IS THE 21ST CENTURY THE PROMISED BAYESIAN CENTURY?

Ali Hebbal

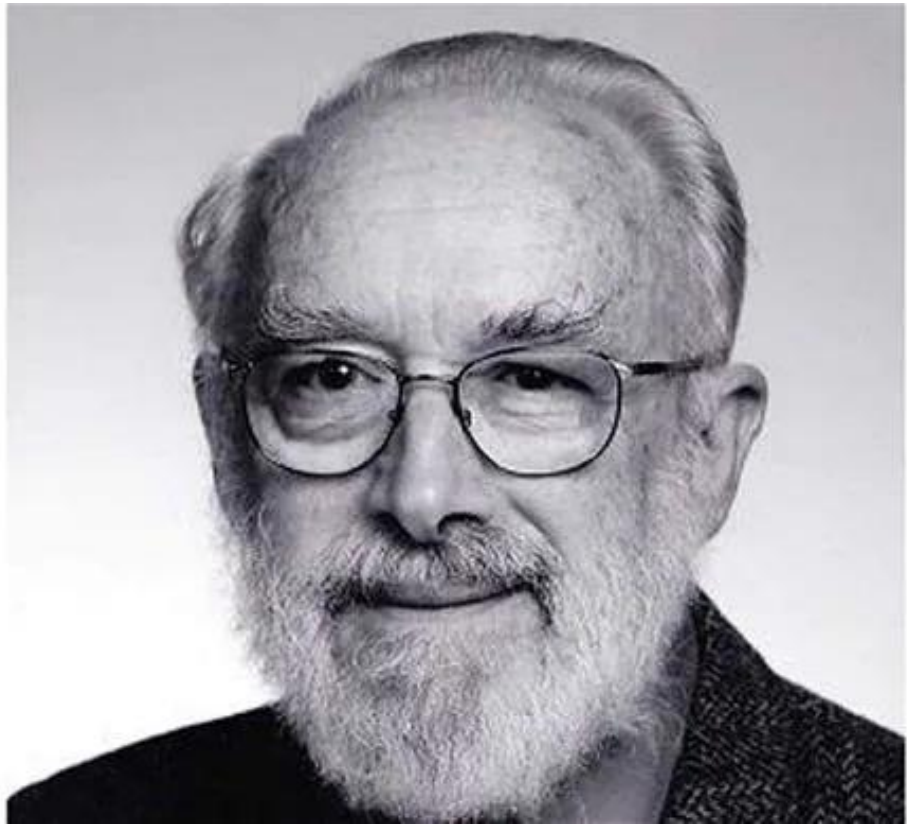
Capgemini engineering

ENGINEERING HORIZONS

A conference for engineers by engineers

Dennis Lindley

"The Future of Statistics: A Bayesian 21st Century " *in Advances in Applied Probability , 1975.*



What this talk is not:

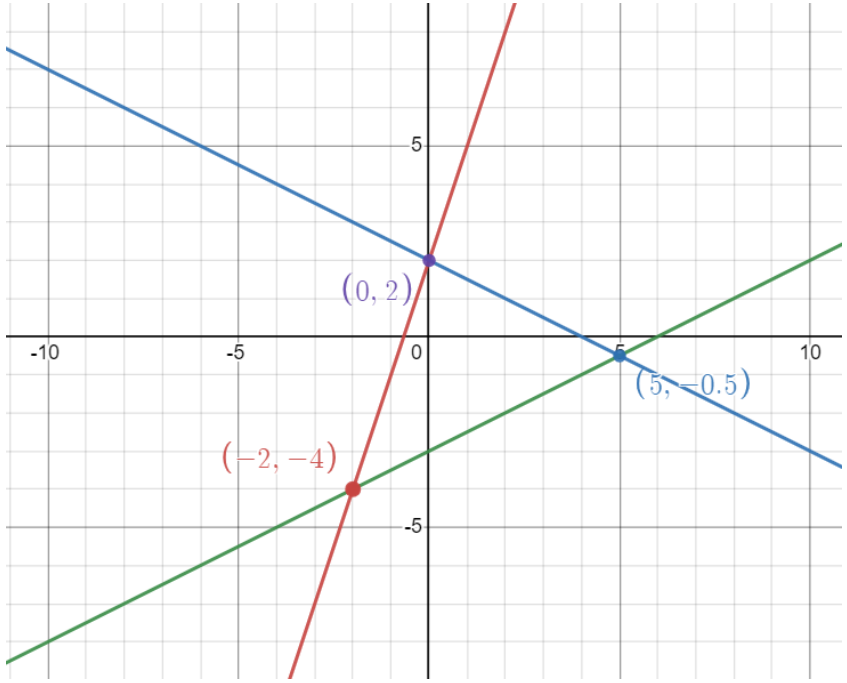
- A Fisherian vs Bayesian debate
- A technical review of Bayesian methods

What this talk is about:

- The importance of uncertainty quantification
- Bridging the gap between Bayesian and deep learning approaches
- **The Bayesian perspective is straight-forward and not that complex**

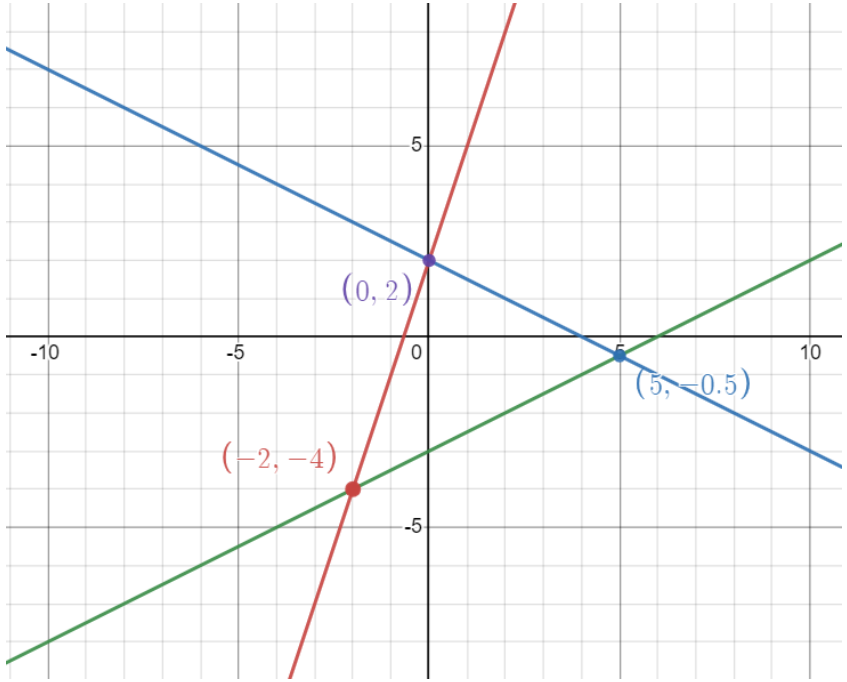
A BIT OF PHILOSOPHY

Fitting data: How did we go from over-determined systems to under-determined systems



A BIT OF PHILOSOPHY

Fitting data: How did we go from over-determined systems to under-determined systems

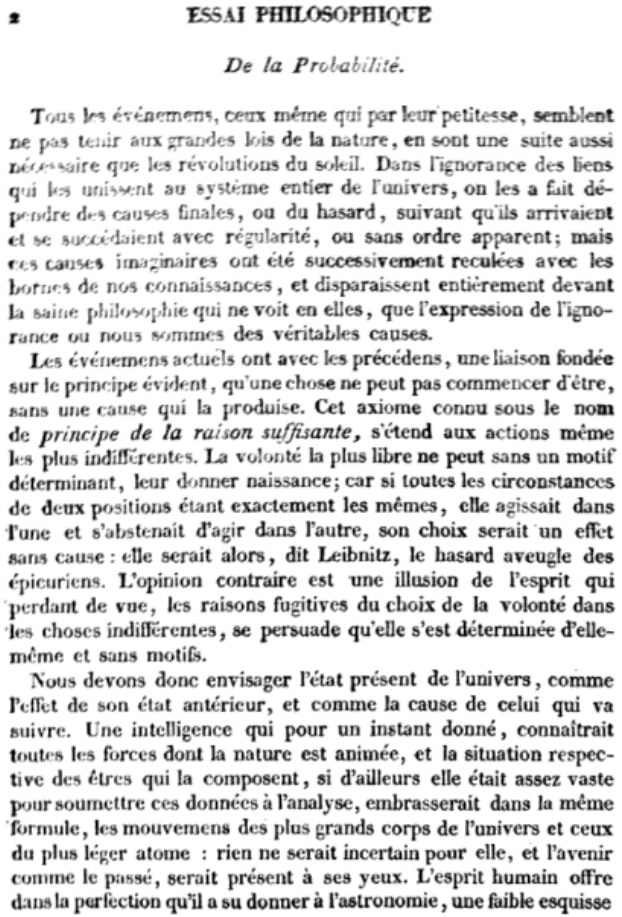


Pierre Simon Laplace

$$y = mx + c + \varepsilon$$

A BIT OF PHILOSOPHY

Essai philosophique sur les probabilités



2 ESSAI PHILOSOPHIQUE

De la Probabilité.

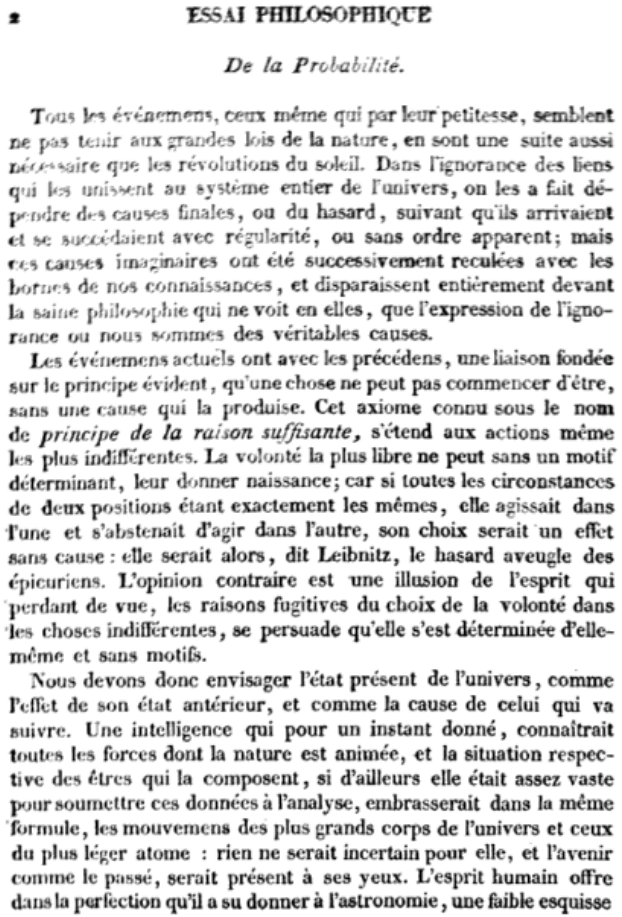
Tous les événemens, ceux même qui par leur petitesse, semblent ne pas tenir aux grandes lois de la nature, en sont une suite aussi nécessaire que les révolutions du soleil. Dans l'ignorance des liens qui les unissent au système entier de l'univers, on les a fait dépendre des causes finales, ou du hasard, suivant qu'ils arrivaient et se succédaient avec régularité, ou sans ordre apparent; mais ces causes imaginaires ont été successivement reculées avec les bornes de nos connaissances, et disparaissent entièrement devant la saine philosophie qui ne voit en elles, que l'expression de l'ignorance ou nous sommes des véritables causes.

Les événemens actuels ont avec les précédens, une liaison fondée sur le principe évident, qu'une chose ne peut pas commencer d'être, sans une cause qui la produise. Cet axiome connu sous le nom de *principe de la raison suffisante*, s'étend aux actions même les plus indifférentes. La volonté la plus libre ne peut sans un motif déterminant, leur donner naissance; car si toutes les circonstances de deux positions étant exactement les mêmes, elle agissait dans l'une et s'abstenait d'agir dans l'autre, son choix serait un effet sans cause : elle serait alors, dit Leibnitz, le hasard aveugle des épicuriens. L'opinion contraire est une illusion de l'esprit qui perdant de vue, les raisons fugitives du choix de la volonté dans les choses indifférentes, se persuade qu'elle s'est déterminée d'elle-même et sans motifs.

Nous devons donc envisager l'état présent de l'univers, comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule, les mouvemens des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux. L'esprit humain offre dans la perfection qu'il a su donner à l'astronomie, une faible esquisse

A BIT OF PHILOSOPHY

Essai philosophique sur les probabilités



2 ESSAI PHILOSOPHIQUE

De la Probabilité.

Tous les événemens, ceux même qui par leur petitesse, semblent ne pas tenir aux grandes lois de la nature, en sont une suite aussi nécessaire que les révolutions du soleil. Dans l'ignorance des liens qui les unissent au système entier de l'univers, on les a fait dépendre des causes finales, ou du hasard, suivant qu'ils arrivaient et se succédaient avec régularité, ou sans ordre apparent; mais ces causes imaginaires ont été successivement reculées avec les bornes de nos connaissances, et disparaissent entièrement devant la saine philosophie qui ne voit en elles, que l'expression de l'ignorance où nous sommes des véritables causes.

Les événemens actuels ont avec les précédens, une liaison fondée sur le principe évident, qu'une chose ne peut pas commencer d'être, sans une cause qui la produise. Cet axiome connu sous le nom de *principe de la raison suffisante*, s'étend aux actions même les plus indifférentes. La volonté la plus libre ne peut sans un motif déterminant, leur donner naissance; car si toutes les circonstances de deux positions étant exactement les mêmes, elle agissait dans l'une et s'abstenait d'agir dans l'autre, son choix serait un effet sans cause : elle serait alors, dit Leibnitz, le hasard aveugle des épicuriens. L'opinion contraire est une illusion de l'esprit qui perdant de vue, les raisons fugitives du choix de la volonté dans les choses indifférentes, se persuade qu'elle s'est déterminée d'elle-même et sans motifs.

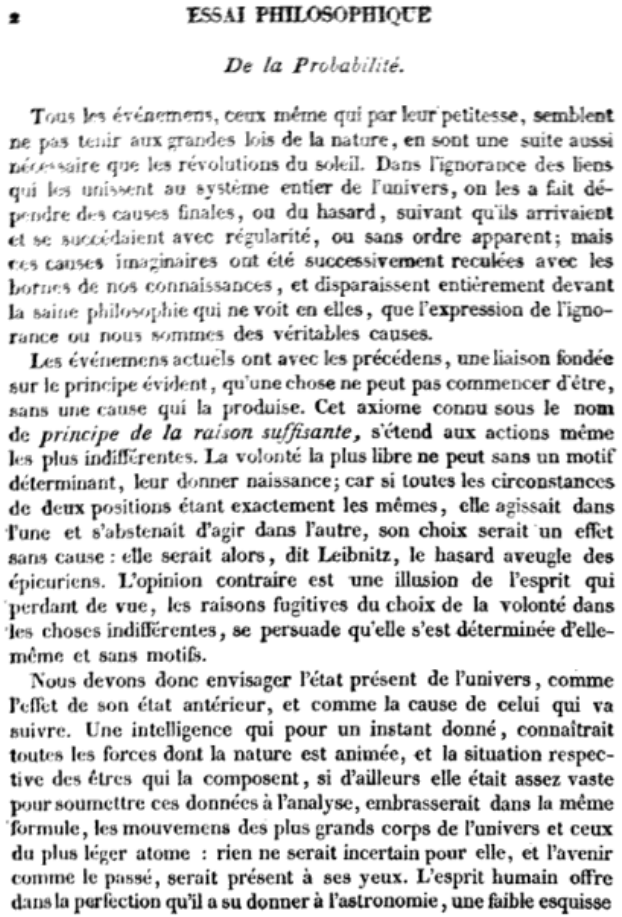
Nous devons donc envisager l'état présent de l'univers, comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule, les mouvemens des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux. L'esprit humain offre dans la perfection qu'il a su donner à l'astronomie, une faible esquisse

Laplace's Demon

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."

A BIT OF PHILOSOPHY

Essai philosophique sur les probabilités



2 ESSAI PHILOSOPHIQUE

De la Probabilité.

Tous les événemens, ceux même qui par leur petitesse, semblent ne pas tenir aux grandes lois de la nature, en sont une suite aussi nécessaire que les révolutions du soleil. Dans l'ignorance des liens qui les unissent au système entier de l'univers, on les a fait dépendre des causes finales, ou du hasard, suivant qu'ils arrivaient et se succédaient avec régularité, ou sans ordre apparent; mais ces causes imaginaires ont été successivement reculées avec les bornes de nos connaissances, et disparaissent entièrement devant la saine philosophie qui ne voit en elles, que l'expression de l'ignorance où nous sommes des véritables causes.

Les événemens actuels ont avec les précédens, une liaison fondée sur le principe évident, qu'une chose ne peut pas commencer d'être, sans une cause qui la produise. Cet axiome connu sous le nom de *principe de la raison suffisante*, s'étend aux actions même les plus indifférentes. La volonté la plus libre ne peut sans un motif déterminant, leur donner naissance; car si toutes les circonstances de deux positions étant exactement les mêmes, elle agissait dans l'une et s'abstenait d'agir dans l'autre, son choix serait un effet sans cause : elle serait alors, dit Leibnitz, le hasard aveugle des épicuriens. L'opinion contraire est une illusion de l'esprit qui perdant de vue, les raisons fugitives du choix de la volonté dans les choses indifférentes, se persuade qu'elle s'est déterminée d'elle-même et sans motifs.

Nous devons donc envisager l'état présent de l'univers, comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule, les mouvemens des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux. L'esprit humain offre dans la perfection qu'il a su donner à l'astronomie, une faible esquisse

The model

Data

Computation

Prediction

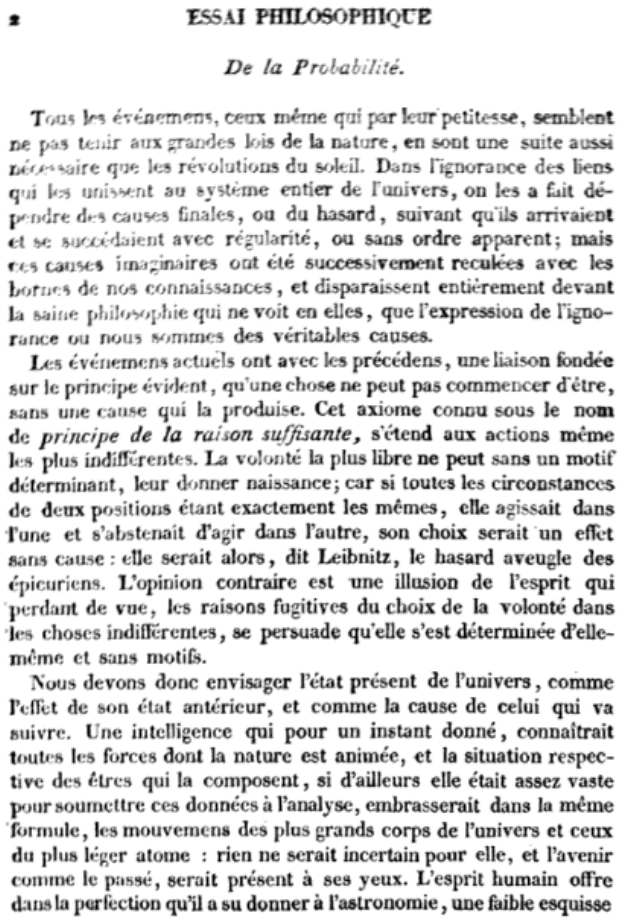
Laplace's Demon

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."

A BIT OF PHILOSOPHY

Neil Lawrence Awesome Introduction in Gaussian process summer school. https://gpss.cc/gpss18/

Essai philosophique sur les probabilités



2 ESSAI PHILOSOPHIQUE

De la Probabilité.

Tous les événemens, ceux même qui par leur petitesse, semblent ne pas tenir aux grandes lois de la nature, en sont une suite aussi nécessaire que les révolutions du soleil. Dans l'ignorance des liens qui les unissent au système entier de l'univers, on les a fait dépendre des causes finales, ou du hasard, suivant qu'ils arrivaient et se succédaient avec régularité, ou sans ordre apparent; mais ces causes imaginaires ont été successivement reculées avec les bornes de nos connaissances, et disparaissent entièrement devant la saine philosophie qui ne voit en elles, que l'expression de l'ignorance où nous sommes des véritables causes.

Les événemens actuels ont avec les précédens, une liaison fondée sur le principe évident, qu'une chose ne peut pas commencer d'être, sans une cause qui la produise. Cet axiome connu sous le nom de *principe de la raison suffisante*, s'étend aux actions même les plus indifférentes. La volonté la plus libre ne peut sans un motif déterminant, leur donner naissance; car si toutes les circonstances de deux positions étant exactement les mêmes, elle agissait dans l'une et s'abstenait d'agir dans l'autre, son choix serait un effet sans cause : elle serait alors, dit Leibnitz, le hasard aveugle des épicuriens. L'opinion contraire est une illusion de l'esprit qui perdant de vue, les raisons fugitives du choix de la volonté dans les choses indifférentes, se persuade qu'elle s'est déterminée d'elle-même et sans motifs.

Nous devons donc envisager l'état présent de l'univers, comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule, les mouvemens des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux. L'esprit humain offre dans la perfection qu'il a su donner à l'astronomie, une faible esquisse

→ **Laplace's Demon**

Markov property

The model

Data

Computation

Prediction

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."

The model + **Data** —**Computation**→ **Prediction**

A BIT OF PHILOSOPHY

Essai philosophique sur les probabilités

"The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits. The only difference between them is that which comes from our ignorance.

Probability is relative, in part to this ignorance, in part to our knowledge.

The model

Data

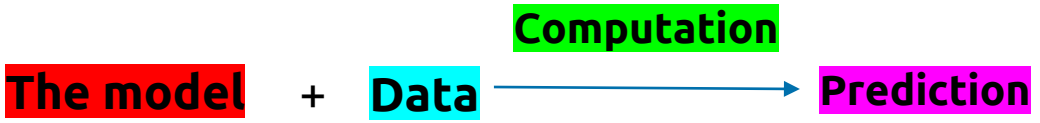
Computation

Prediction

Laplace's Demon

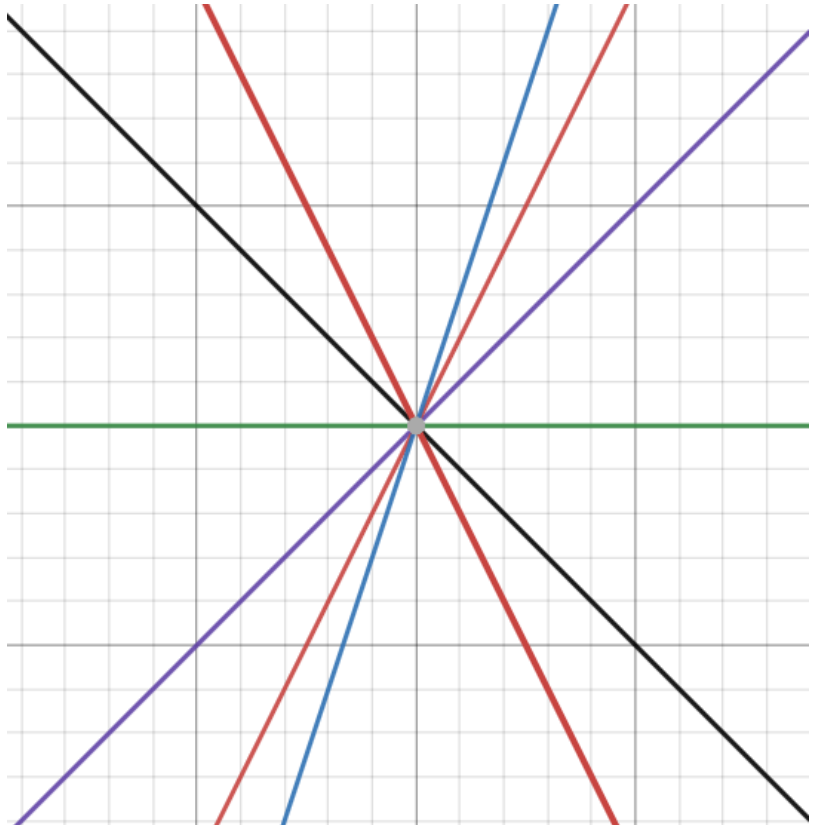
Markov property

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."



A BIT OF PHILOSOPHY

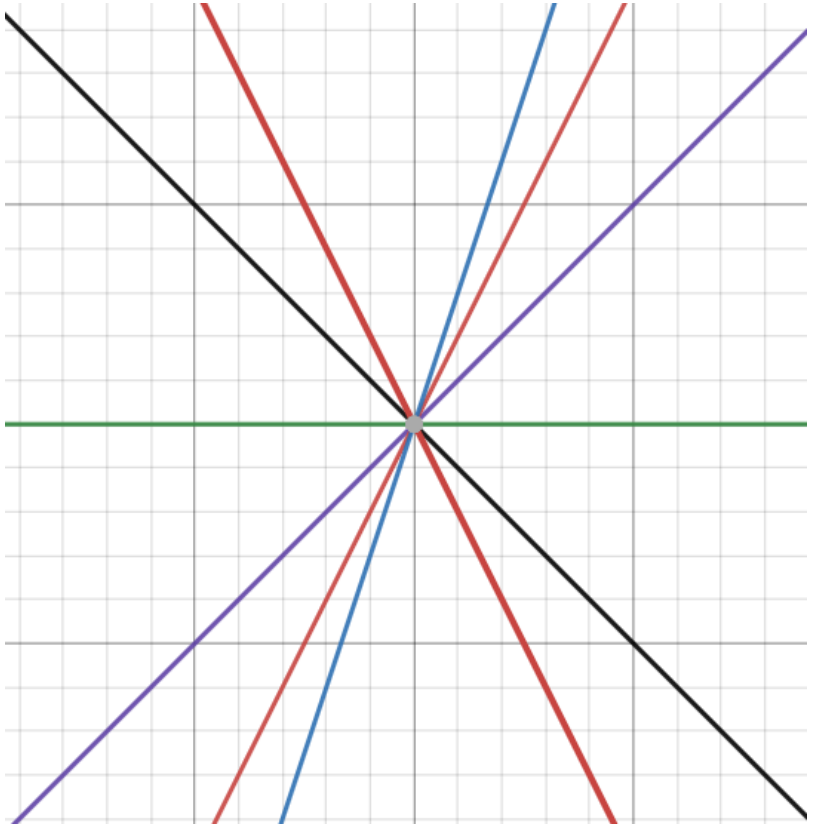
Fitting data: what about under-determined systems



$$y = mx + c$$

A BIT OF PHILOSOPHY

Fitting data: what about under-determined systems



- ***Aleatoric uncertainty*** captures noise inherent in the data
- ***Epistemic uncertainty*** captures model's lack of knowledge

Introduce probability distribution for parameter *c*

Bayesian Treatment

WHY CARE ABOUT UNCERTAINTY IN OUR MODELS?

IT'S ALL ABOUT TRUST

WHY CARE ABOUT UNCERTAINTY IN OUR MODELS?

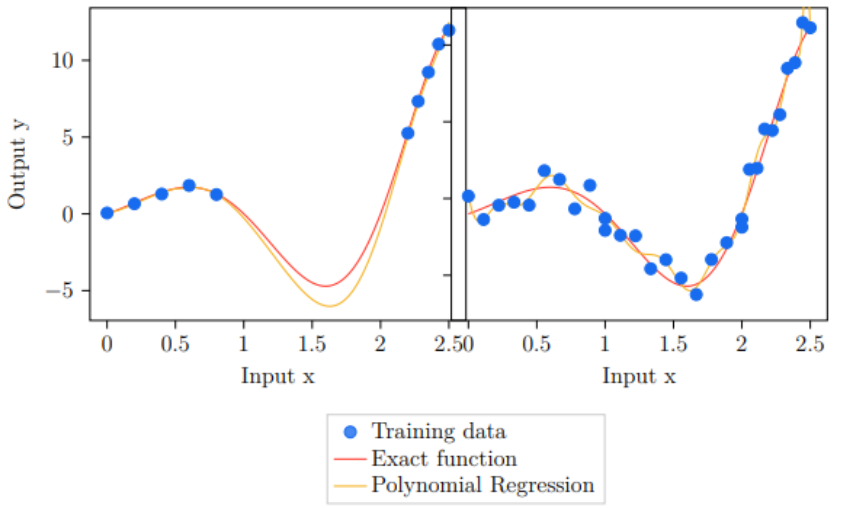
Quantify the trust in our model,

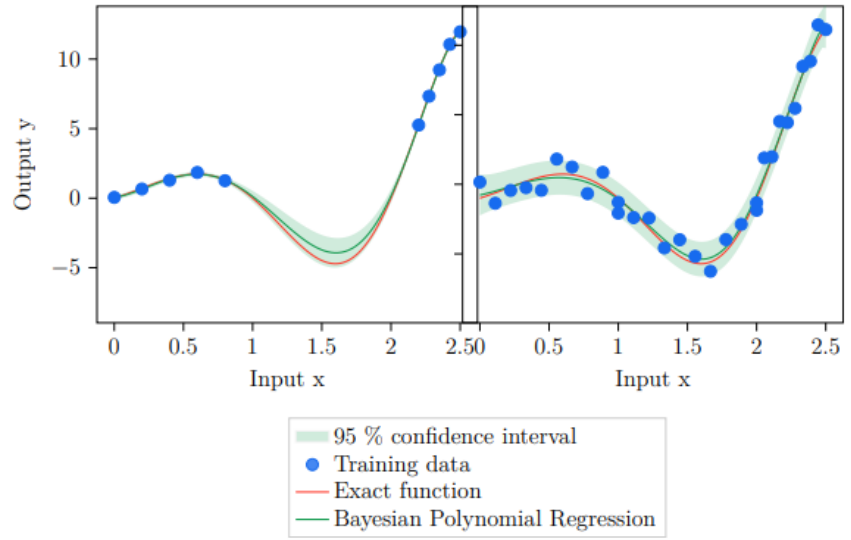
A natural Occam's razor effect

A natural paradigm for out-of-domain data detection

A way to diagnostic the behavior of black-box models

Exploration criteria for design of experiments





WHY CARE ABOUT UNCERTAINTY IN OUR MODELS?

Quantify the trust in our model,

A natural Occam's razor effect

A natural paradigm for out-of-domain data detection

A way to diagnostic the behavior of black-box models

Exploration criteria for design of experiments

(a) Confident Prediction (b) High data uncertainty (c) Out-of-distribution

Malinin, Andrey, and Mark Gales. "Predictive uncertainty estimation via prior networks." *Advances in neural information processing systems* 31 (2018).

WHY CARE ABOUT UNCERTAINTY IN OUR MODELS?

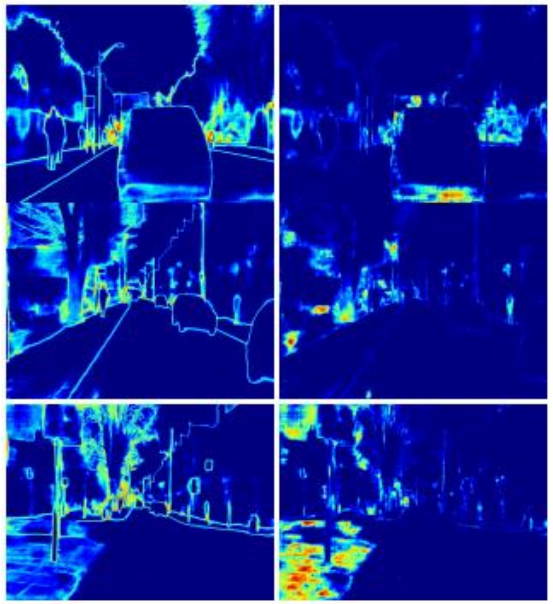
Quantify the trust in our model,

A natural Occam's razor effect

A natural paradigm for out-of-domain data detection

A way to diagnostic the behavior of black-box models

Exploration criteria for design of experiments



(d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Kendall, Alex, and Yarin Gal. "What uncertainties do we need in Bayesian deep learning for computer vision?." *Advances in neural information processing systems* 30 (2017).

WHY CARE ABOUT UNCERTAINTY IN OUR MODELS?

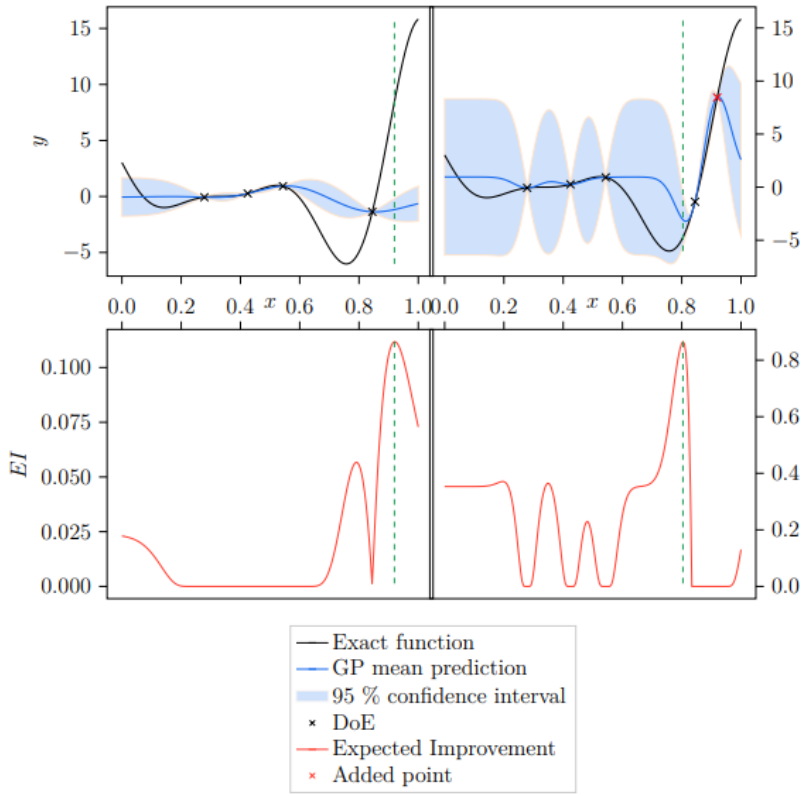
Quantify the trust in our model,

A natural Occam's razor effect

A natural paradigm for out-of-domain data detection

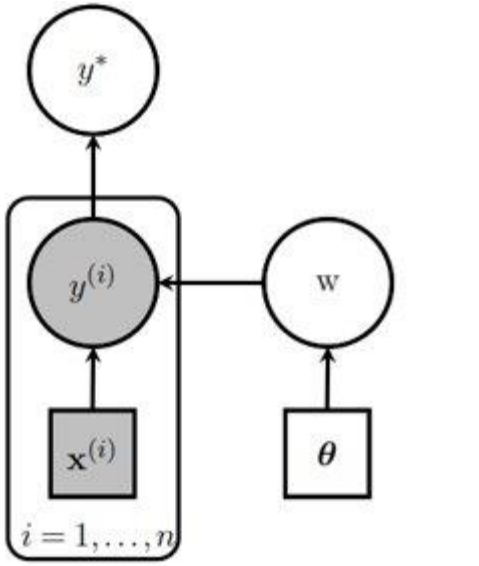
A way to diagnostic the behavior of black-box models

Exploration criteria for design of experiments



PROBABILITY THEORY IS THE LANGUAGE OF UNCERTAINTY

"Do your data analysis but remember, to make sense, you must never forget the rules of coherent behavior (Bayesian rules), any more than an engineer can forget Newton's laws." Lindley, D.V. (1975)



Empirical Bayes graph representation

Bayes rule is the bread-and-butter of Bayesian modeling. In the inference step, the posterior is inferred using Bayes rule.

Likelihood

Posterior $p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$ Prior / Evidence

$$= \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}}$$

APPROXIMATE BAYESIAN INFERENCE

Approach	Concept	Advantages	Drawbacks
MAP estimate	Mode of the posterior	Easy to compute	Not Bayesian, pathologies of the mode
Laplace Approximation	Gaussian approximation around the MAP	for $n \to \infty$ posterior $\to$ Gaussian	Computation of the Hessian, scarce data case, pathologies of the mode
Variational Inference	Minimization of the reverse KL	Flexible, ELBO, different variants	Mean-field approximations, under-estimate the variance, log expectation term
Expectation Propagation	Minimization of the direct KL	Highly parallelizable, the exponential family, Fast to converge	Multi-modal posterior, scarce data case, high-dimensionnal problems
Monte-Carlo Markov-Chain	Sampling through a defined Markov chain	Easy to implement, Adaptation to problems	Computationally intensive, stopping criteria

DEEP LEARNING IS COOL BUT …

How to trust the model?

How can we interpret the model?

Lacks solid mathematical background

Relies on big data

BAYESIAN DEEP LEARNING

A BAYESIAN PERSPECTIVE TO DEMYSTIFY DEEP LEARNING

BAYESIAN DEEP LEARNING MODELS

Bayesian neural networks and dropout

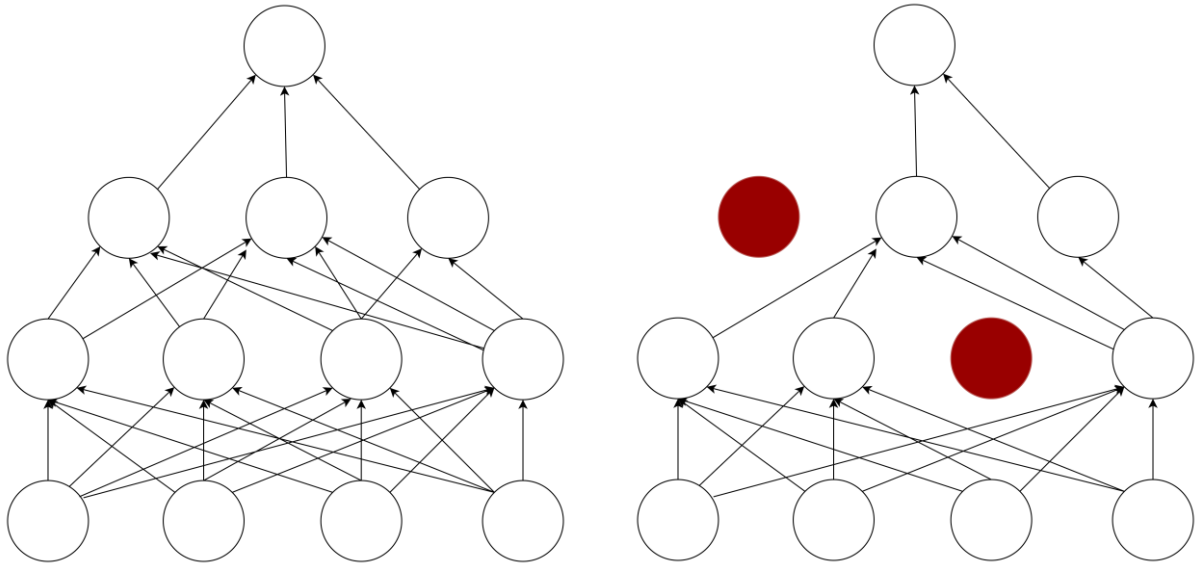
Gaussian processes,

Deep Gaussian processes

Variational auto-encoder

BAYESIAN NEURAL NETWORKS AND DROPOUT

Dropout



Standard neural net

Neural net with dropout

What it does:

Works by setting units to zero given some probability *p*.

Circumvent over-fitting.

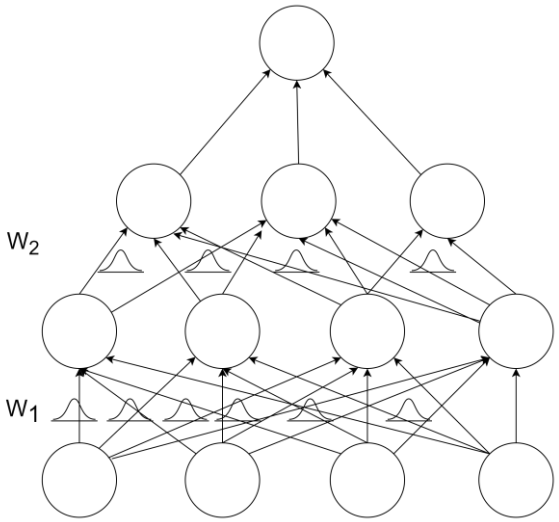
Improves generalization performance of the network

Why does it work:

?

BAYESIAN NEURAL NETWORKS AND DROPOUT

Bayesian neural network



Inference is intractable:

Given a Gaussian prior over each weight parameter:

$$p(\mathbf{w}) \sim \mathcal{N}(0, I)$$

The inference of the posterior distribution is analytically not tractable.

Variational inference:

Minimize the reverse KL divergence

$$\hat{\boldsymbol{\theta}}_q = \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}}\ \mathbb{KL}(q_{\boldsymbol{\theta}_q} || \tilde{p})$$

$$= \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}} \int_{\mathbf{w}} q_{\boldsymbol{\theta}_q}(\mathbf{w}) \log \frac{q_{\boldsymbol{\theta}_q}(\mathbf{w})}{\tilde{p}(\mathbf{w})} \mathrm{d}\mathbf{w}$$

$$\hat{\boldsymbol{\theta}}_q = \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p(\mathbf{y}, \mathbf{w})} \mathrm{d}\mathbf{w} + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{y}|X) d\mathbf{w}$$

$$= \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}} \int_{\mathbf{w}} -q(\mathbf{w}) \log p(\mathbf{y}|\mathbf{w}) \mathrm{d}\mathbf{w} + \mathbb{KL}[q||p] + C$$

$$= \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}}\ \mathbb{E}_q[-\log p(\mathbf{y}|\mathbf{w})] + \mathbb{KL}[q||p]$$

BAYESIAN NEURAL NETWORKS AND DROPOUT

Specification of the variational distribution:

Define:

$$q_M(W_i) = M\text{diag}(\text{Bernouli}(p_i))$$

With variational parameters M

With this variational distribution we have exactly the objective function of a neural network with dropout.

Variational inference:

Minimize the reverse KL divergence

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_q &= \underset{\boldsymbol{\theta}_q}{\text{argmin}}\ \mathbb{KL}(q_{\boldsymbol{\theta}_q}||\tilde{p}) \\
&= \underset{\boldsymbol{\theta}_q}{\text{argmin}}\ \int_{\mathbf{w}} q_{\boldsymbol{\theta}_q}(\mathbf{w}) \log \frac{q_{\boldsymbol{\theta}_q}(\mathbf{w})}{\tilde{p}(\mathbf{w})} \mathrm{d}\mathbf{w} \\
\hat{\boldsymbol{\theta}}_q &= \underset{\boldsymbol{\theta}_q}{\text{argmin}} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p(\mathbf{y}, \mathbf{w})} \mathrm{d}\mathbf{w} + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{y}|X) d\mathbf{w} \\
&= \underset{\boldsymbol{\theta}_q}{\text{argmin}} \int_{\mathbf{w}} -q(\mathbf{w}) \log p(\mathbf{y}|\mathbf{w}) \mathrm{d}\mathbf{w} + \mathbb{KL}[q||p] + C \\
&= \underset{\boldsymbol{\theta}_q}{\text{argmin}}\ \mathbb{E}_q\left[-\log p(\mathbf{y}|\mathbf{w})\right] + \mathbb{KL}[q||p]
\end{aligned}$$

BAYESIAN NEURAL NETWORKS AND DROPOUT

Specification of the variational distribution:

Define:

$$q_M(W_i) = M\,\mathrm{diag}(\mathrm{Bernouli}(p_i))$$

With variational parameters M

With this variational distribution we have exactly the objective function of a neural network with dropout.

Bayesian interpretation --> predictive distribution function

Variational inference:

Minimize the reverse KL divergence

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_q &= \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}}\ \mathbb{KL}(q_{\boldsymbol{\theta}_q}||\tilde{p}) \\
&= \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}}\ \int_{\mathbf{w}} q_{\boldsymbol{\theta}_q}(\mathbf{w}) \log \frac{q_{\boldsymbol{\theta}_q}(\mathbf{w})}{\tilde{p}(\mathbf{w})} \mathrm{d}\mathbf{w} \\
\hat{\boldsymbol{\theta}}_q &= \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p(\mathbf{y},\mathbf{w})} \mathrm{d}\mathbf{w} + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{y}|X) d\mathbf{w} \\
&= \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}} \int_{\mathbf{w}} -q(\mathbf{w}) \log p(\mathbf{y}|\mathbf{w}) \mathrm{d}\mathbf{w} + \mathbb{KL}[q||p] + C \\
&= \underset{\boldsymbol{\theta}_q}{\operatorname{argmin}}\ \mathbb{E}_q\left[-\log p(\mathbf{y}|\mathbf{w})\right] + \mathbb{KL}[q||p]
\end{aligned}$$

BAYESIAN NEURAL NETWORKS AND DROPOUT

Specification of the variational distribution:

Define:

$$q_M(W_i) = M\,\mathrm{diag}(\mathrm{Bernouli}(p_i))$$

With variational parameters M

With this variational distribution we have exactly the objective function of a neural network with dropout.

Bayesian interpretation --> predictive distribution function

MC Dropout:

First moment for mean prediction:

$$\mathbb{E}(\mathbf{y}^*) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t$$

Second moment for uncertainty:

$$\mathrm{Var}(\mathbf{y}^*) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t^T\hat{\mathbf{y}}_t - \mathbb{E}(\mathbf{y}^*)^T\mathbb{E}(\mathbf{y}^*) + \tau^{-1}\mathbf{I}$$

Where:

$$\hat{\mathbf{y}}_t \sim \mathrm{DropoutNetwork}(\mathbf{x}^*)$$

Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." international conference on machine learning. PMLR, 2016.

BAYESIAN NEURAL NETWORKS AND DROPOUT

- A Bayesian perspective allows us to :
  - explain why a certain model works well
  - obtain well-calibrated uncertainty (MCdropout)

GAUSSIAN PROCESSES

Definition

Gaussian process [Rasmussen, 2004]

A Gaussian process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function $\mu(.)$ and covariance function $k^{\Theta}(.)$ (Kernel) :
$f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

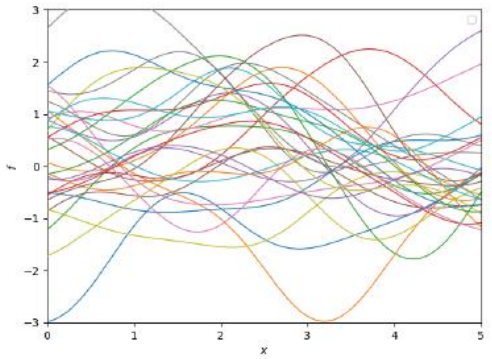
GAUSSIAN PROCESSES

Definition

Gaussian process [Rasmussen, 2004]

A Gaussian process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function $\mu(.)$ and covariance function $k^{\Theta}(.)$ (Kernel) :
$f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

GP with a squared exponential kernel **prior** samples



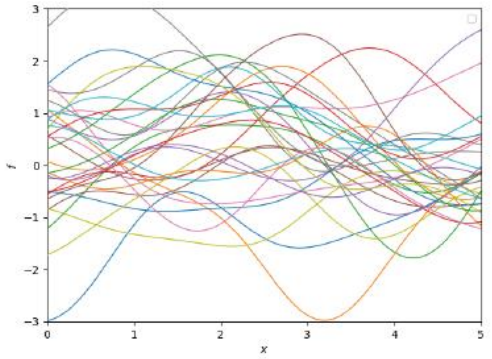
GAUSSIAN PROCESSES

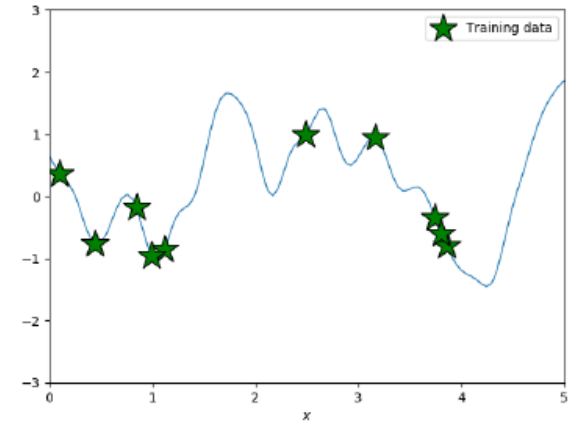
Definition

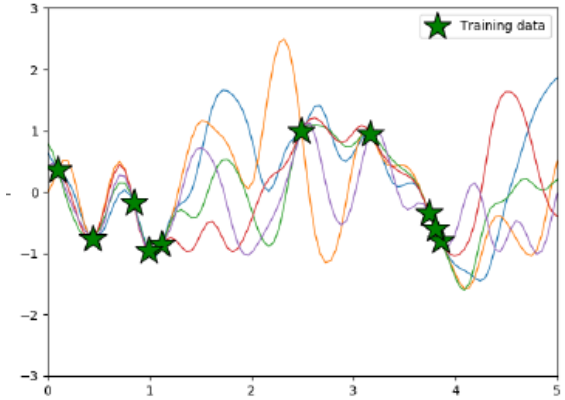
Gaussian process [Rasmussen, 2004]

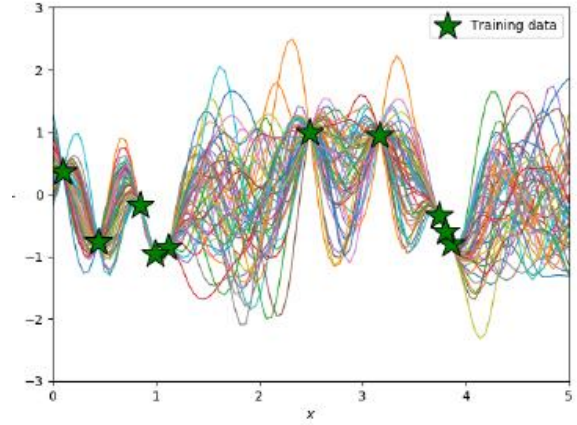
A Gaussian process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function $\mu(.)$ and covariance function $k^{\Theta}(.)$ (Kernel) :
$f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

GP with a squared exponential kernel **prior** samples









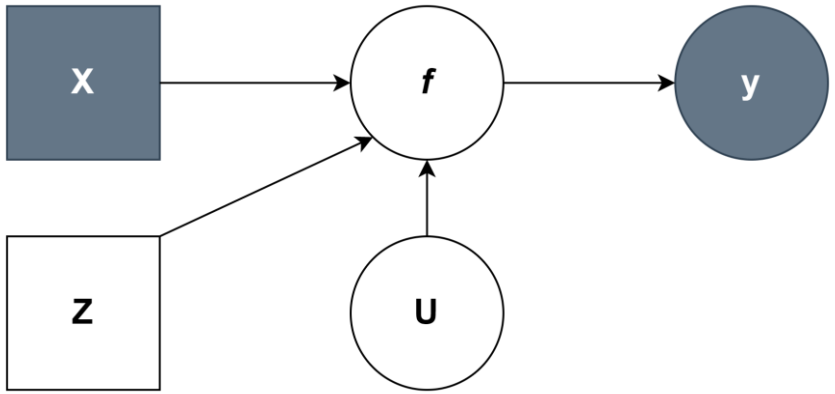
SPARSE GAUSSIAN PROCESSES

Gaussian process training and prediction complexity :

$$\mathcal{O}(n^3)$$

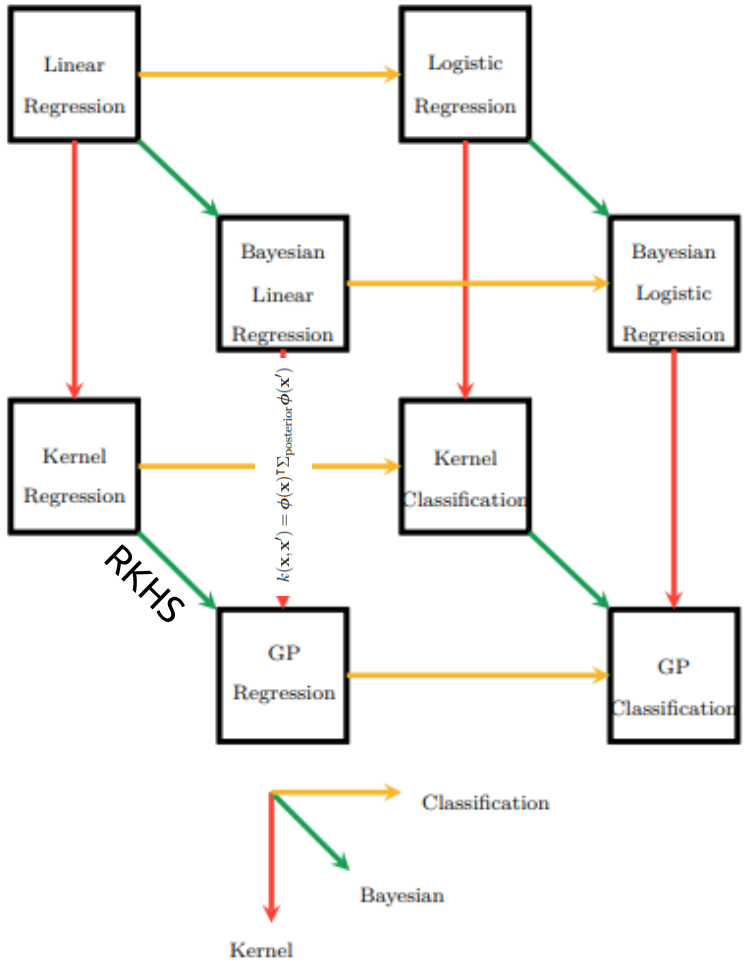
Sparse Gaussian processes introduce latent variables to obtain a low rank approximation of the covariance matrix:

$$\mathcal{O}(nm^2)$$



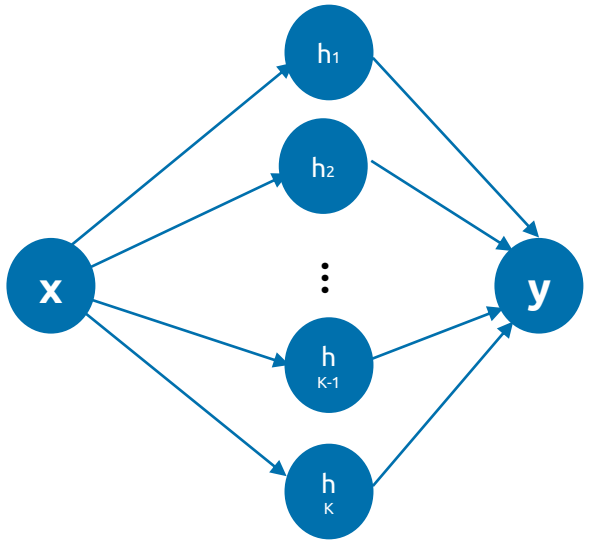
FROM LINEAR REGRESSION TO GAUSSIAN PROCESSES

The Ghahramani cube

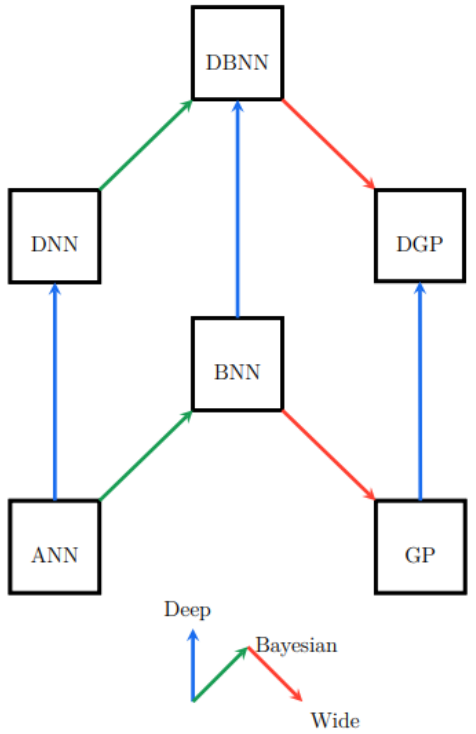


FROM NEURAL NETWORKS TO GAUSSIAN PROCESSES

$$f(\mathbf{x}) = \frac{1}{K}\mathbf{w}^{\mathsf{T}}\mathbf{h}(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K} w_i h_i(\mathbf{x}).$$



The GP/NN spectrum



GAUSSIAN PROCESSES

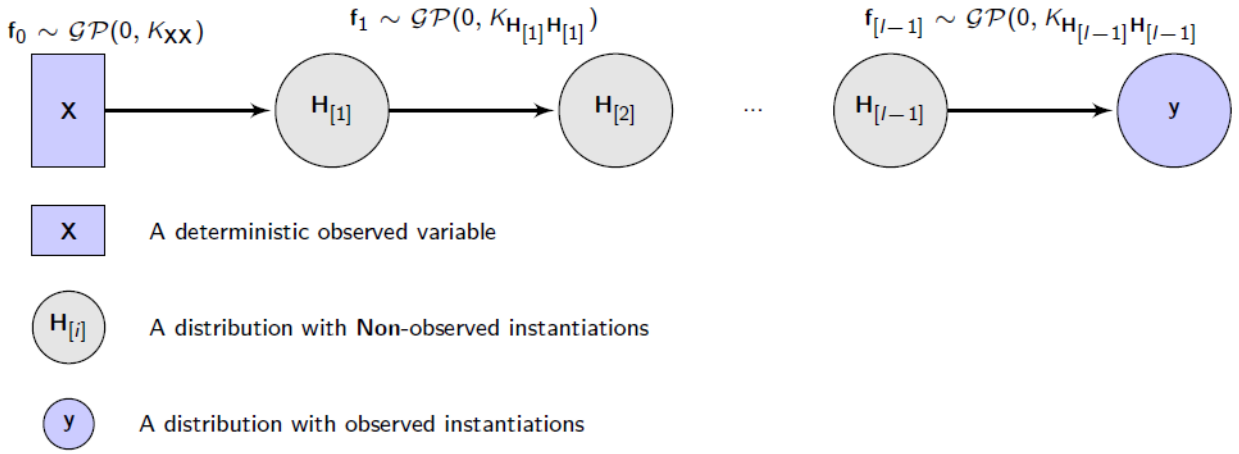
- Bayesian model can be well suited for big data (sparse GP)
- Non-parametric Bayesian models have an increased power of representation (asymptotic convergence of NN to GP)

DEEP GAUSSIAN PROCESSES

Definition

Deep Gaussian processes are a hierarchical generalization of Gaussian processes.

It considers the statistical relationship between the inputs and the response as a functional composition of Gaussian processes.

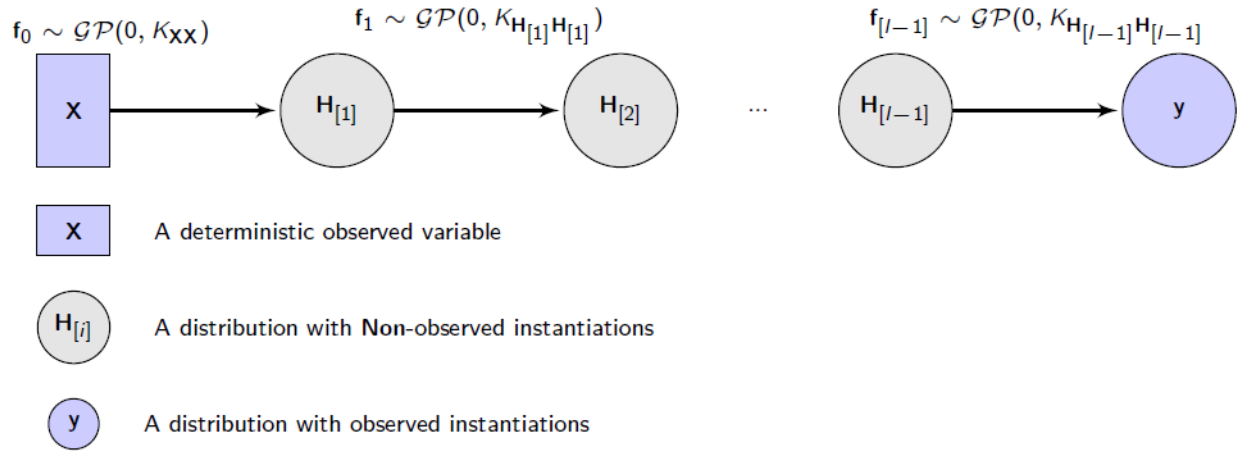


DEEP GAUSSIAN PROCESSES

Definition

Deep Gaussian processes are a hierarchical generalization of Gaussian processes.

It considers the statistical relationship between the inputs and the response as a functional composition of Gaussian processes.



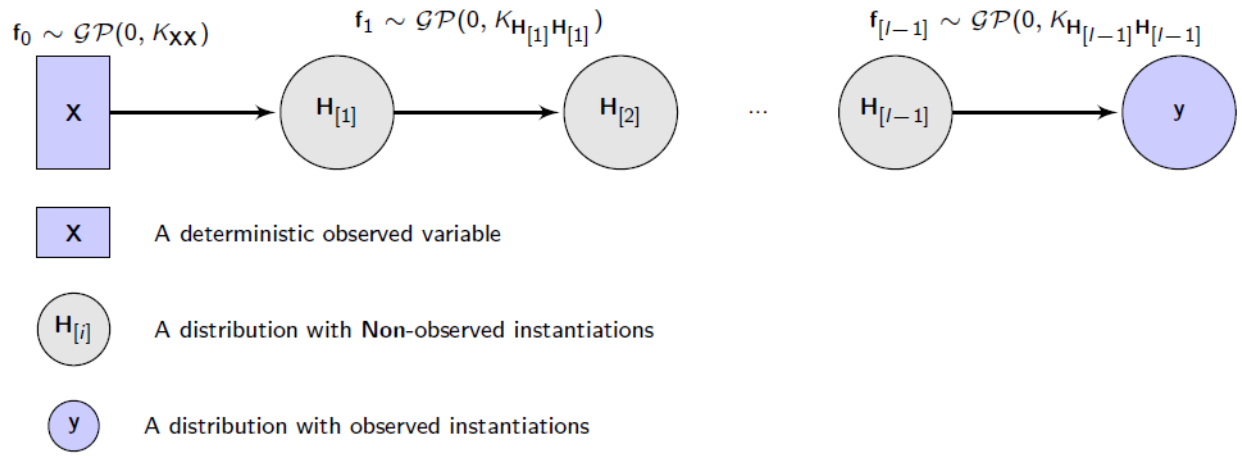
$$p(\mathbf{y}|\mathbf{X}) = \int_{\{H_{[i]}\}_1^{l-1}} p(\mathbf{y}|\mathbf{H}_{[l-1]}) \dots p(\mathbf{H}_{[1]}|\mathbf{X}) \mathrm{d}\{\mathbf{H}_{[i]}\}_1^{l-1}$$

Analytically Intractable

DEEP GAUSSIAN PROCESSES

Deep Gaussian process inference

Approach	Inference approach	Approximation approach
[Damianou and Lawrence, 2013]	Variational inference	Variational sparse GPs
[Dai et al., 2015]	Variational inference	Variational sparse GPs
[Salimbeni and Deisenroth, 2017]	Variational inference	Variational sparse GPs
[Haibin et al., 2019]	Variational inference	Variational sparse GPs
[Cutajar et al., 2017]	Variational inference	Random feature-based GP
[Bui et al., 2016]	Expectation propagation	Fully independent training conditional GPs
[Havasi et al., 2018]	Markov-Chain Monte-Carlo	Variational sparse GPs
[Rossi et al., 2020]	Markov-Chain Monte-Carlo	Variational sparse GPs

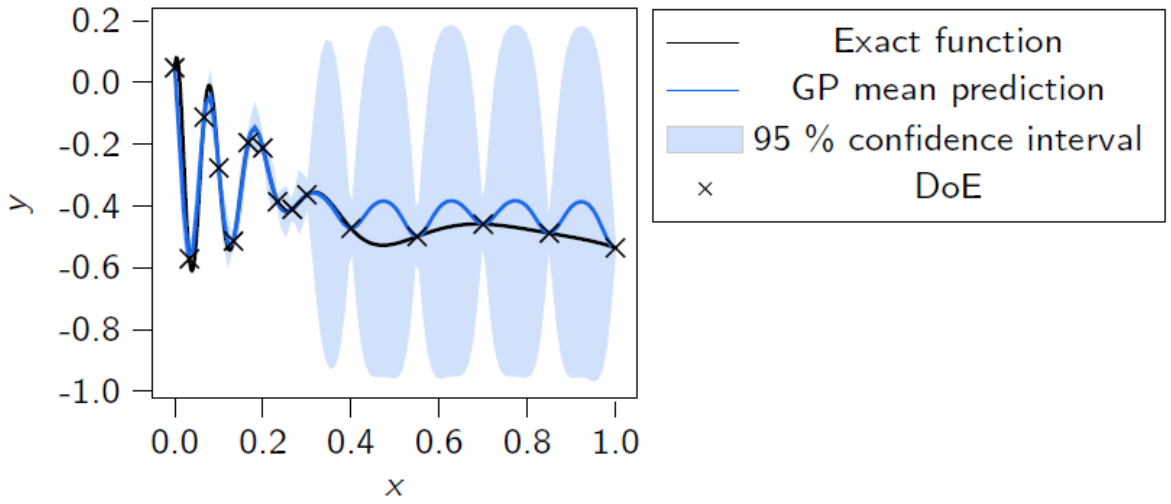


$$p(\mathbf{y}|\mathbf{X}) = \int_{\{H_{[i]}\}_1^{l-1}} p(\mathbf{y}|\mathbf{H}_{[l-1]}) \dots p(\mathbf{H}_{[1]}|\mathbf{X}) \mathrm{d}\{\mathbf{H}_{[i]}\}_1^{l-1}$$

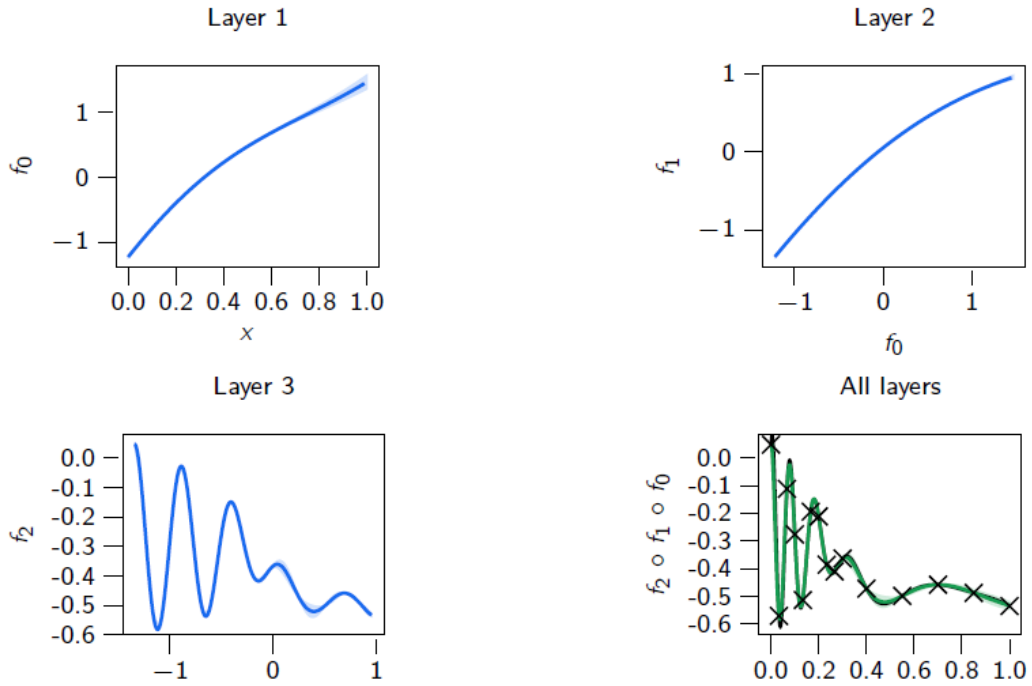
Analytically Intractable

DEEP GAUSSIAN PROCESSES

A Gaussian process prediction



A deep Gaussian process prediction

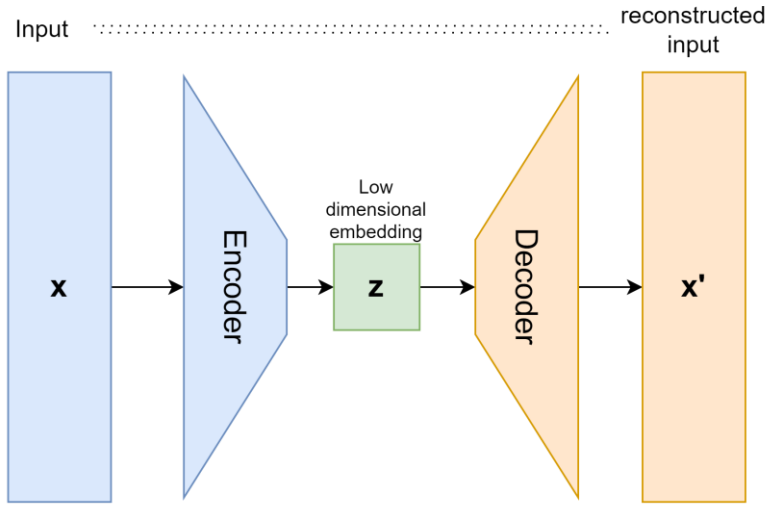


DEEP GAUSSIAN PROCESSES

- A non-parametric equivalent to deep neural network
- The hierarchical generalization increases the power of representation

VARIATIONAL AUTO-ENCODER

Auto-encoder



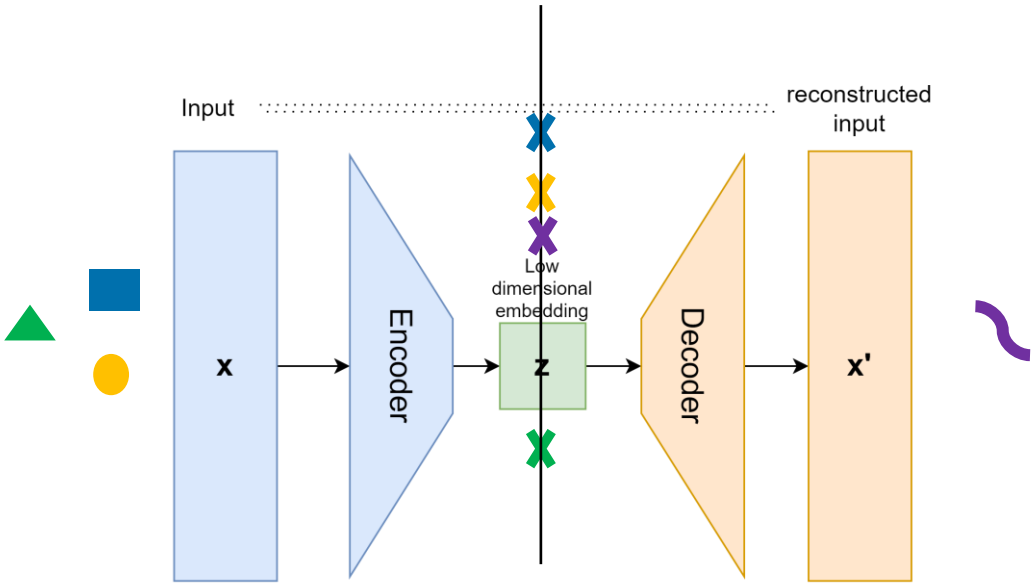
$$L_{\mathrm{AE}}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - f_\theta(g_\phi(\mathbf{x}^{(i)})))^2$$

Limitations

- Regularity of the latent space
- Over-fitting

VARIATIONAL AUTO-ENCODER

Auto-encoder



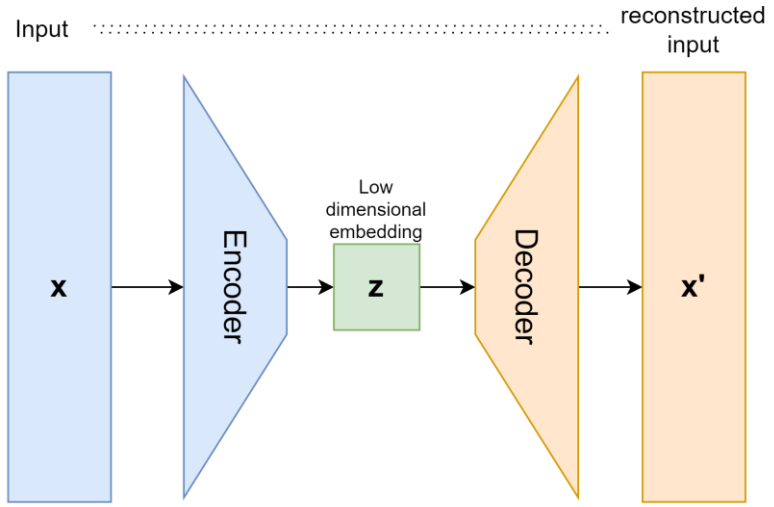
$$L_{\mathrm{AE}}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - f_\theta(g_\phi(\mathbf{x}^{(i)})))^2$$

Limitations

- Regularity of the latent space
- Over-fitting

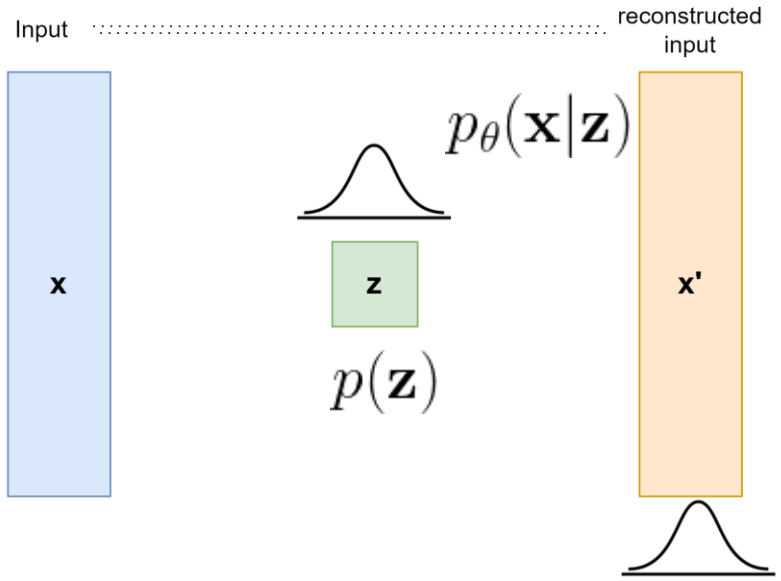
VARIATIONAL AUTO-ENCODER

Auto-encoder



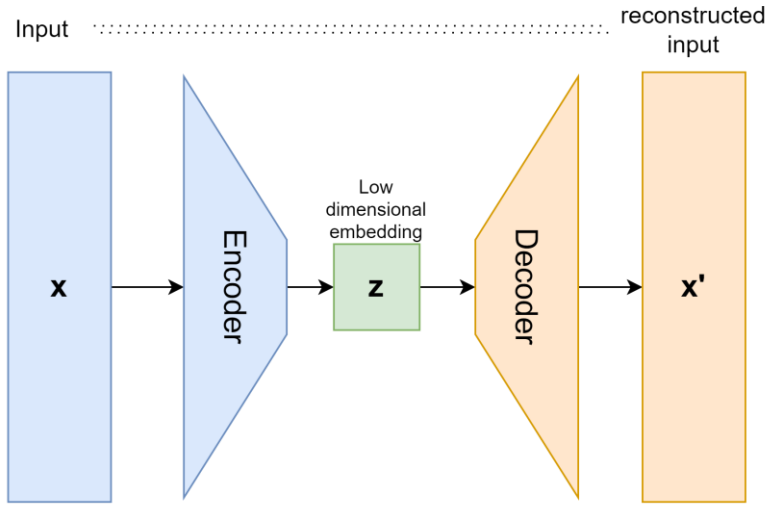
$$L_{\mathrm{AE}}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - f_\theta(g_\phi(\mathbf{x}^{(i)})))^2$$

Variational auto-encoder



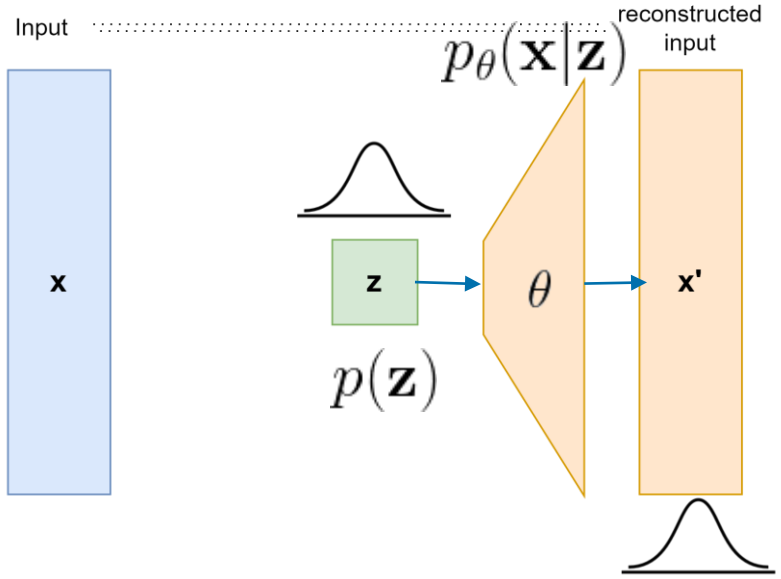
VARIATIONAL AUTO-ENCODER

Auto-encoder



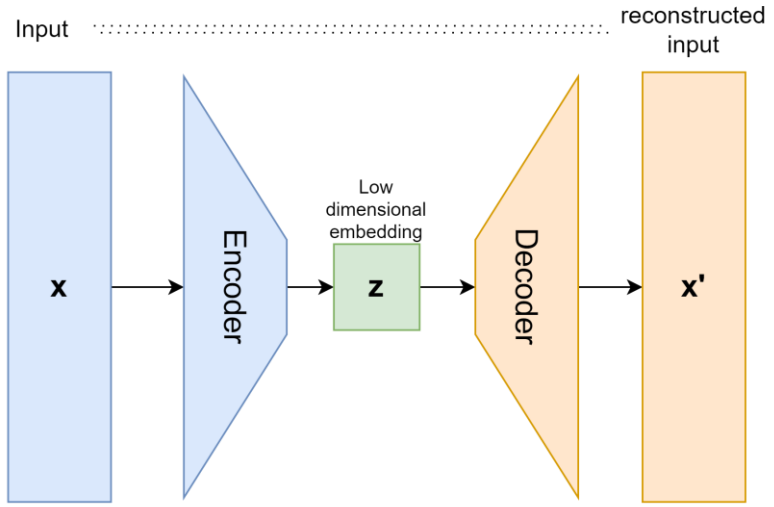
$$L_{\mathrm{AE}}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - f_\theta(g_\phi(\mathbf{x}^{(i)})))^2$$

Variational auto-encoder



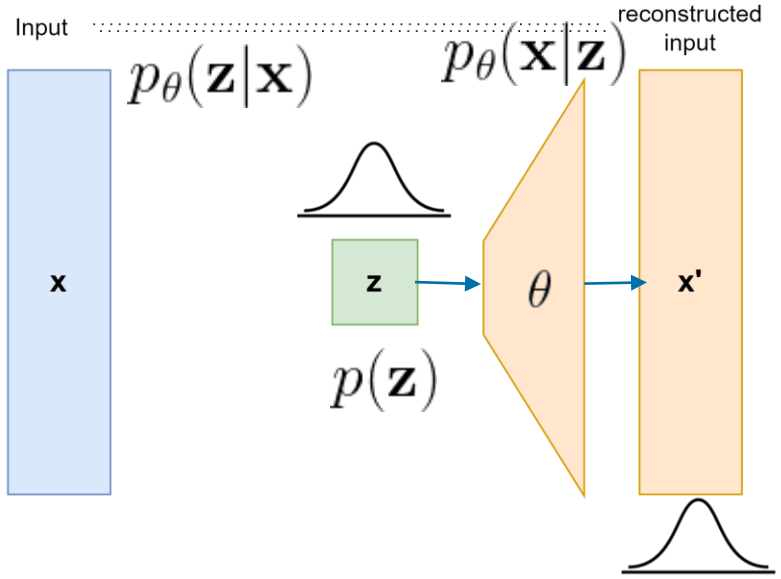
VARIATIONAL AUTO-ENCODER

Auto-encoder



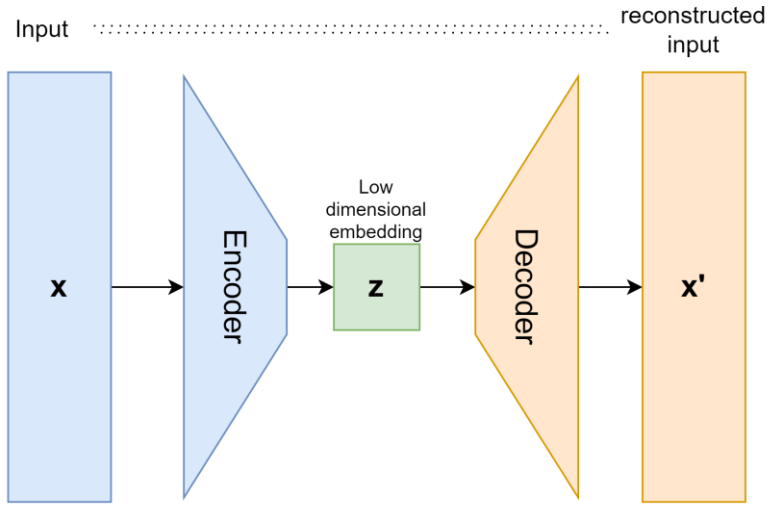
$$L_{\mathrm{AE}}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - f_\theta(g_\phi(\mathbf{x}^{(i)})))^2$$

Variational auto-encoder



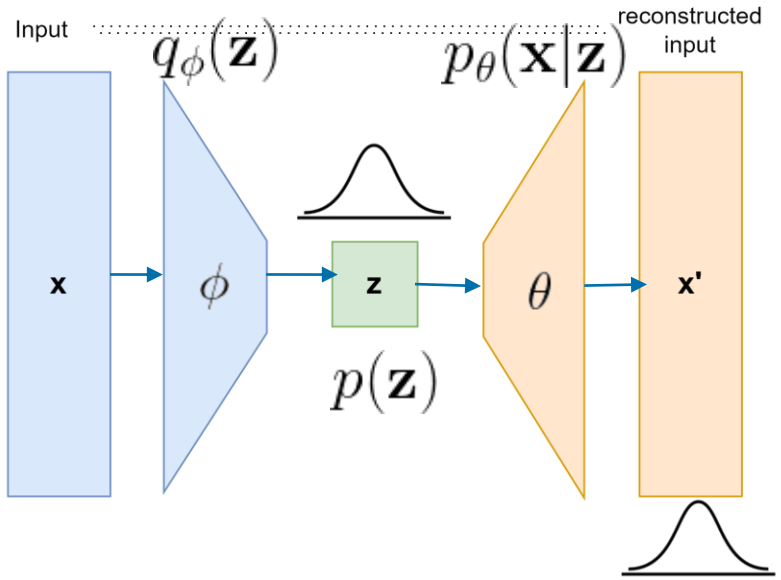
VARIATIONAL AUTO-ENCODER

Auto-encoder



$$L_{\mathrm{AE}}(\theta,\phi) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - f_\theta(g_\phi(\mathbf{x}^{(i)})))^2$$

Variational auto-encoder



$$L_{VAE}(\theta,\phi) = \mathbb{E}_{q_\phi}(\log(p(\mathbf{x}|\mathbf{z}))) - \mathbb{KL}(q_\phi(\mathbf{z}), p(\mathbf{z}))$$

VARIATIONAL AUTO-ENCODER

- Encoding and decoding are Bayesian transformation
- A Bayesian perspective induce a regularized latent space

CONCLUSIONS

THE 21ST CENTURY IS INDEED THE BAYESIAN CENTURY

SUMMARY OF THE REVIEWED MODELS

- **A Bayesian perspective allows us to :**
  - Explain the mechanism behind a certain model
  - Obtain well-calibrated uncertainty
  - Increase the power of representation
  - Regularize the predictive space and the latent space

CURRENT LIMITATIONS

- Prior selection

- The optimization of the ELBO for variational inference (non-Euclidian space)

- MCMC approaches still computationally intensive

CURRENT LIMITATIONS

- Prior selection
- The optimization of the ELBO for variational inference (non-Euclidian space)
- MCMC approaches still computationally intensive

The best is yet to come …

- An increasing interest by the ML research community
- Research on more efficient Bayesian approximations,
- More combinations of Bayesian and deep learning,
- Mysterious results in deep learning are resolved by thinking about model construction and generalization from a probabilistic perspective.

THAT'S ALL FOLKS

BAYESIAN IMPLEMENTATIONS

- **Bayesian inference:**
  - *PyMC3*
- ***Gaussian Processes:***
  - *GPflow/GPtorch*
- ***Deep Gaussian processes:***
  - *GPFlux*