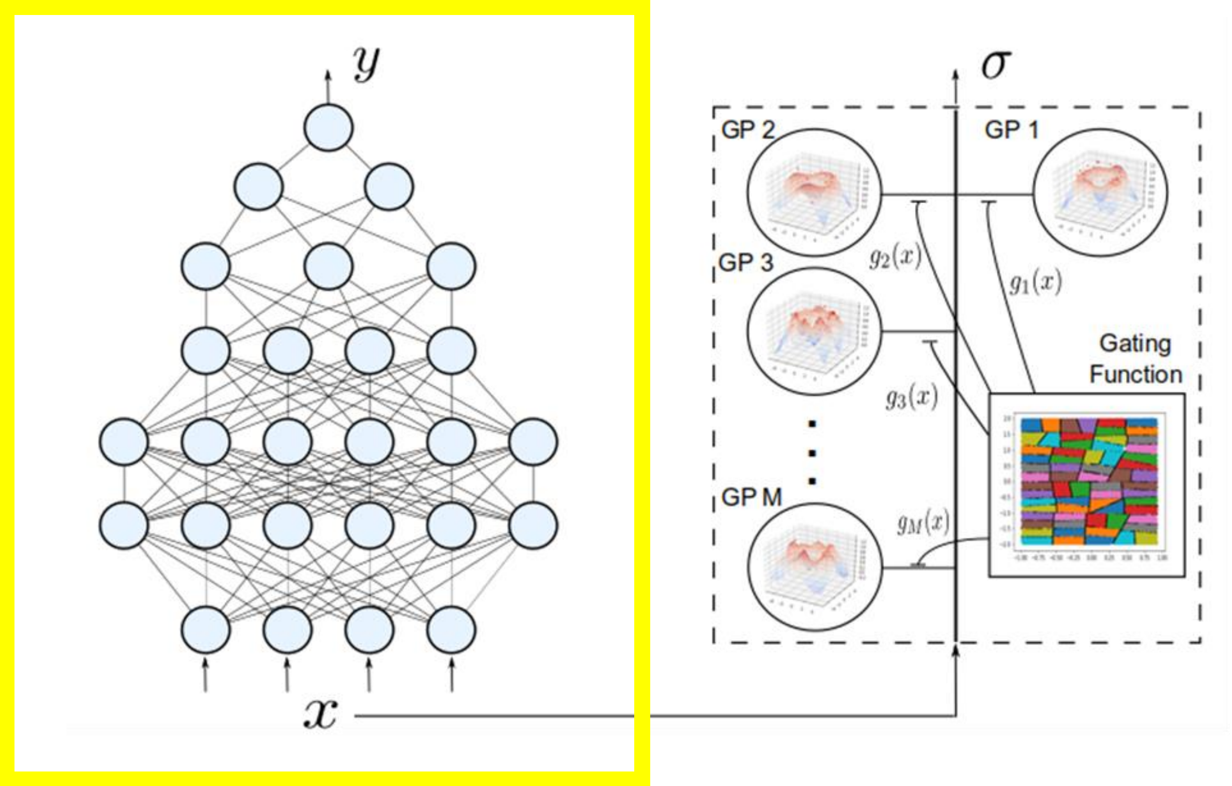
$$\|\boldsymbol{K}(\mathcal{X},\mathcal{X}) - \boldsymbol{K}_{true}(\mathcal{X},\mathcal{X})\|_{\mathrm{F}}^2 = \sum_{ij} \boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)^2 - \sum_{m=1}^{M} \sum_{ij\in\mathfrak{B}_m} \boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)^2$$

- This means closer data points in kernel space should be together, and otherwise, the data points can be separated apart.

- Revealing how a variant of sparse GPs can provably approximate uncertainty of DNN predictions.



The resulting predictive model



Main advantages by design:

- Neural Networks for accurate most-likely predictions.

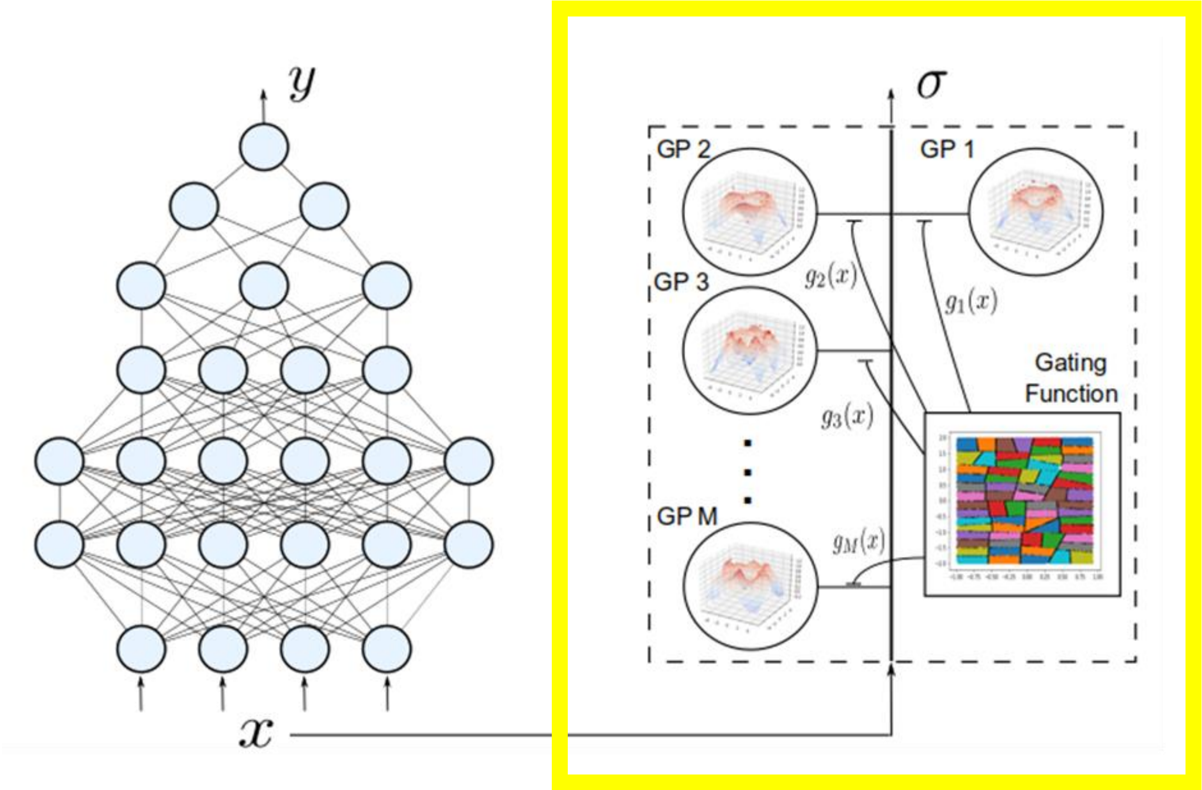
$y = f_w(x)$

- Sparse Gaussian Processes for well-calibrated uncertainty estimates:

$$\tilde{y} = \sum_{m=1}^{M} g_m(x) \tilde{f}_m(\mathbf{x}) + \epsilon_m$$

$$\tilde{f}_m(\mathbf{x}) \sim \mathrm{GP}\left(0, \frac{1}{\delta_m} \mathbf{J}_{\mathbf{f}_m}^{\mathrm{T}}(\mathbf{x}) \mathbf{J}_{\mathbf{f}_m}(\mathbf{x})\right)$$

The resulting predictive model



Main advantages by design:

- Neural Networks for accurate most-likely predictions.

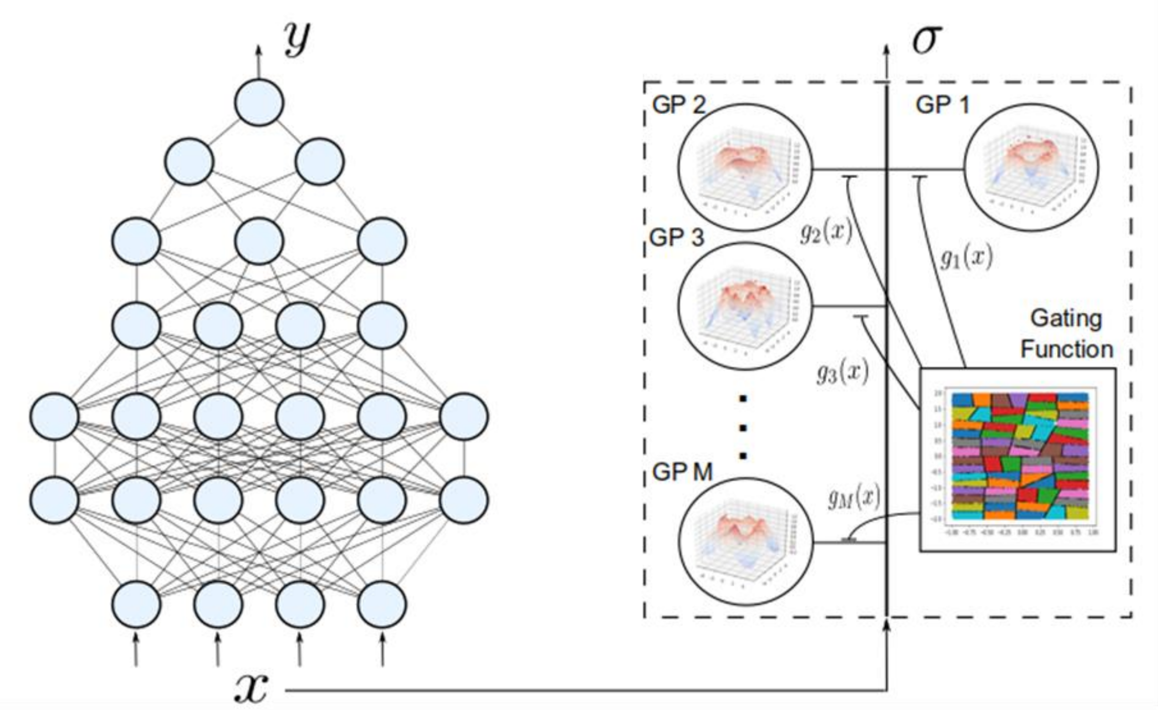
y*= *f*$_w$*(x)

- Sparse Gaussian Processes for well-calibrated uncertainty estimates:

$$\tilde{\mathbf{y}} = \sum_{m=1}^{M} g_m(x)\tilde{f}_m(\mathbf{x}) + \epsilon_m$$

$$\tilde{f}_m(\mathbf{x}) \sim \mathbf{GP}(0, \frac{1}{\delta_m} \mathbf{J}_{\mathbf{f_m}}^{\mathbf{T}}(\mathbf{x})\mathbf{J}_{\mathbf{f_m}}(\mathbf{x}))$$

The resulting predictive model



Note

$\mathbf{y} \neq \tilde{\mathbf{y}}$ but one can prove: $\boldsymbol{\sigma}(\mathbf{y}) = \boldsymbol{\sigma}(\tilde{\mathbf{y}})$

with $p(w; D) \approx \prod p(f_m; \widetilde{D_m})$

Main advantages by design:

- Neural Networks for accurate most-likely predictions.

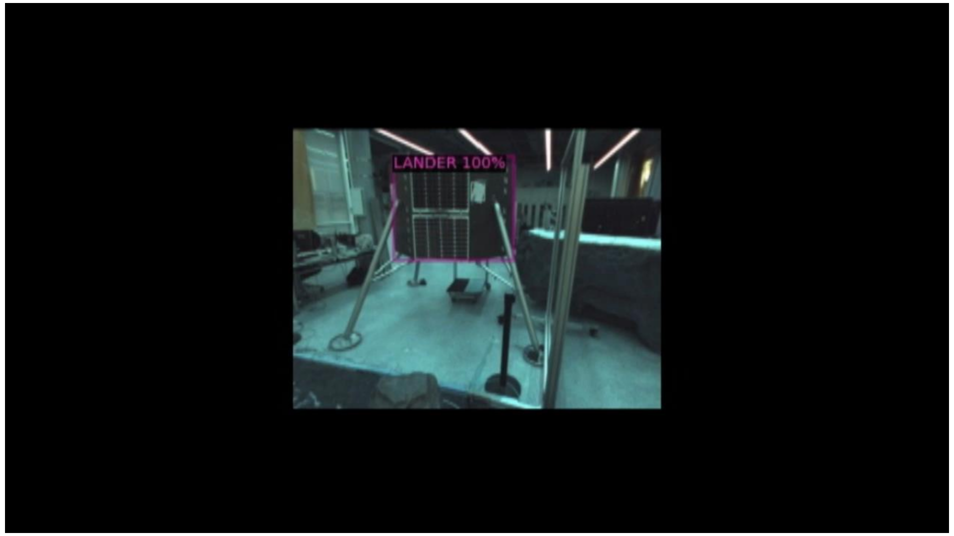
$y = f_w(x)$

- Sparse Gaussian Processes for well-calibrated uncertainty estimates:

$$\tilde{y} = \sum_{m=1}^{M} g_m(x)\tilde{f}_m(\mathbf{x}) + \epsilon_m$$

$$\tilde{f}_m(\mathbf{x}) \sim \mathbf{GP}(0, \frac{1}{\delta_m}\mathbf{J}_{\mathbf{f_m}}^{\mathbf{T}}(\mathbf{x})\mathbf{J}_{\mathbf{f_m}}(\mathbf{x}))$$

Results



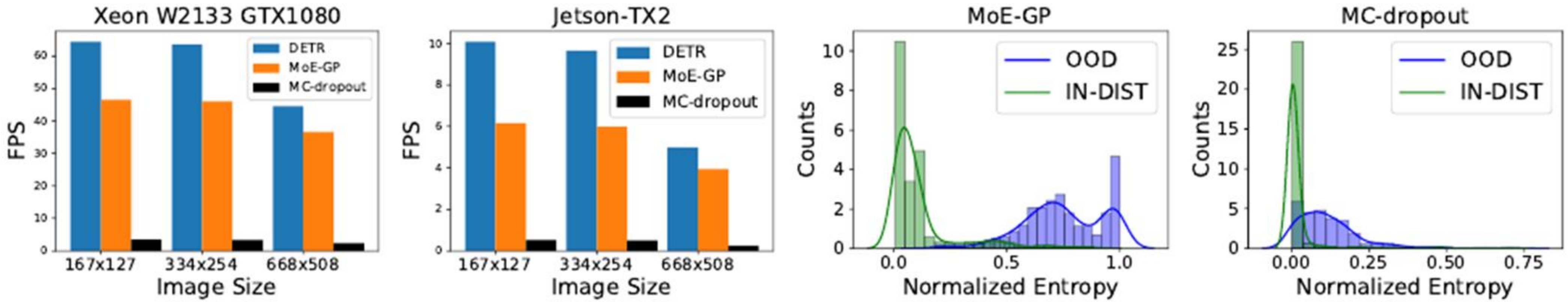
Nominal case (similar to training data)

Simulating failures (distributional shift)

Live demo at CoRL 2022 with a GPU-laptop to demonstrate real-time uncertainty estimates.

First real-time demo of deep learning uncertainty to our knowledge!

Results



- Run-time comparison on a GPU-desktop and an embedded GPU. Higher FPS the faster.
- Entropy histogram. More separable, better calibrated the uncertainty estimates.

Scalability test upto approx. 2 million data-points, ablation studies, comparison to five state-of-the-art methods across 12 evaluation setting, and toy examples are provided in the paper.

Main take-away/use-cases: when sparse GPs can scale, real-time uncertainty estimates from a GP formulation of neural networks can be obtained, improving over the state-of-the-art methods.

Conclusion

- The problem of sampling-free uncertainty estimation.
- Theoretic connection between neural networks and mixtures of GP experts through the neural tangent kernel → predictive model!
- Use-case: if sparse GPs can be tamed, faster and better uncertainty.

Neural Networks as Sparse Gaussian Processes for Uncertainty Quantification

J. Lee et al, “Trust Your Robots! Predictive Uncertainty Estimation of Neural Networks with Sparse Gaussian Processes”, CoRL 2021.