Bay(e)sics for Deep Learning and Bayesian Neural Nets

Elena Sellentin
Assistant Prof & Oort Fellow

Universiteit Leiden

The appeal of deep learning

'It is so complicated, I am going to use a neural net.'

Want:

Controls ('inputs') $\boldsymbol{x}$ mapped to target variables ('outputs') $\boldsymbol{y}$.

Sadly:

$\boldsymbol{y} = g(h(f(\boldsymbol{x})))$ extremely complicated/unknown/probably not injective/slow/or not interesting.

Thankfully:

The Universal Approximation Theorem:

For every function f, there exists a (convolutional) neural net architecture A which can approximate f to arbitrary accuracy[1]. The inverse is also true: For every fixed A, there exists a function f which A cannot approximate accurately.

[1] In the sense of $|A(\boldsymbol{x}) - f(\boldsymbol{x})| < \epsilon \; \forall \boldsymbol{x}$

Black magic or maths?

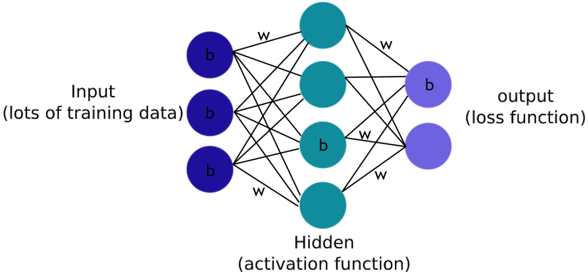
The Univ. App. Theorem is intuitively quickly understood:

1. Want to compose an arbitrary function out of simple sub-functions f_l.
2. Ergo: must define connections between the functions f_l, such that $f_{l_1} \otimes f_{l_2} \otimes ... \otimes f_{l_n}$ does not leave the realm of functions I am interested in.[2]
3. $\Rightarrow$ need to pick mathematical operation '$\otimes$' cleverly.
4. $\Rightarrow$ Pick: $\otimes = (+, *)$ (addition; and convolution = multiplication).
5. Those have inverse operations/elements.

→ Can now patch together any function. 4 and 5 directly imply there will exist many networks who perform the same task (insert identities).

[2] Q: Physicists, what does this maths remind you of?

Maths in pictures



Q: How many free parameters does this NN introduce?
How many training samples will you minimally need?

Q: How many free parameters did Planck or Gaia use?

Q: How many nuisance parameters will Euclid introduce?

Mini Review Neural Nets

Pairs (x_1, y_1), (x_2, y_2)... (x_n, y_n).

Feed-forward network, initial activation is the (observable) input variables:

$$a_j^l = A(w_{ji}^l a_i^{l-1} + b_j^l), \quad a_i^0 = x_i \tag{1}$$

Compare (observable) outputs with a generic[3] loss function:

$$L = \sum_i |a_i^R - y_i|^2 \tag{2}$$

Train:

$$w_{mn}^l(e+1) = w_{mn}^l(e) + r\frac{\partial L}{\partial w_{mn}^l(e)}, \qquad b_n^l(e+1) = b_n^l(e) + r\frac{\partial L}{\partial b_n^l(e)}$$

Compose functions from addition and multiplication $\Rightarrow$ sequential layers + convolution $\Rightarrow$ Universal Approximation Theorem.

[3] As a Bayesian, you'll often define your own loss function (teacher function).

Frequent questions

1. Can I optimize the procuration of training data?
2. Can I shrink my NN without losing accuracy?
3. Can I even make my NN convex?
4. How do I put an NN into a physics problem?
5. What has my net learned?
6. How can I achieve a reliable an uncertainty quantification?

→ All of these questions are answered by (Bayesian) thinking.

To set the mood

'With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.' (John von Neumann)

'In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.' (R. T. Rockafellar)[4]

Engineering 1-0-1:
Dissect complex problems into manageable bits you can understand!

[4] R. T. R.: Lagrange multipliers and optimality. SIAM Review, 1993.

Mini Review Bayesian statistics

The Sleeping Beauty Problem: Waking up, which non-trivial events can you reconstruct about your past?

1. Irrelevant information, parameters and data factor out.
2. State a prior π; learning and inference *require* priors.
3. Observe data $\boldsymbol{x}$ passively. What is their sampling distrib $\mathcal{P}(\boldsymbol{x})$?
4. Embolden yourself to 'explain' the data: $\mathcal{P}(\boldsymbol{x}) \Rightarrow \mathcal{P}_M(\boldsymbol{x}|\boldsymbol{\theta})$.
5. Wonder: what can I say about $\boldsymbol{\theta}$? $\Rightarrow$ 'Inverse Problem'.
6. Maybe shouldn't have been so bold? $\Rightarrow$ assign 'credibility' to model M.

Q: By which step can you already *generate* fake data?

Sleeping beauty problems in astronomy

1. Given a stellar surface, what can we deduce about the stellar interior?
2. Given data from today, what can we say about the past/future of an astronomical system beyond our lifespan?
3. Given the CMB map, how old is the Universe?
4. Given the Kepler data, is there life elsewhere in the Universe?

Boolean Logic

Boolean algebra with true and false:

- (A and B) is true iff A true and B true.
- (A or B) is true if A true or B true or both true.
- Not-A is true if A is not true.

Bayesian logic (inference logic)

Replace 'true' and 'false' by 'possibly' and 'possibly not'.

1. If A likely, and B frequently occurs when A occured, then B likely.
2. If B occured, and B frequent when A occured, then A likely to have occured, even though I cannot observe A.
3. If A likely no matter what, and B likely no matter what, then observing B teaches me nothing about A.
4. If B impossible when A occurs, then observing A tells me B definitely did not occur.
5. Iff A causes B and iff B causes C, then observing C means A occured.
6. If A causes B and if B causes C, then observing C means A possibly occured, but unlikely so.

Bayes Theorem

- Prior $\pi(\boldsymbol{\theta})$: penalty to pay for having introduced params.
- Likelihood $\mathcal{L}$: factors signal and noise apart, then fits.
- $\epsilon = \pi(\boldsymbol{x})$ is the evidence (doubts model M).

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x})} \tag{3}$$

Bayes Theorem is not akin to Bayesian Inference! Also not if you simply multiply flat priors and a Gaussian likelihood because 'that is what people do'.

Marginals and Conditionals

Marginals drop information

$$\mathcal{P}(x) = \int \mathcal{P}(x, y)\mathrm{d}y$$

Conditionals provide information

$$\mathcal{P}(x|y) \subset \mathcal{P}(x)$$

The aim of Bayesian inference it to put in as many conditional-marginals as possible:

$$\mathcal{P}(x) = \int \mathcal{P}(x, y)\mathrm{d}y = \int \mathcal{P}(x|y)\pi(y|I_y)\mathrm{d}y$$

The space of all possibilities is vast, but not if you logically condition on all information you have.

Inference vs Learning

Learning: learns the map between *two observables*, which may include simulations and/or penalties. ($\boldsymbol{x}_{\text{in}}\boldsymbol{y}_{\text{out}}$; loss function!)

Inference: may *include* learning but goes significantly beyond. Given *observables* it attempts to draw non-trivial conclusions on the *unobservable*, which might not even exist/be correct.

Credibility: Inference assigns a ‘credibility’ (probability) to on an elusive, not-observable ‘latent’ quantity.

Example I (inference)

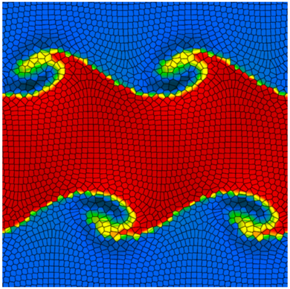
Who has ever seen a Higgs particle?

Example I (inference)

Who believes there exists a Higgs particle?

Example II (learning)

Solving a (beastly) system of differential equations



Volker Springel et al. (AREPO code)

$$\begin{aligned} \dot{\rho} + \vec{\nabla} \cdot (\rho \vec{v}) &= 0, \\ \dot{\vec{v}} + (\vec{v} \cdot \vec{\nabla})\vec{v} &= -\frac{\vec{\nabla} p}{\rho}, \\ \dot{\epsilon} + \vec{\nabla} \cdot (\epsilon \vec{v}) &= -p \vec{\nabla} \cdot \vec{v}. \end{aligned} \tag{4}$$

Initial conditions $\Rightarrow$ later state (learnable; not an inference problem)

Example III (learning and inference)

Image Credit: ESO/D. Minniti/VVV Team

Q: If you know nothing about physics/astronomy, how would you make one (or thousands of) fake stellar cluster like this one?

Example III (learning and inference)

Want: Formation history and initial conditions.

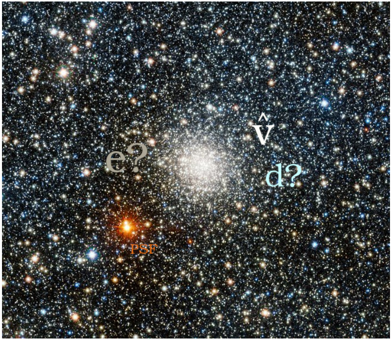


Image Credit: ESO/D. Minniti/VVV Team

Voilà: a combined inference & learning problem.

A few words on 'uncertainty'

UNCERTAINTY IS MATHEMATICALLY A CONSERVED QUANTITY.

Once you have it, you don't get rid of it.

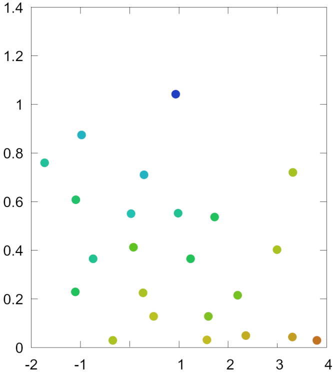
By being non-smart, you can *increase* it, by being smart you can preserve the minimal level.[5]

1. Engineering uncertainty: finite integration precision $\hat{f} = f \pm \epsilon$
2. Finite instrumental precision $\boldsymbol{x} \sim \mathcal{P}(\boldsymbol{x})$
3. Population variability (the *Universe* is random) $\boldsymbol{s} \sim \mathcal{P}(\boldsymbol{s})$
4. Scientific uncertainty: We really do not know sth, and want to deduce as much as possible about it: $\boldsymbol{\theta} \sim \mathcal{P}(\boldsymbol{\theta})$ while M potentially wrong?

[5] The fun being that until you were top-smart, it is often unclear how small 'minimal' is :)

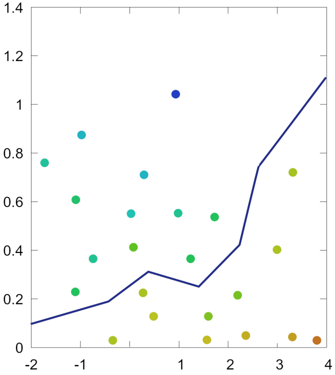
Bayesian NN setup

Classification problem:



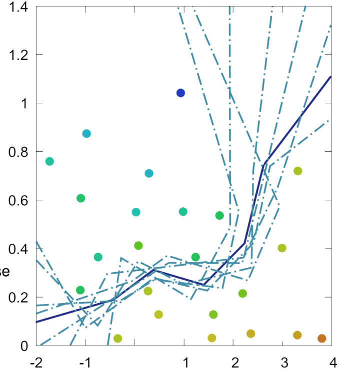
Bayesian NN setup

Traditional Learning (Adams optimizer)



Bayesian NN setup

Bayesian Neural Net

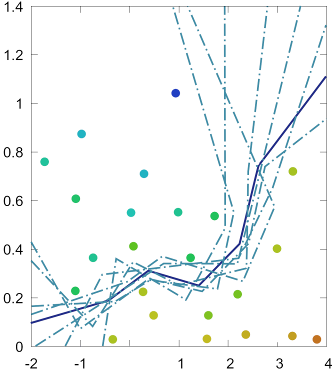


Large uncertainty where no training data.

Many solutions, because finite amount of data!

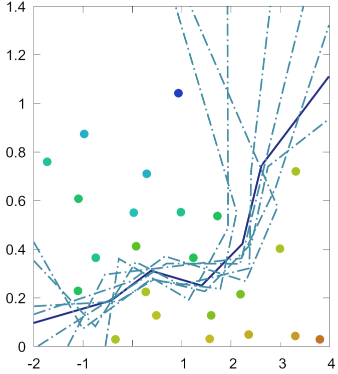
Bayesian NN setup

Q: How many more training data do we need?



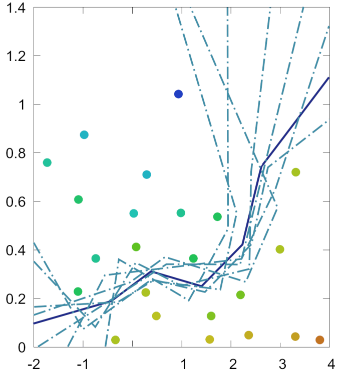
Bayesian NN setup

Q: Which training data can we remove?

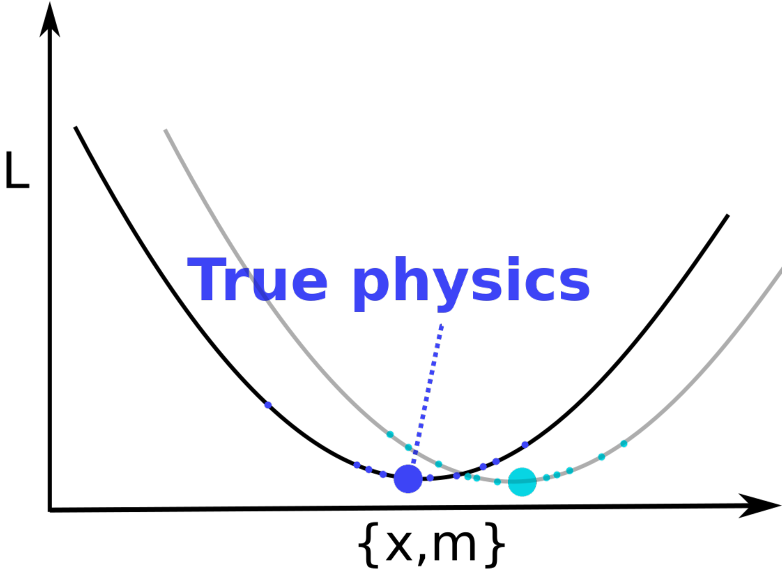


Bayesian NN setup

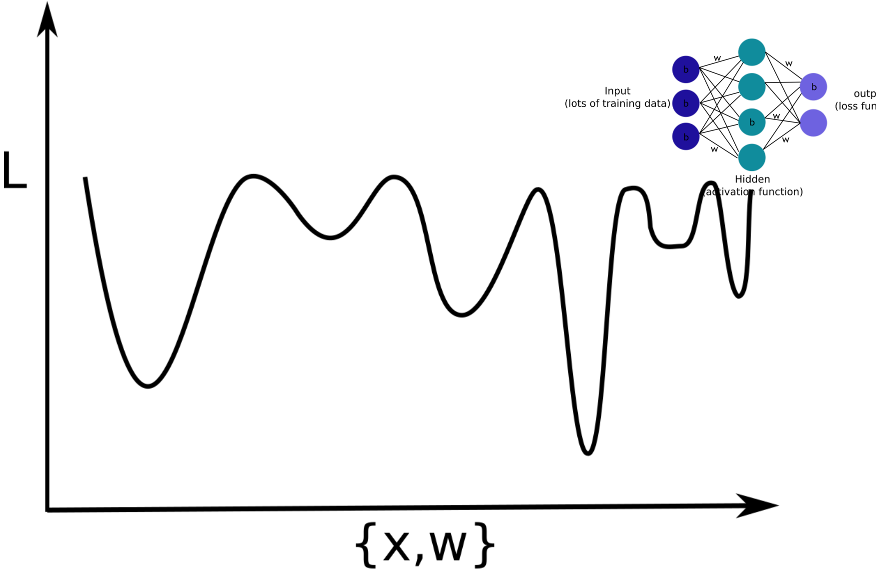
Q: Can we make the network convex? E.g. because we want an auto-encoder and imagine the bottleneck is some parameters?



Convex problem

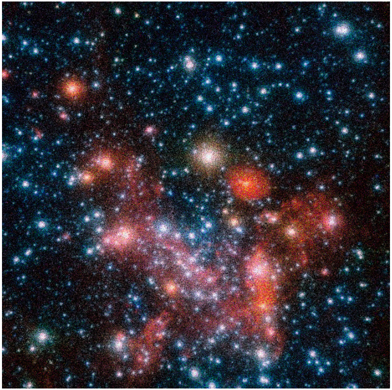


Non-convex problem



The research question in words

Given the Stefan Boltzmann law, $L = \sigma_T A T^4$ and this stellar field, which value for σ_T do you infer?



$\Rightarrow$ Which computation do you have to set up?

Want: σ_T, given everything we know.

$$
\begin{aligned}
\mathcal{P}(\sigma_T|\hat{L}) &= \int \mathcal{P}(\sigma_T, A, T, L|\hat{L})\, \mathrm{d}A\mathrm{d}T\mathrm{d}L \\
&\propto \mathcal{P}(\sigma_T, A, T, L)\mathcal{P}(\hat{L}|L)\, \mathrm{d}A\mathrm{d}T\mathrm{d}L \\
&= \mathcal{P}(L|A, T, \sigma_T)\mathcal{P}(A, T, \cancel{\sigma_T})\mathcal{G}(\hat{L}|L)\, \mathrm{d}A\mathrm{d}T\mathrm{d}L \\
&= \int \delta_D(L - \sigma_T A T^4)\mathcal{P}(A, T)\mathcal{G}(\hat{L}|L)\, \mathrm{d}A\mathrm{d}T\mathrm{d}L
\end{aligned}
\tag{5}
$$

$$
\begin{aligned}
\mathcal{P}(A, T) &= \int \mathcal{P}(A, T, Z, M)\mathrm{d}Z\mathrm{d}M \\
&= \int \mathcal{P}(A, T|Z, M)\mathcal{P}(Z)\mathcal{P}(M)\, \mathrm{d}Z\mathrm{d}M
\end{aligned}
\tag{6}
$$

Q: What is $\mathcal{P}(A, T|Z, M)$? And what is this $\mathcal{P}$?

1. If $\mathcal{P}$ is a delta function $\delta_D[f_{\rm out}(A,Z) - f_{\rm in}(Z,M)]$, then you can train a NN for it to 'sufficient accuracy'.
2. If $\mathcal{P}$ not a delta function, then either train a Bayesian-NN, or go back to (1).

Bayesian Neural Network

- Has the structure of a neural network and still approximates functions.
- Is additionally setup to yield a *distribution* over outputs.
- Trainable parameters drawn from prior distribution $\pi(\boldsymbol{w}, \boldsymbol{b})$.
- Getting posterior $\mathcal{P}(\boldsymbol{w}, \boldsymbol{b}|\boldsymbol{x}_{\mathrm{train}})$ intractable.
- Workflow may include initial grid-search for architecture (depth and layer-sizes).
- All architectures may be marginalized.

An ensemble of nets estimates uncertainty more accurately.
(Drops modelling (aka architecture) dependency.)

BNNs

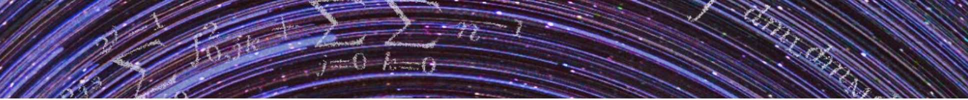
Have: input variables $\boldsymbol{x} \sim \mathcal{P}(\boldsymbol{x})$, but unobservable $\boldsymbol{y}$ interesting.[6]

Want: $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$

Subject to the map $\boldsymbol{x} \to \boldsymbol{y}$ be a NN using weights $\boldsymbol{w}$, $\boldsymbol{b}$, and architecture A.

Q: What would a good Bayesian do?

[6] Astronomical example: $\boldsymbol{x}$ observed stellar surfaces, $\boldsymbol{y}$ stellar interior.



The Bayesian marginalizes everything that is not interesting.

Marginalize over $\boldsymbol{w}, \boldsymbol{b}$:

$$\mathcal{P}_A(\boldsymbol{y}|\boldsymbol{x}) = \int \int \int \dots \int_{w_r} \dots \int_{b_q} \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b}) \pi(\boldsymbol{w}, \boldsymbol{b}) \, \mathrm{d}^r w \mathrm{d}^q b \tag{7}$$

Then model averaging to balance good and bad architectures (implies running an ensemble of neural nets):

$$\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) = \sum_i \epsilon_i \mathcal{P}_{A_i}(\boldsymbol{y}|\boldsymbol{x}) \tag{8}$$

Now you really have $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$[7] (And a numerical nightmare!)[8]

[7] Astronomical example: $\boldsymbol{x}$ observed stellar surfaces, $\boldsymbol{y}$ stellar interior.

[8] Ask A.MY (10^6) & J.B. (target: 10^9)

Summary

1. Even if you don't do BNNs, maybe remember what it really takes to infer unobservables.
2. BNNs nonetheless solve the *important problem* of (un)certainty quantification $\Rightarrow$ at least train a 'swarm' of nets.
3. Inference is never a no-brainer, also not with "machine learning", because there is something we really do not know, and we therefore need logic to reason about it.
4. By being a smart astronomer, you can replace a BNN with an NN in a delta function $\delta_D(\cdot)$...
5. ... and put that $\delta_D(\cdot)$ into a Bayesian Hierarchical Model.

→ Result: You combined learning with inference!