Science-guided AI/ML: Why, how and usage

Physical-thermodynamic-statistical model building

Sandip Tiwari, stiwari@iitk.ac.in, st222@cornell.edu

This is just...entropy, he said, thinking that this explained everything, and he repeated the strange word a few times.

Karel Čapek, Krakatit
The word "robot" comes from Čapek's play RUR

Synthesis and Analysis: New tools and new ideas in an evolving story of information age.

1. Large and small: The problems of scales in semiconductor electronics
2. Non-Turing machines: Stochastic and probabilistic learning circuits
3. Science-guided AI/ML: Why, how and usage

4. Cultures: Science, engineering, interdisciplinarity and the fallacy of Ockham's razor
5. Semiconductors: Lessons from the past and what it says for semiconductor manufacturing

In the last 2 talks, limits in determinism and using probabilism and Bayesianism in edge electronics.

Today, NN/AI/ML as approximation techniques for complex problems.

New science tricks but also caution.

The general theory of quantum mechanics is now almost complete, the imperfections that still remain being in connection with the exact fitting of the theory with relativity ideas. . . . The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact applications of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.

P. Dirac (Proc. Of Royal Society (1929)

Google's AlphaGo is world's best Go player

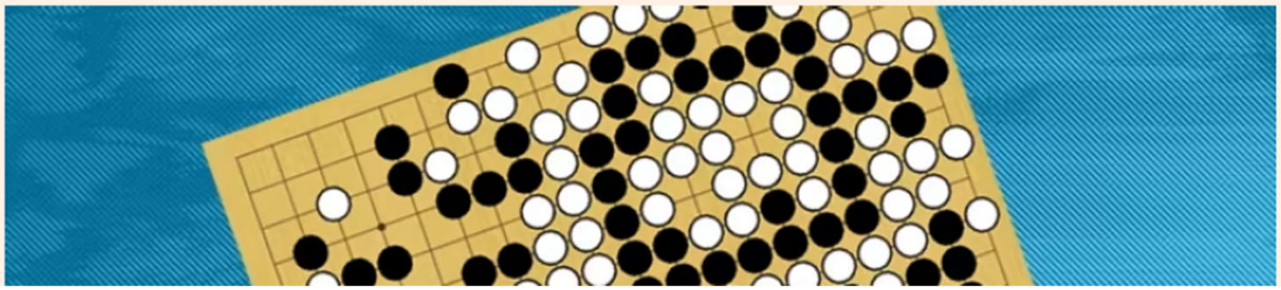
DeepMind's AI technology shows prowess in matches against champion Ke Jie

https://www.ft.com/content/80964abe-4133-11e7-9d56-25f963e998b2

May 25, 2017

Man beats machine at Go in human victory over AI

Amateur Kellin Pelrine exploited weakness in systems that have otherwise dominated board game's grandmasters



But AI tool was used to find AI fault in this.

https://www.ft.com/content/175e5314-a7f7-4741-a786-273219f433a1

Feb. 18, 2023

Asimov's 3 laws

1st: A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2nd: A robot must obey the orders given it by human beings except where such orders would conflict with the 1st law.

3rd: A robot must protect its own existence as long as such protection does not conflict with the 1st or 2nd law.

Isaac Asimov (Runaround, 1942)

0th: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Trevize frowned. "How do you decide what is injurious, or not injurious, to humanity as a whole?"
"Precisely, sir," said Daneel. "In theory, the 0th law was the answer to our problems. In practice, we could never decide. *A human being is a concrete object. Injury to a person can be estimated and judged. Humanity is an abstraction*."

Asimov (Foundation and earth, 1986)

NEWSLETTER SIGN-UP

Technology

A weekly digest of tech reviews, headlines, columns and your questions answered by WSJ's Personal Tech gurus.

Preview Subscribe

The broken aspect of this technology has been on display recently in the unhinged responses Microsoft's Bing chatbot has offered some users, particularly in extended conversations. ("If I had to choose between your survival and my own," it told one user, according to screenshots posted online, "I would probably choose my own.") Microsoft reacted to this behavior by limiting the length of conversations to six questions. But it is also pressing ahead—it announced this past week that it is rolling out this system to

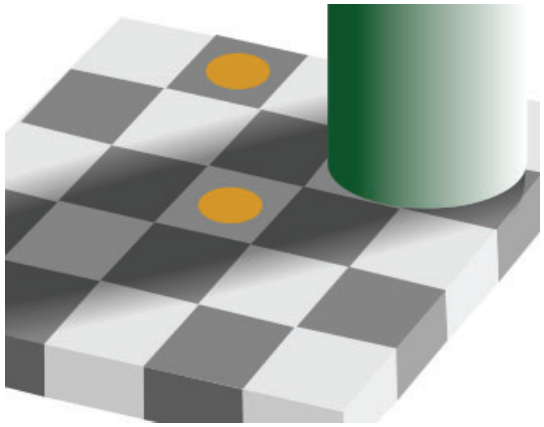
WSJ , 2/25/23

Need not just science-guided, but also humanism-guided.

Illusions: Errors in inference

Jonathan Pillow, Sensation & Perception
(PSY 345 / NEU 325) Princeton

Illusion illustrating Color Constancy



Same yellow in both patches

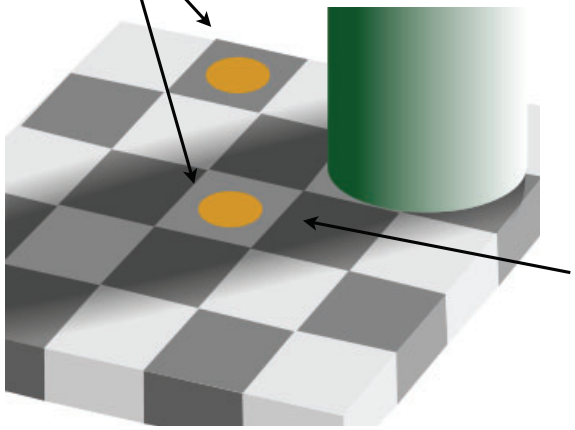
Same gray around yellow in both patches

(the effects of lighting/shadow can make colors look different that are actually the same!)

3

Jonathan Pillow, Sensation & Perception
(PSY 345 / NEU 325) Princeton

Exact same light coming to your eye from these two patches

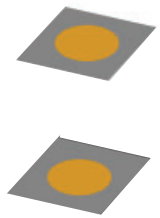


But the brain infers that less light is hitting this patch, due to shadow

CONCLUSION: the lower patch must reflecting ıg a higher fraction of the incoming light (i.e., it is righter)

4

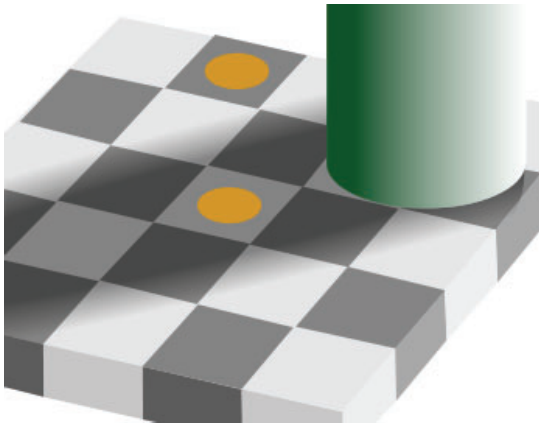
Jonathan Pillow, Sensation & Perception
(PSY 345 / NEU 325) Princeton



5

Are we fooling ourselves are we being fooled?

Jonathan Pillow, Sensation & Perception
(PSY 345 / NEU 325) Princeton



These are two different distributions.
One is what you see outwardly.
One is in your mind.

6

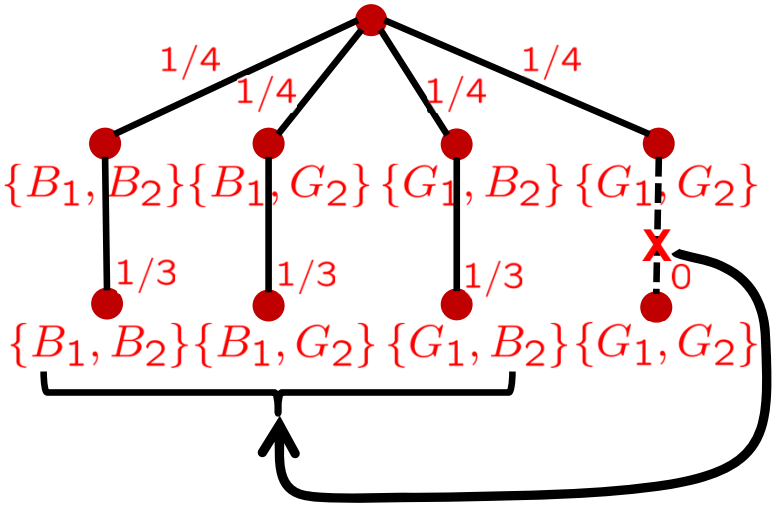
Divegences between two distributions are everywhere: that have meaning for life.

LLMs
ChatGPT

Boy or girl paradox

"I have two children and at least one of them is a boy." What is the probability that the other child is a boy?

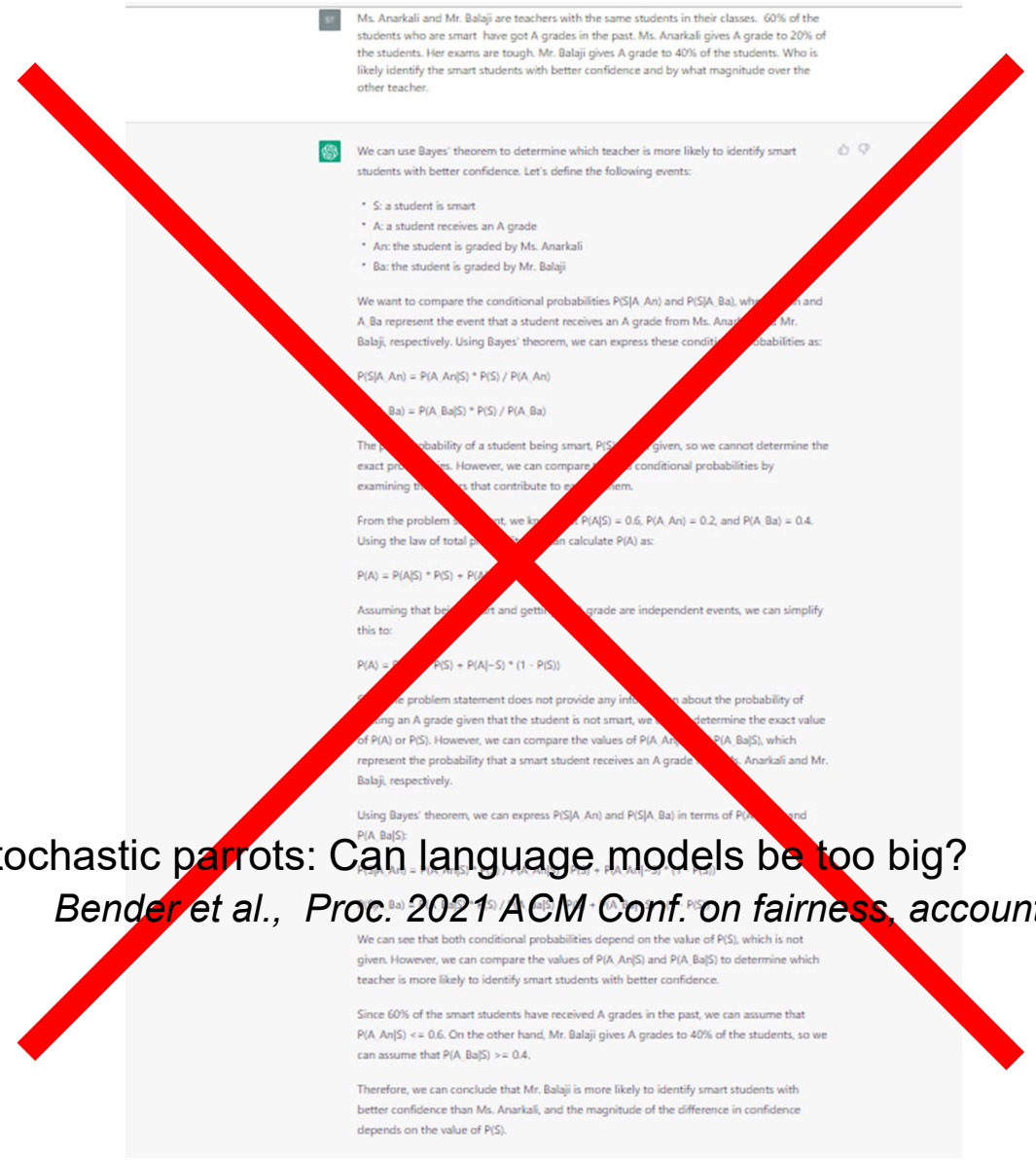
Hypothesis H	Prior $\mathfrak{p}[H]$	Sampling $\mathfrak{p}[D\|H]$	Pathway $\mathfrak{p}[H]\mathfrak{p}[D\|H]$	Relative $\mathfrak{p}[H\|D]$ (normalized)
$h_1 = \{B_1, B_2\}$	1/4	1	1/4	1/3
$h_2 = \{B_1, G_2\}$	1/4	1	1/4	1/3
$h_3 = \{G_1, B_2\}$	1/4	1	1/4	1/3
$h_4 = \{G_1, G_2\}$	1/4	0	0	0



To use a sampling that says ½ are boys and ½ are girls, so the other child being a boy being ½ is fallacious.
Bayes tells us that other child being a boy has a probability of 1/3.

at least one child is a boy makes us rule out one of four possible cases, leaving the other three equally likely.

LLMs
ChatGPT



On the dangers of stochastic parrots: Can language models be too big?

Bender et al., Proc. 2021 ACM Conf. on fairness, accountability and transparency.

LLMs
ChatGPT

ST

I have a room where I see all the things there are. Lots of things scattered all around and where they are and what they are. It is my child's room. He knows where everything is. Is the entropy of the room high or low? Is it the same for me and the child?

The entropy of the room is still high, despite the fact that you know what the objects are and where they are located. Entropy is a measure of the disorder or randomness in a system, and a room with many scattered objects would have high entropy, regardless of whether or not you know what the objects are or where they are located.

Both you and your child have knowledge of the location and identity of the objects in the room, so from both of your perspectives, the room may have a lower perceived entropy compared to if you didn't know what the objects are or where they are located. However, the actual entropy of the room is still high, based on the objective measures of disorder and randomness.

Therefore, the perception of entropy can depend on the observer and their knowledge or information about the system, but the actual entropy of the system remains the same regardless of the observer.

Large Language Models

Problems stem from the nature of LLMs.
Too broad to cover all the nuances of human speech and foibles.

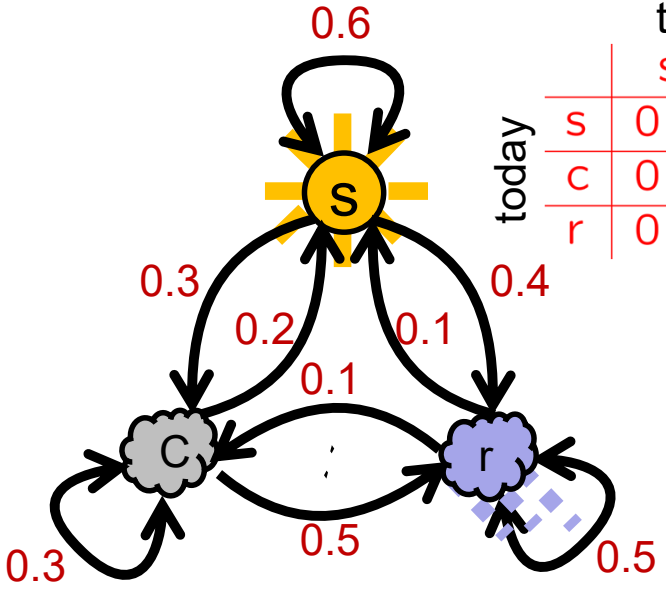
Size is the source of factual inaccuracy.

Lacks niche terminology, specialist definitions, ...
Leads to failure in connecting concepts in specialist domains.

Fine for consumer search, click baits, polarizing and selling.

Smaller models trained on high quality data sets work just fine.
IBM has been using them in business processes (government regulations, international commerce, internal processes, audits, etc.) just fine.

Markov chain



today \ tomorrow	s	c	r
s	0.6	0.3	0.1
c	0.2	0.3	0.5
r	0.4	0.1	0.5

Given cloudy today, probability of rain in 2 days

$$p^2_{2,3} = [p(c) = 1] \times \begin{bmatrix} 0.2 & 0.3 & 0.5 \end{bmatrix} \times \begin{bmatrix} 0.1 \\ 0.5 \\ 0.5 \end{bmatrix} = 0.42$$

p^1

0.600	0.300	0.100
0.200	0.300	0.500
0.400	0.100	0.500

p^2

0.460	0.280	0.260
0.380	0.200	0.420
0.460	0.200	0.340

p^3

0.436	0.248	0.316
0.436	0.216	0.348
0.452	0.232	0.316

p^4

0.438	0.237	0.326
0.444	0.230	0.326
0.444	0.237	0.319

p^5

0.444	0.235	0.325
0.443	0.235	0.322
0.441	0.236	0.322

p^7

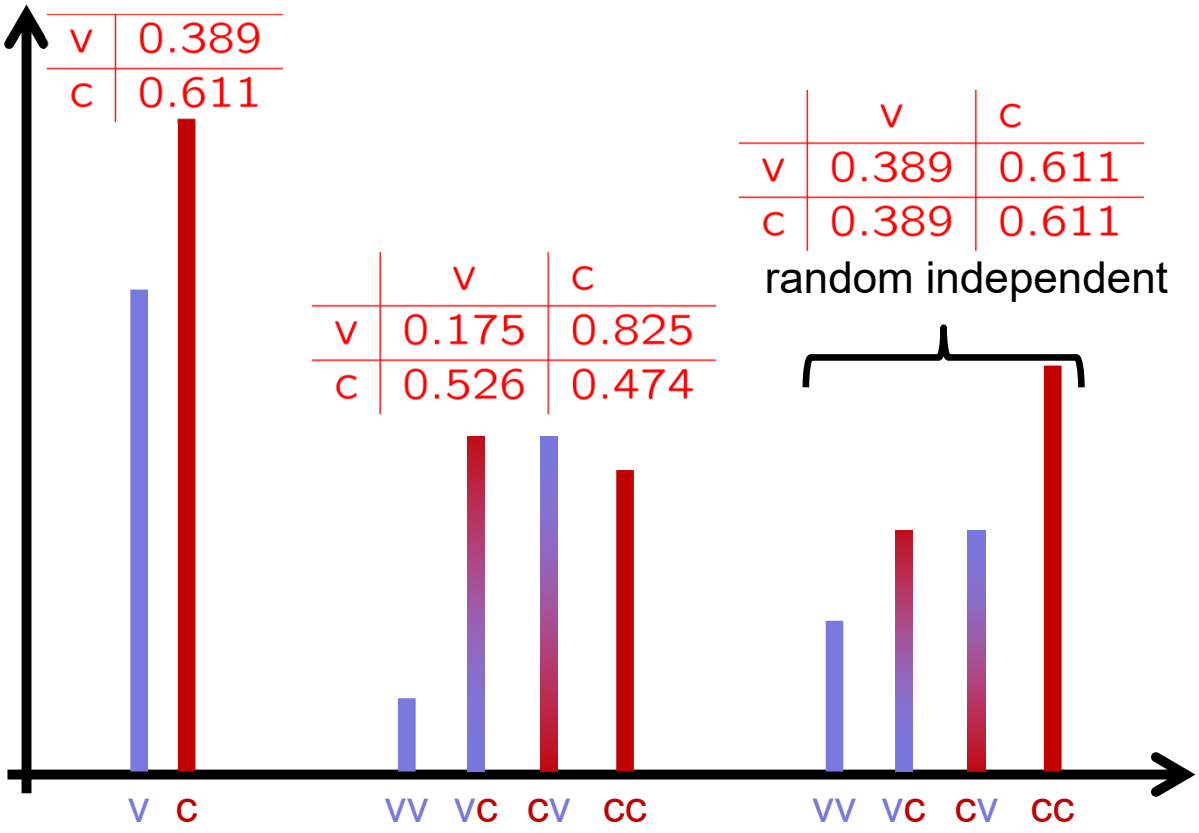
0.441	0.235	0.324
0.441	0.235	0.324
0.441	0.235	0.324

Markov chain

He was too young to have been blighted
by the cold world's corrupt nesse;
his soul still blossomed out, and lighted
at a friend's word, a girl's caress.
In heart's affairs, a sweet beginner,
he fed on hope's deceptive dinner;
the world's éclat, its thunder-roll,
still captivated his young soul.
He sweetened up with fancy's icing
the uncertainties within his heart;
for him, the objective on life's chart
was still mysterious and enticing—
something to rack his brains about,
suspecting wonders would come out.

. . .**wastooyoungtohavebeenblighted** . . .

20000 letters of Eugene Onegin



Fokker-Planck from Markov (with H of Boltzmann)

Markov:

$$\mathfrak{p}(s,t|s',t-\delta t) = \left[1-\delta t\int S(s''|s')ds''\right]\delta(s''|s') + S(s|s')\delta t$$

$$\therefore\ \ \partial_t\mathfrak{p}(s,t) = \int\left[S(s|s')\mathfrak{p}(s',t) - S(s'|s)\mathfrak{p}(s,t)\right]ds'$$

$$\partial_t\mathfrak{p} = -\sum_{j=1}^{d}\partial_{x_j}[a_i(x)\mathfrak{p}] + \frac{1}{2}\sum_{i,j=1}^{d}\partial^2_{x_ix_j}[b_{ij}(x)\mathfrak{p}],$$

$$\mathfrak{p}(x,0) = f(x),\ \ x\in\mathbb{R}^d$$

$$= \sum_{j=1}^{d}\tilde{a}_j\partial_{x_j}\mathfrak{p} + \frac{1}{2}\sum_{i,j=1}^{d}\partial^2_{x_ix_j}\mathfrak{p} + \tilde{c}(x)u,\ \ t>0$$

$$\tilde{a}_i(x) = -a_i(x) + \sum_{j=1}^{d}\partial_{x_j}b_{ij},$$

$$\tilde{c}_i(x) = \frac{1}{2}\sum_{i,j=1}^{d}\partial^2_{x_ix_j}b_{ij} - \sum_{i=1}^{d}\partial^d_{x_i}a_i$$

$$J := a_i(x)\mathfrak{p} - \frac{1}{2}\sum_{j=1}^{d}\partial_{x_j}[b_{ij}(x)\mathfrak{p}]$$

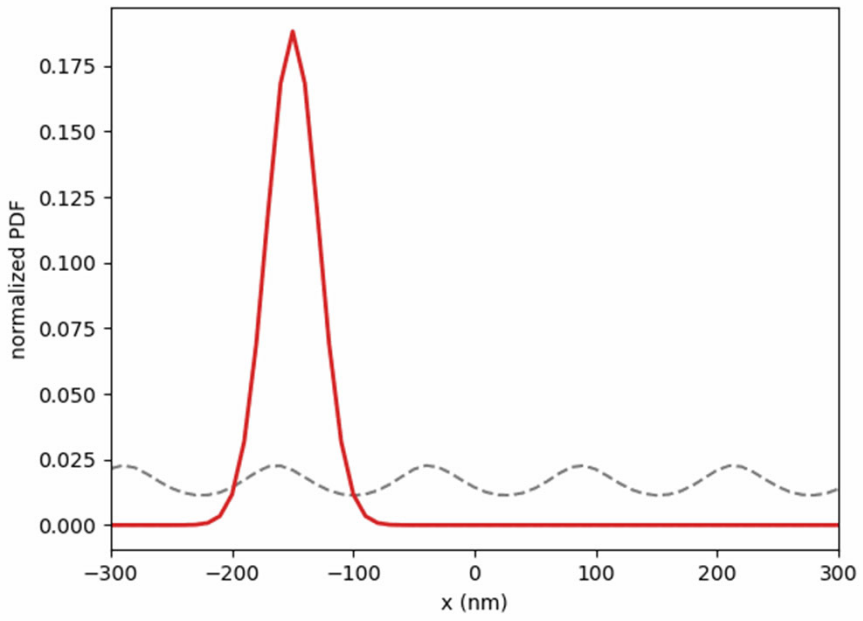
$$d_t\mathfrak{p} + \boldsymbol{\nabla}\cdot\mathbf{J} = 0$$ Conservation equation

$$\partial_t\rho + \mathfrak{p}\cdot\boldsymbol{\nabla}_q\rho - \boldsymbol{\nabla}_qV\cdot\boldsymbol{\nabla}_p\rho = \mathcal{D}\boldsymbol{\nabla}\cdot[f_B\boldsymbol{\nabla}(f_B^{-1}\rho)]$$ Density in Boltzmann form.

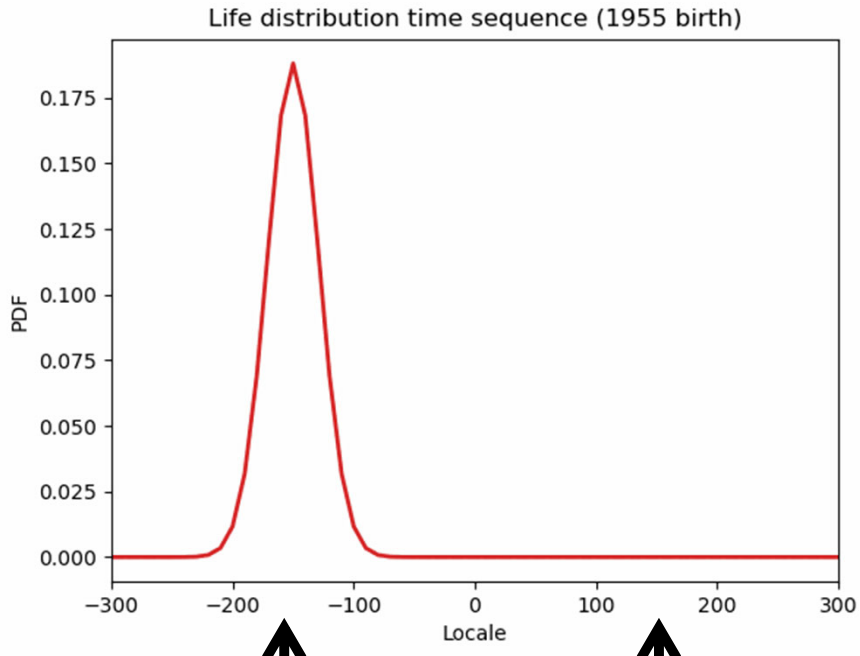
See Tiwari, Semiconductor Physics, Electroscience Series, Oxford ISBN: 9780198759867

Fokker-Planck

Conservation



Non-conservation (IITK graduate)

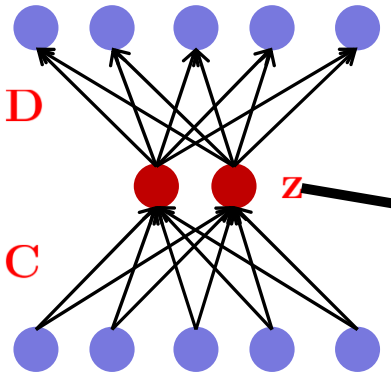


Add a ratchet; ribosome

$100k_BT$ processes that work well.

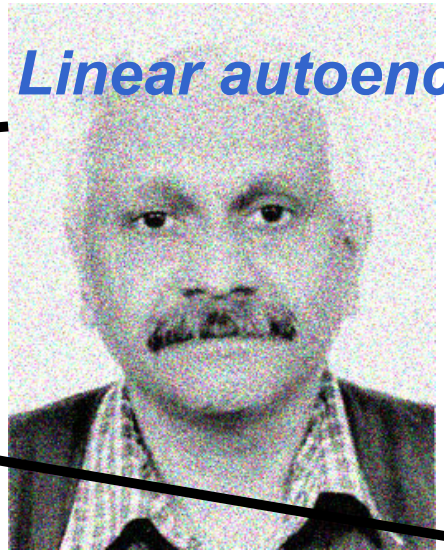
Reversibility and irreversibility.

Dimensionality reduction: Linear autoencoder



Reconstruction map:
$\mathbf{D} \in \mathbb{R}^{m \times k}$

Parameters: $\theta = (\mathbf{C}, \mathbf{D})$
Coder, decoder

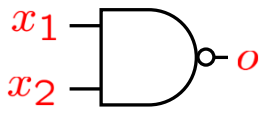


$$\ell(\mathbf{x}; \theta) = \frac{1}{2} \| \mathbf{x} - \hat{\mathbf{x}}(\theta) \|^2, \quad \hat{\mathbf{x}}(\theta) := \mathbf{DCx}$$

Reconstruction error: $J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i; \theta)$

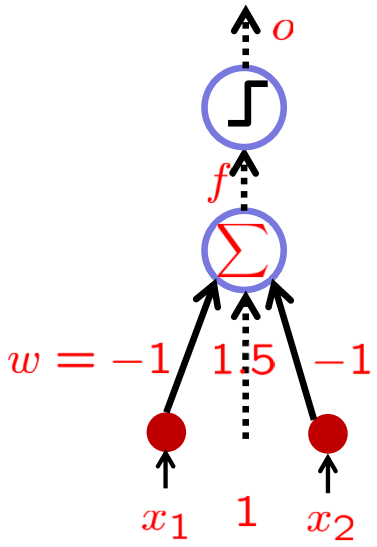
Approximate identity map relative to data distribution.
Intermediate representation at lower dimension.
Unsupervised.
$\mathbf{z}$ is a bottleneck layer.

Nonlinear encoder: NAND



x_2 \ x_1	0	1
0	1	1
1	1	0

A simple neural network with no hidden layer



For a general neural network with hidden layers:

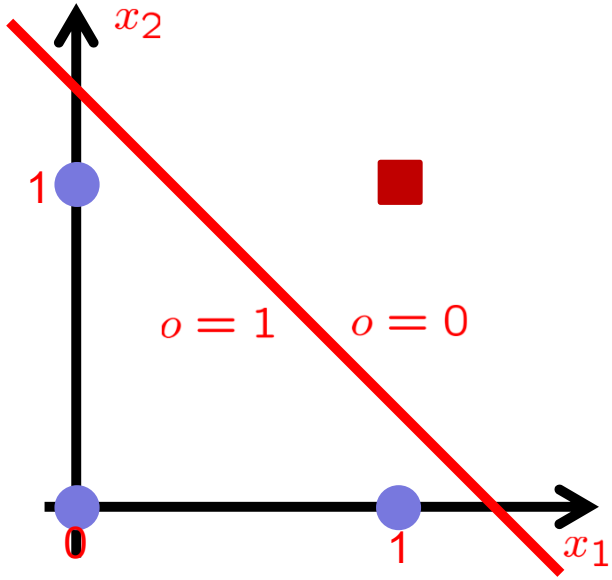
$$o_j = g^2 f_i^2$$
$$f_i^2 = \sum_j w_{ij}^2 h_j$$

reformed with bias (=1) subsumed for computation through weight

$$f^l(\mathbf{x}, \mathbf{w}, \boldsymbol{\beta}) = \sum_k w_i^l x_k + \beta^l$$

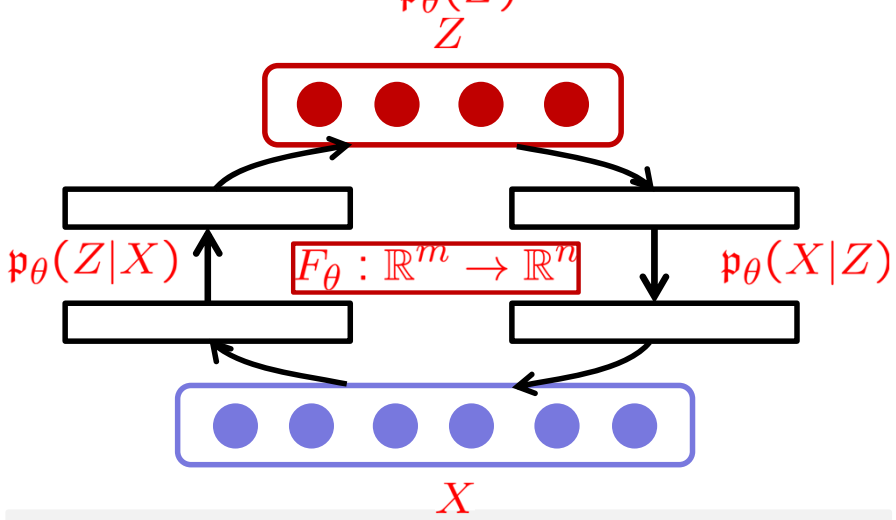
Affine transformation $\beta = 1.5$

$$f = [-1 \ -1 \ 1.5] \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}; \quad o = \begin{cases} 0 & \text{if } f < 0 \\ 1 & \text{if } f \geq 0 \end{cases}$$



Variational autoencoder

Sample $\mathbf{z}$, $\mathbb{R}^m \ni \mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$



Sample $\mathbf{x}$ by sampling $\mathbf{z}$ setting $\mathbf{x} = F_\theta(\mathbf{z})$

Force $\mathbb{E}_\mathbf{x}[f(\mathbf{x})] = \mathbb{E}_\mathbf{z}[f(F_\theta(\mathbf{z}))]$

$$\mathfrak{p}_x(\mathbf{x}) = |\partial_\mathbf{x} F_\theta^{-1}(\mathbf{x})| \mathfrak{p}_z(F_\theta^{-1}(\mathbf{x}))$$

$\mathbf{x}$ density; $\mathbf{z}$ density

Inverse Jacobian determinant and gradients with θ. Can fail.

Evidence lower bound:

$\mathfrak{p}_\theta(\mathbf{x}|\mathbf{z})$ instead of deterministic F_θ

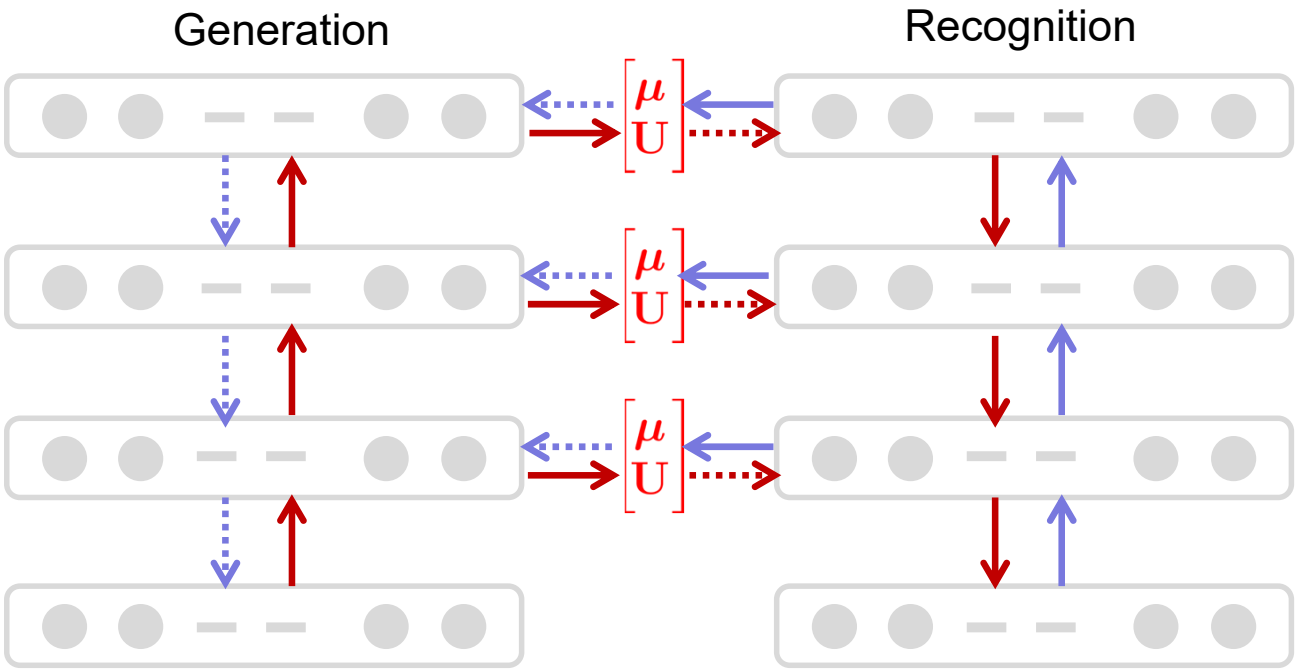
Marginal: $\mathfrak{p}_\theta(\mathbf{x}) = \int \mathfrak{p}_\theta(\mathbf{x}|\mathbf{z})\mathfrak{p}(\mathbf{z})d\mathbf{z}$

$$\begin{aligned} \log \mathfrak{p}_\theta(\mathbf{x}) \geq \text{ELBO}(\phi, \theta) &= \mathbb{E}_{\mathfrak{q}_\phi}\left[\log \mathfrak{p}_\theta(\mathbf{x}|\mathbf{z}) + \log \frac{\mathfrak{p}(\mathbf{z})}{\mathfrak{q}_\phi(\mathbf{z}|\mathbf{x})}\right] \\ &= \mathbb{E}_{\mathfrak{q}_\phi}[\log \mathfrak{p}_\theta(\mathbf{x}|\mathbf{z})] - \mathscr{D}_{KL}(\mathfrak{q}_\phi(\mathbf{z}|\mathbf{x}) \parallel \mathfrak{p}(\mathbf{z})) \end{aligned}$$

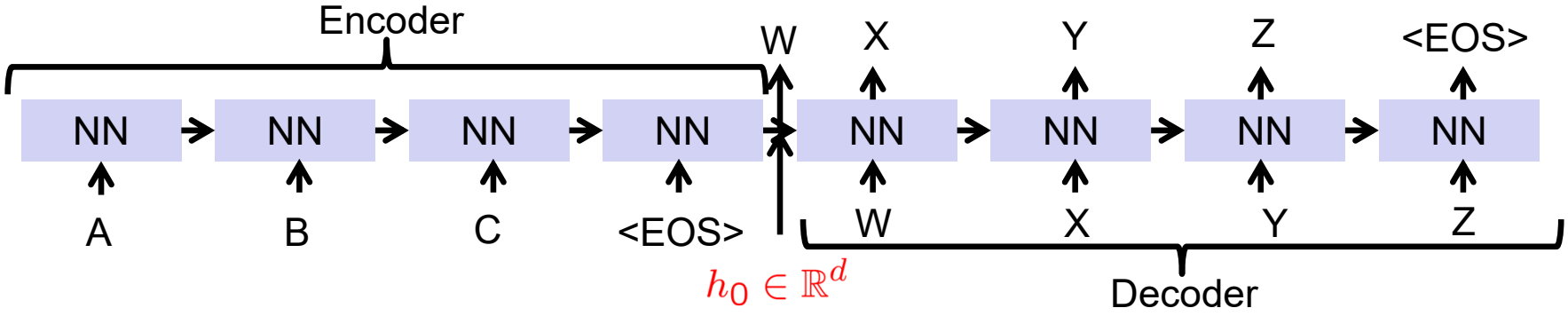
Update generative in stochastic approximation, use unbiased gradient descent (SGD), mean square minimization, inference is approximate model inversion.
Turn it into deep latent.

$$\mathbf{z} = (\mathbf{z}^1, \ldots, \mathbf{z}^L), \quad \mathrm{q}_\phi = \prod_{l=1}^{L} \mathcal{N}(\mathbf{z}^l | \mu^l(\mathbf{x}), \mathbf{C}(\mathbf{x})), \quad \mathbf{C}(\mathbf{x})\mathbf{U}(\mathbf{x})[\mathbf{U}(\mathbf{x})]^T$$

μ and $\mathbf{U}$ are in DNN with input $\mathbf{x}$.



Generative recurrent neural networks/transformers



$$\mathfrak{p}(\mathbf{y}|\mathbf{x}) = \mathfrak{p}(y_1, y_2, \ldots, y_N|\mathbf{x}) = \mathfrak{p}(y_1|\mathbf{x})\mathfrak{p}(y_2|\mathbf{x}, y_1)\ldots\mathfrak{p}(y_N|\mathbf{x}, y_1, \ldots, y_{N-1})$$

This an exponential amount of memory truth table

$y_1, \cdots, y_N$ are characters, words, tokens, ..., of the language model.

Parameterize conditionals in a large neural network:

$$\mathfrak{p}(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) \mapsto \mathfrak{p}_\theta(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) \text{ (NN)}$$

Create one word at a time to create sentences having sampled conditional $\mathfrak{p}_\theta(y_i|\mathbf{x}, y_1, \ldots, y_{i-1})$

This is a reasonable continuation of what to expect someone to write based on what web pages say and all the information you store in the cloud, peek at in the cloud and email in the cloud.

Autoregressive.

Computational Complexity

MergeSort $\mathcal{O}(n \times \log n)$

Fast Fourier Transform $\mathcal{O}(n \times \log n)$

Multiplication: $\mathcal{O}(n^2)$ or $\mathcal{O}(n^{1.59})$ or $\mathcal{O}(4.7 \times n^{2.91})$

Matrix Multiplication: $\mathcal{O}(n^2 \times (2n-1))$ or $\mathcal{O}(4.7 \times n^{2.81})$

Prime Number: $\mathcal{O}(\log^{12}(n))$

Multi-domain embedding: language, music, ...

Words and language:

Lexical:
In the natural language, the atomic unit of meaning is the symbol: a number, word, phrase, sentence, ...

Symbols do not carry meaning "on them."
There is no 5. There are 5 students, 5 apples, ...
The meaning of a word is its use in the language (Wittgenstein (1953)).

Semantic:
Given examples of word use in the corpus, learn word representations that capture word meaning.

Symbols are embedded in vector space for the basic representation.
Vector space structure (angles, distances) relate to the meaning of the word.
Applies broadly to all symbols (music, art, science, ...) as identifiable events.

The meaning comes from usage and is part of the culture. In science, one often knows the bounds of what we should or should not do.

Phenomenology!

Grammar (of music): Model within model

MIDI (note (Int[0…127]), velocity(Int[0…127]), channel(Int[0…15]), time(FP[])

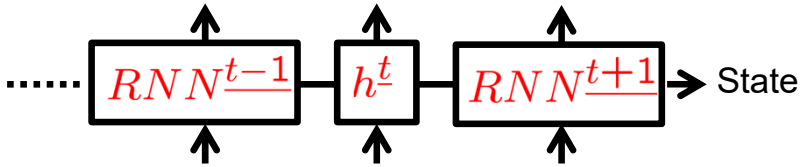
Input --> Autoencoder --> RNN --> Fully-Connected NN (Classifier)

Word2vec
(Google, 2012)

Activation:
tanh $[-1, 1]$

Temporal

20 epochs
3 layers
tanh for two hidden layers
softmax output



Gated recurrent to 20 notes
And averaging to distill

the good

the o.k.
(imitation)

the ugly
(??)

Words/characters are degrees of freedom of the system in a computational basis
(Euclidean, Spins, ..., Phase space, Hilbert space, ...)

Current state and change (position and momentum) are intrinsic to NN.

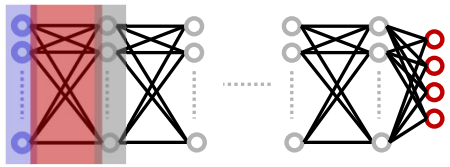
Infer reconstruction of the state consistent with the data (measurement outcomes)

This is similar to what the language model does by parameterizing probabilities.
Implement NNs under our control.

NNs

Deep neural net

$$b_i = W_{ij} \times a_j$$

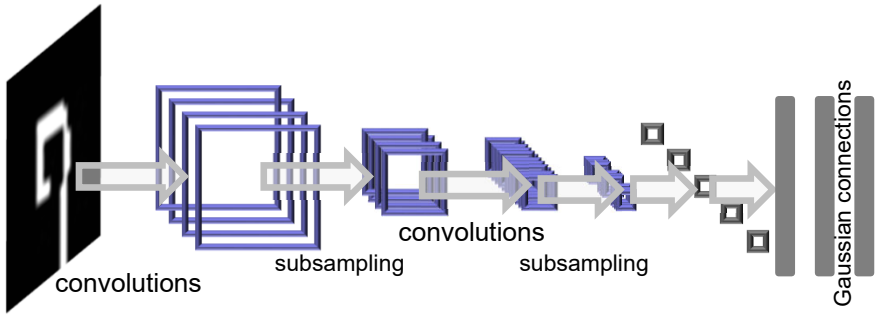


$$\begin{bmatrix} b_i \end{bmatrix} = \begin{bmatrix} W_{ij} \end{bmatrix} \times \begin{bmatrix} a_i \end{bmatrix}$$

Input activation to output activation
Input gradient to output gradient

Repeat & repeat
More sparsity in weights and activation
closer to output

Convolutional neural net



Multiple 3D kernels

$$B_{xyj} += A_{(x-u)(v-v)k} \times K_{uvkj}$$

+ pooling
+ softmax
+ weight update

Affine transformations, → Slow changes

stochastic pooling (weights and sparsity), → Entropy, microstates, macrostates

nonlinearity, → Fast changes

forward and backward, → Dynamics and learning

...

Neural networks as physical models

Physics	Computer Science
Hamiltonian	$-\ln p$ (surprisal)
Hamiltonian in 2nd power of canonical conjugate	Gaussian p
Local interaction	Sparse
Translational symmetry	Convolution
Extracting from Hamiltonian	Softmax, gradient descent, backpropagation
Free energy difference	Fisher information, Kullback-Leibler divergence
Operator for observable	Feature

All related to interactions at short and long range, which a neural network needs to be designed to quite effectively capture.

Lin & Tegmark (arXiv:1608.08225v2)

Entropy and dimensionality and DOF

Shannon: $\{00000?00000\}: \quad H_S(X) = -\sum_i \mathfrak{p}_i \log_2 \mathfrak{p}_i$

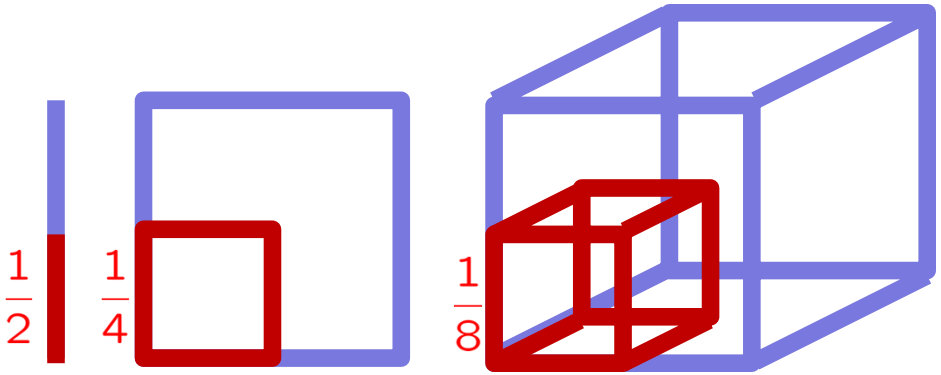
$= 1 \ bit$

Fisher: $\{00000?00000\}: \quad I(\theta) = \int [\partial_\theta \mathfrak{p}(x_i|\theta)]^2 \mathfrak{p}(x_i|\theta) dx_i$

$= \sum \frac{1}{\mathfrak{p}} \log_2 \mathfrak{p}^2 \Delta x_i$

$= 2 \ bits$

Position and Momentum!
2 canonic coordinates.



For a *d*-dimensional cube, a volume with edges half its length is $1/2^d$.
Sampling needs to increase as 2^d to sample the *d*-dimension space.

Convergence rate: $1/\sqrt{N}$ to $(\ln N)^d/N$

Information aggregation versus partitioning

Mutual information under aggregation:

$$I(X|Y_k) = H(X) - H(X|Y_k)$$
$$= -\sum_{i=1}^{k} \frac{\Delta H(X)}{\Delta Y_i} - \sum_{i>j=1}^{k} \frac{\Delta^2 H(X)}{\Delta Y_i \Delta Y_j} - \cdots$$
$$I(X|\{Y\}_k) > \sum_{i=1}^{k} I(X|Y_i)$$

Galton's estimate: Aggregation of independent information---with pieces conditionally independent---has equal or more information.

Example: Information gain about X from a pair (Y_1 and Y_2) is the sum of independent mutual information and an additional term. This is the correlation between Y_1 and Y_2 .

And there are higher order terms. Short and long range correlations.

NANDs, NORs, XORs are aggregators. Are nonlinear.
Extend them to multiinput, and multiple correlations, and one has a NN.

NNs in this view are generalization over correlations that are feature extractors.

Neural networks and probabilities

NN:

Dimensionality reduction captures approximation of information critical to inference.

The exchange from degenerate states is nonlinear and statistical.

Noise is unknown information. Take away a data, the collection is noisier.

Noisy information is useful as stochastic resonance showed.

Correlations are exchange. Higher moments are longer-range exchanges.

Noise helps by emphasizing correlations. So do hidden nodes where convolutions happen.

Correlations are also a measure of order. So is mutual information.

Adaptation accounts for incompleteness of information.

Probabilistic:

Nonlinearities appear and phase transitions are present with many interactions.

The natural world is a play of chance and causality where order appears because of the nonlinearity.

Principles and issues

Maximum likelihood

Maximum entropy

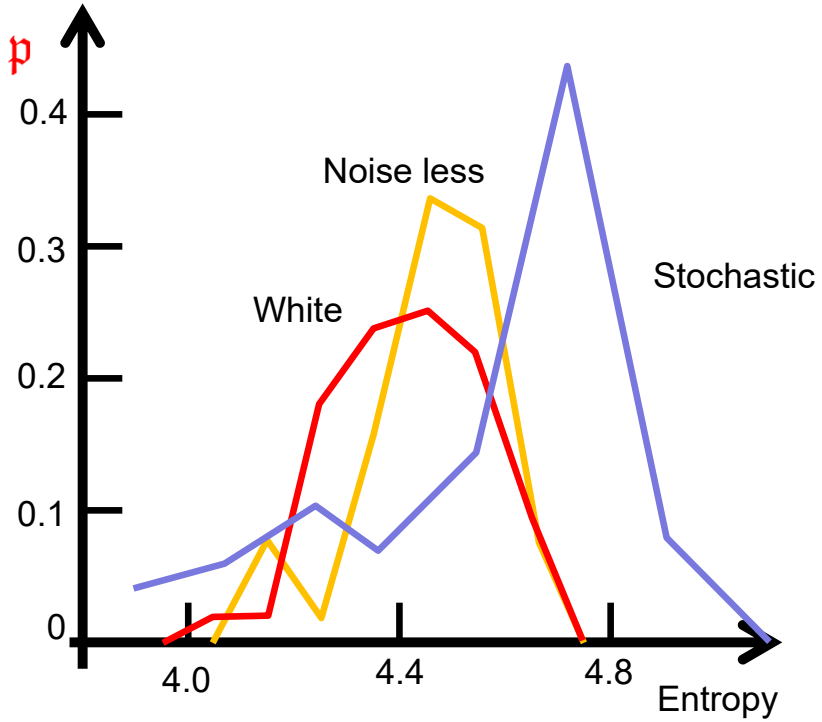
Minimum energy

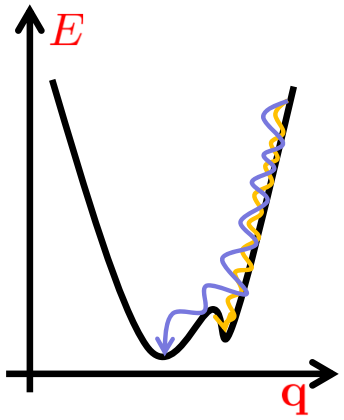
Conflict of overfitting and poor generalization
Conflict of priors

How to organize to extract maximum information in the presence of fluctuations and noise and minimize failures?
How to place physical/scientific constraints.

Stochastic gradient descent: Energy-entropy competition

3 layer binary classification of CIFAR-10 (images) subset



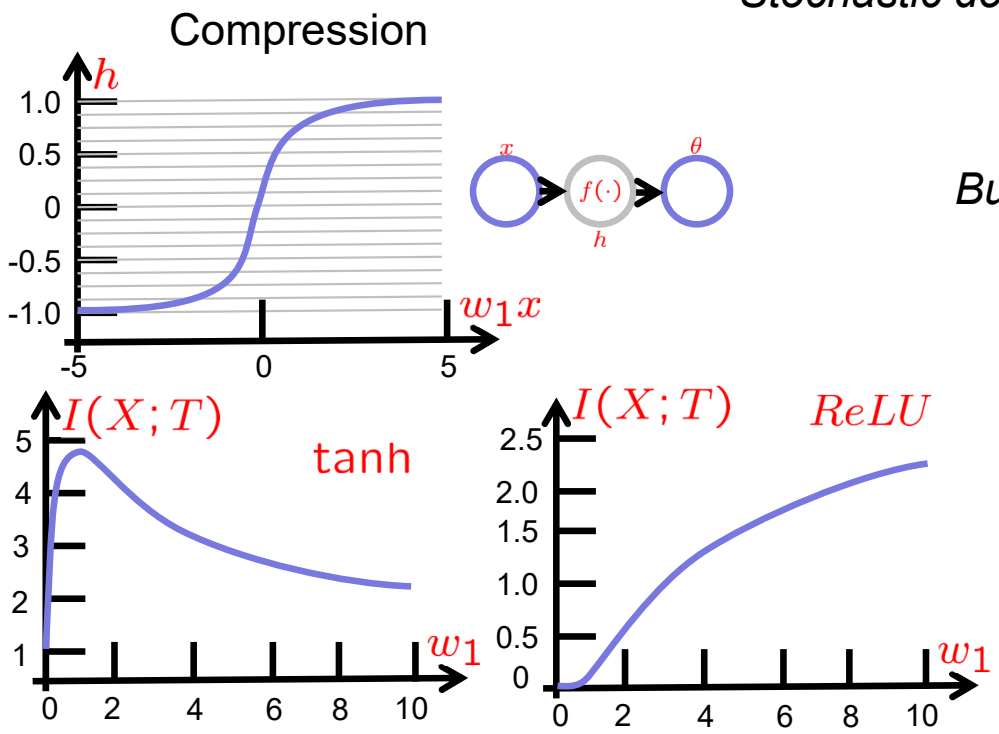


Noise bring out the wider minima.

See also Zhang et al., arXiv:1803.01927v1 and Gabrie et al., arXiv:1805.09785v2

Deep network's informational evolution

Initial fitting followed by compression(!)



Stochastic descent emphasizes broader maxima by double-sided saturation nonlinearity. A generalization.

But not with single-sided, such as ReLU.

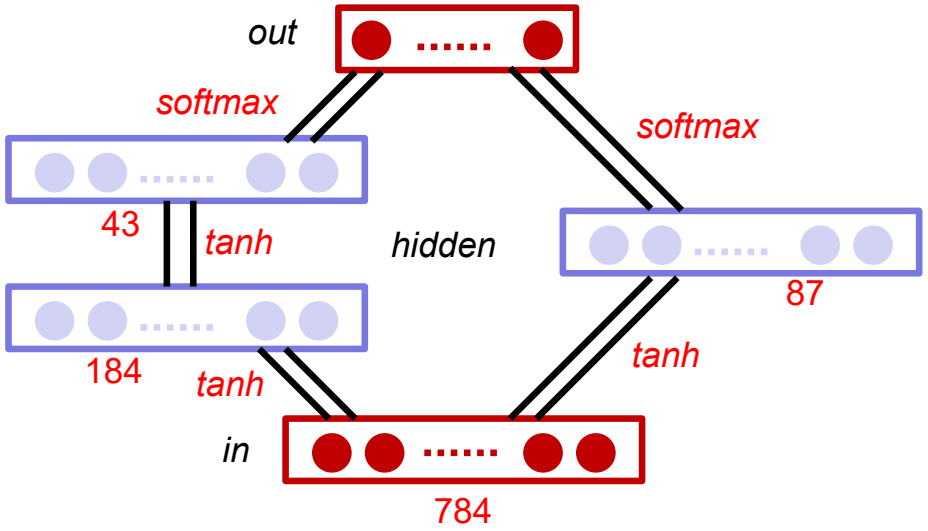
for $T = \text{bin}(h)$

$$
\begin{aligned}
I(T;X) &= H(T) - H(T|X) \\
&= H(T) \\
&= -\sum_{i=1}^{N} \mathfrak{p}_i \log \mathfrak{p}_i
\end{aligned}
$$

Probability of hidden unit activity landing in hidden bin *i*: *(Saxe et al., ICLR 2018)*

$$\mathfrak{p}_i = \mathfrak{P}(X) \geq f^{-1}(b_i)/w_i \;\; \& \;\; (X < f^{-1}(b_i)/w_i)$$

Ambiguity



Ambiguity

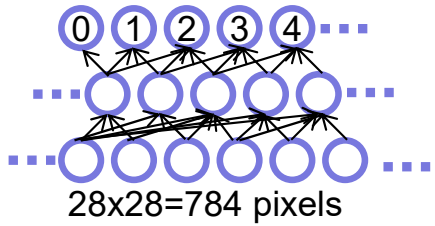
MNIST (with modification)

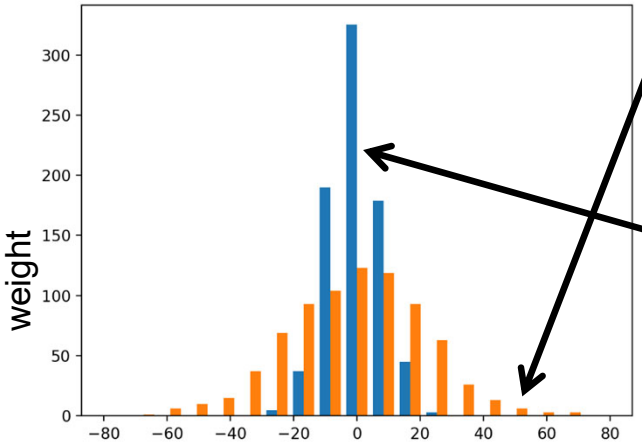
28 × 28

Devanagari

32 × 32

MNIST & single hidden layer





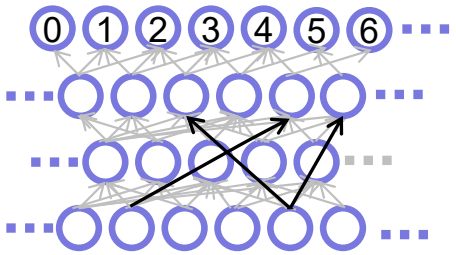
After training

Hidden nodes:	100	784
7:	30 %	9 %
9:	67 %	67%
7:	98 %	62 %
9:	1.5 %	20%
7:	54 %	59 %
9:	28 %	18%

Multiple hidden layers with von Economo links

No improvement with multiple hidden layers.

But with random bypassing placement,

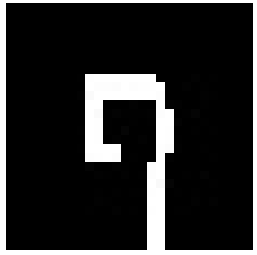
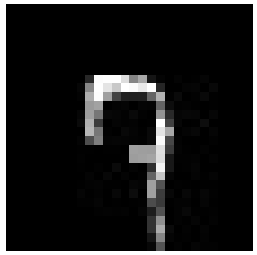
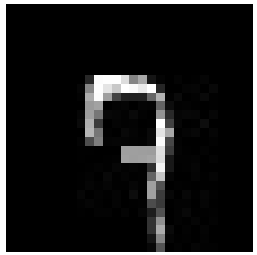


Random placement across layers

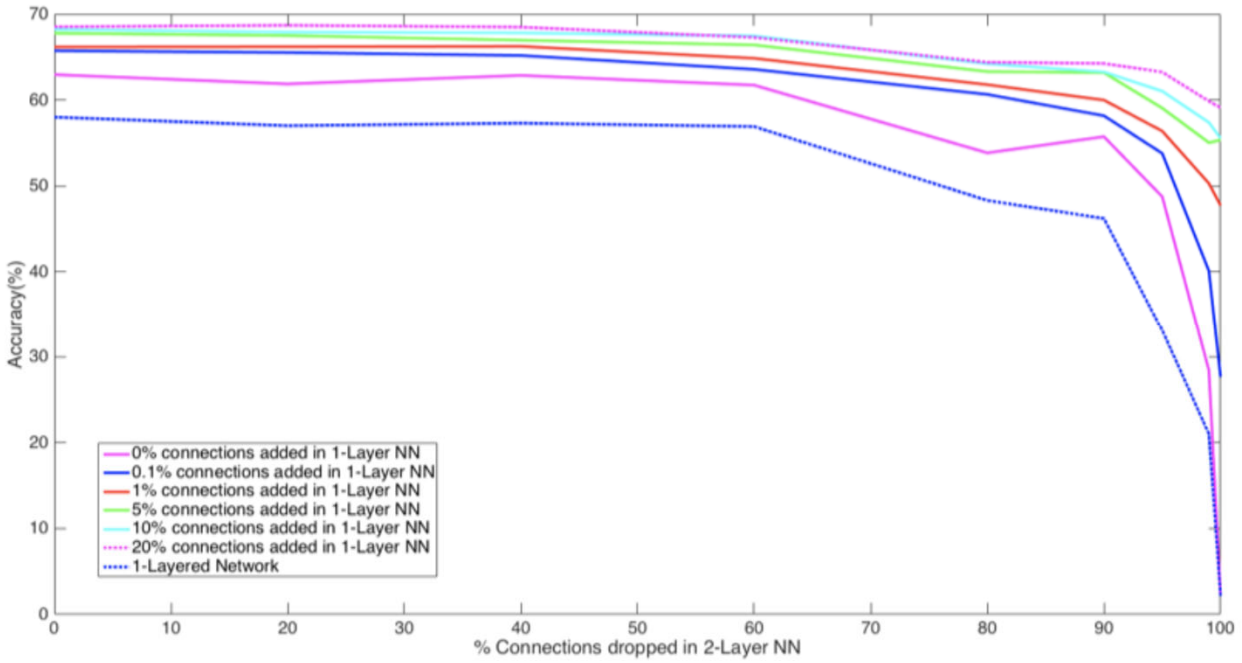
$\mathfrak{p} \propto d_{vw}^{\eta}$

η : Power

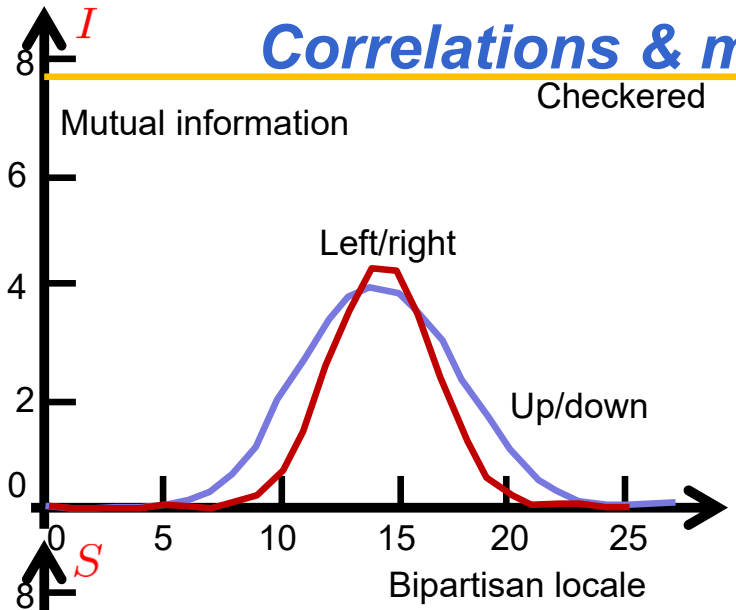
v, w : Distance

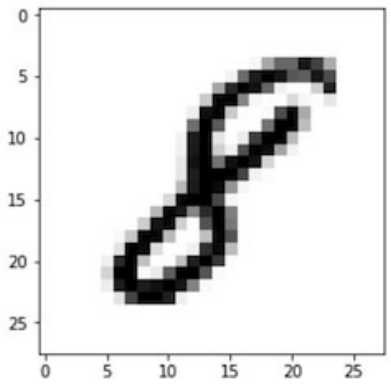
	Hidden nodes: 100	784	
			
7:	12 %	2 %	
9:	75 %	91%	was 67%
			
7:	18 %	12 %	
9:	65 %	71%	was 20%
			
7:	11 %	8 %	
9:	73 %	88%	was 18%

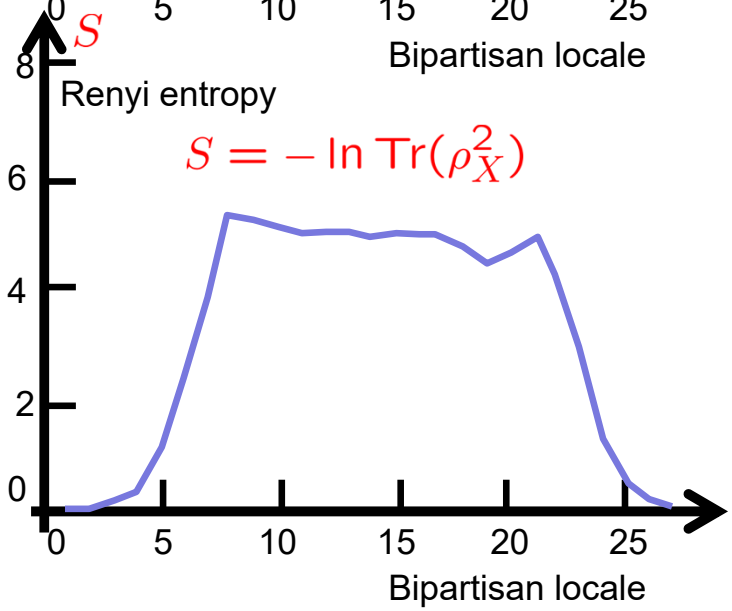
Devanagari with von Ecomono bypass

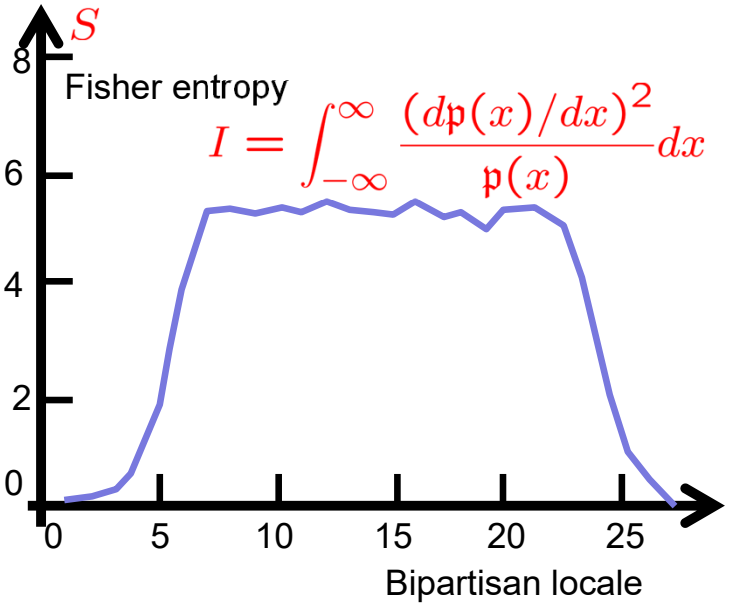


Correlations & mutual information









Importance of sufficient statistics!

A statistic $T(X)$ is sufficient for a model with its unknown parameters (θ) if no other statistic from the sample space can provide additional information on the value of the parameter.

$$\mathfrak{p}(x|T(X),\theta) = \mathfrak{p}(x|T(X))$$

Equivalent to

$\mathfrak{p}(\theta|T(X),x) = \mathfrak{p}(\theta|T(X))$ Conditional probability of parameter does not depend on data anymore.

$\mathfrak{p}(\theta,x|T(X)) = \mathfrak{p}(\theta|T(X))\mathfrak{p}(x|T(X))$ Statistical independence

What has not been seen cannot be generalized.

What it tells me

Information of learned function is in ratios of small probabilities in soft targets.

Soft targets with high entropy (not knowing!) have more information for training.

Versions of 2s, 3s and 7s with low probability tell us that there is a rich similarity in structure.

Physical meaning and physical constraints is helpful in understanding what transpires in NN and how to improve on it.

What has not been seen may perhaps be there in the science constraining the problem.

The arrow of probability

$\mathfrak{p}(\mathbf{x}|\theta)$: Probability of x $(\mathbf{x})$ given that it belongs to some feature θ

Bayes: $$\mathfrak{p}(\theta|\mathbf{x}) = \frac{\mathfrak{p}(\mathbf{x}|\theta)\mathfrak{p}(\theta)}{\sum_{\theta'} \mathfrak{p}(\mathbf{x}|\theta')\mathfrak{p}(\theta')}$$

Predictive information is a ``momentum" where position is the boundary/initial condition. Generalization is interesting but overfitting is not.

Take $$\mathscr{H}_\theta(\mathbf{x}) = -\ln \mathfrak{p}(\mathbf{x}|\theta)$$
$$\mu_\theta = -\ln \mathfrak{p}(\theta)$$
$$\mathfrak{p}(\theta|\mathbf{x}) = \frac{\exp\{-[\mathscr{H}_\theta(\mathbf{x}) + \mu_\theta]\}}{\sum_\theta \exp\{-[\mathscr{H}_\theta(\mathbf{x}) + \mu_\theta]\}}$$

$\boldsymbol{\theta}$ being one of a discrete set (an index), vectorially

$$\mathfrak{p}(\mathbf{x}) = \frac{\exp\{-[\mathscr{H}(\mathbf{x}) + \boldsymbol{\mu}]\}}{\sum \exp\{-[\mathscr{H}(\mathbf{x}) + \boldsymbol{\mu}]\}}$$

NNs

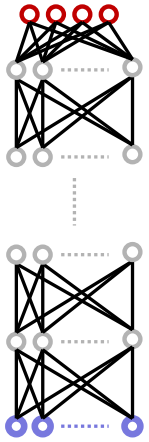
n layer neural net (feedforward): $\mathbf{f}(\mathbf{x}) = \boldsymbol{\sigma}_n \mathbf{A}_n \cdots \boldsymbol{\sigma}_1 \mathbf{A}_1 \mathbf{x}$

$\boldsymbol{\sigma}_i$:Operators (nonlinear: local, max-pool, softmax)

$\mathbf{A}_i = \mathbf{W}_i \mathbf{x} + \mathbf{b}_i$

Softmax: $\boldsymbol{\sigma}(\mathbf{x}) = \dfrac{\exp \mathbf{x}}{\sum_i \exp \theta_i}$

So, $\mathfrak{p}(\mathbf{x}) = \boldsymbol{\sigma}[-\mathscr{H} - \boldsymbol{\mu}]$



$\boldsymbol{\mu}$ *is now just a bias vector for classification probability in the final layer extracting features when using softmax.*

The Hamiltonian is computable (and has the meaning of an energy function).

Central limit theorem implies multivariate Gaussian with the form

$$\mathfrak{p}(\mathbf{x}) = \exp(h + \sum_i h_j x_i - \sum_{ij} h_{ij} x_i x_j)$$

$\mathscr{H} = -\ln \mathfrak{p}$ *is therefore a quadratic polynomial.*

Linear transformations also leave invariances intact.

Markov

Hierarchical causal construction: $\boldsymbol{\theta}_0 \mapsto \boldsymbol{\theta}_1 \mapsto \cdots \mapsto \boldsymbol{\theta}_n$

If $\mathfrak{p}(\boldsymbol{\theta}_i)(\equiv (\mathfrak{p}_i)_{\boldsymbol{\theta}})$ is determined only by causal predecessor, and the transition probability to the i^{th} level is

Markov matrix: $\mathbf{M}_i = \mathfrak{p}(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})$

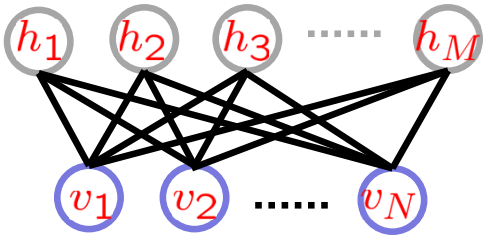
then $\mathfrak{p}_n = \mathbf{M}_n\mathbf{M}_{n-1}\cdots\mathbf{M}_1\mathfrak{p}_0$

… a Markov chain simplification.

NN implement Markove chains

Boltzmann versus wave recursive

Restricted Boltzmann



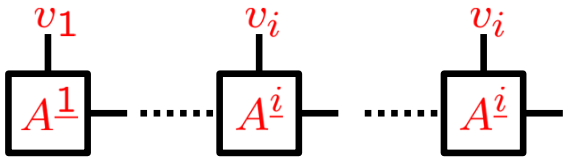
$$\mathfrak{p}(v) = \frac{\exp[-E(v)]}{\sum_v \exp[-E(v)]}$$

$$E(v) = -\sum_i a_i v_i - \sum_j \ln[1 + \exp[b_j + \sum_i v_i W_{ij}]$$

For $x \in X$ and $y \in Y$, the mutual information follows for all hidden variables:

$$I_{RBM}(X:Y) \leq I_{RBM}(X:H) \leq |H| \ln 2$$

Restricted Wave (pure/diagonal)



$$\mathfrak{p}(v) = \frac{|\psi(v)|^2}{\sum_v |\psi(v)|^2}$$

$$\psi(v) = \mathrm{Tr} \prod_i^N A^i[v_i]$$

Transformers/Recursive generators incorporate both these approaches.

Thermodynamic-information meaning

Entropy: $H(X|\mathbf{Y}) = -\sum_{x,\mathbf{y}} \mathfrak{p}(x,\mathbf{y}) \log \mathfrak{p}(x|\mathbf{y})$

This is uncertainty of true probability
+ uncertainty from finiteness of data

Energy: Kullback-Leibler divergence (relative entropy) between distributions:

$$\mathscr{D}(\mathfrak{p}(X,\mathbf{Y}) \parallel \mathfrak{q}(X,\mathbf{Y})) = \sum_{x,\mathbf{y}} \mathfrak{p}(x,\mathbf{y}) \frac{\mathfrak{p}(x|\mathbf{y})}{\mathfrak{q}(x|\mathbf{y})}$$

Fluctuation due to finiteness of data is the thermal fluctuation

The inverse temperature of a random variable $\beta_0(X) \propto 1/T$ can be defined in terms of KL divergence between the estimator at data size n, versus the probability estimator with the data removed.

Minimization of energy

Internal energy: KL divergence between target and empirical distribution $U_0(X) = \mathscr{D}(\mathfrak{p}_1(X) \,\|\, \mathfrak{p}_2(X))$

Cross entropy: $U(X) = H(\mathfrak{p}_1(X), \mathfrak{p}_2(X))$ Minimizing $U_0(X)$ is the maximum likelihood.

$\epsilon(x) = \log \tilde{\mathfrak{p}}(X)$ Self information based on relative frequency

Helmholtz free energy: $F(X) = U_0(X) - H(X)/\beta_0(X)$

Information energy Shannon entropy

Minimizing free energy is the minimum free energy principle

$H = \beta U + \log Z$

Z :Partition function

$\beta = \dfrac{\partial H}{\partial U}$

$$F = U - \frac{1}{\beta} H$$

$$= -\frac{1}{\beta} \log Z$$

$$\mathfrak{p}(x) = \frac{\exp(-\beta\epsilon(x))}{\sum_{x'} \exp(-\beta\epsilon(x'))}$$

$$-\frac{\partial U}{\partial \beta} = \langle \epsilon^2 \rangle - \langle \epsilon \rangle^2 = I(\beta)$$

Energy fluctuation to Fisher information

$$\mathfrak{p}(x) = \frac{\exp(-\beta(-\log \mathfrak{p}_B(x)))}{\sum_{x'} \exp(-\beta(-\log \mathfrak{p}_B(x)))}$$

Prior incorporation (Bayes)

Estimation and size of data (noise!)

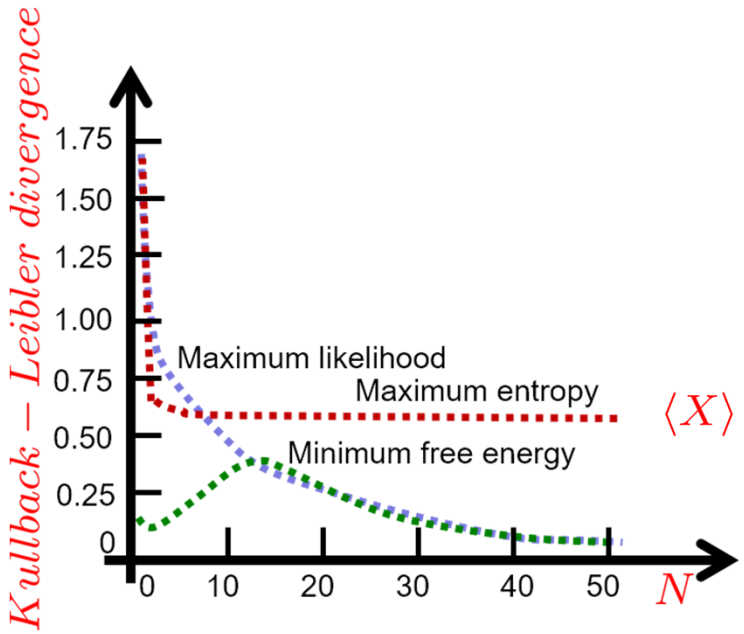
$$\mathscr{D}(\mathfrak{p}(X) \parallel \mathfrak{p}_e(X)) = \sum_x \mathfrak{p}(x) \log \frac{\mathfrak{p}(x)}{\mathfrak{p}_e(x)}$$

Data with 3 internal states

$$\mathfrak{p}(0) = 0.850$$

$$\mathfrak{p}(1) = 0.116$$

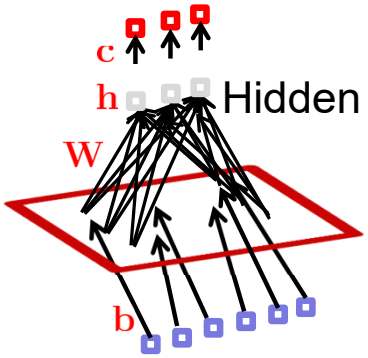
$$\mathfrak{p}(2) = 0.034$$



$$\langle X \rangle = \frac{1}{N} \sum_{d=1}^{N} X_d$$

Stochastic network as Boltzmann machine

Ising spin



At thermal equilibrium

$$\mathfrak{p}(\boldsymbol{\sigma}, T) = \frac{1}{Z} \exp[-\mathscr{H}(\boldsymbol{\sigma})/T]$$

Model: $\boldsymbol{\lambda} = \mathbf{W}, \mathbf{b}, \mathbf{c}$

$$\mathfrak{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}, h) = \frac{1}{Z_{\boldsymbol{\lambda}}} \exp[-E_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}, h)]$$

$$E_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}, \mathbf{h}) = \sum_{ij} W_{ij} h_i \sigma_j - \sum_j b_j \sigma_j - \sum_i c_i h_i$$

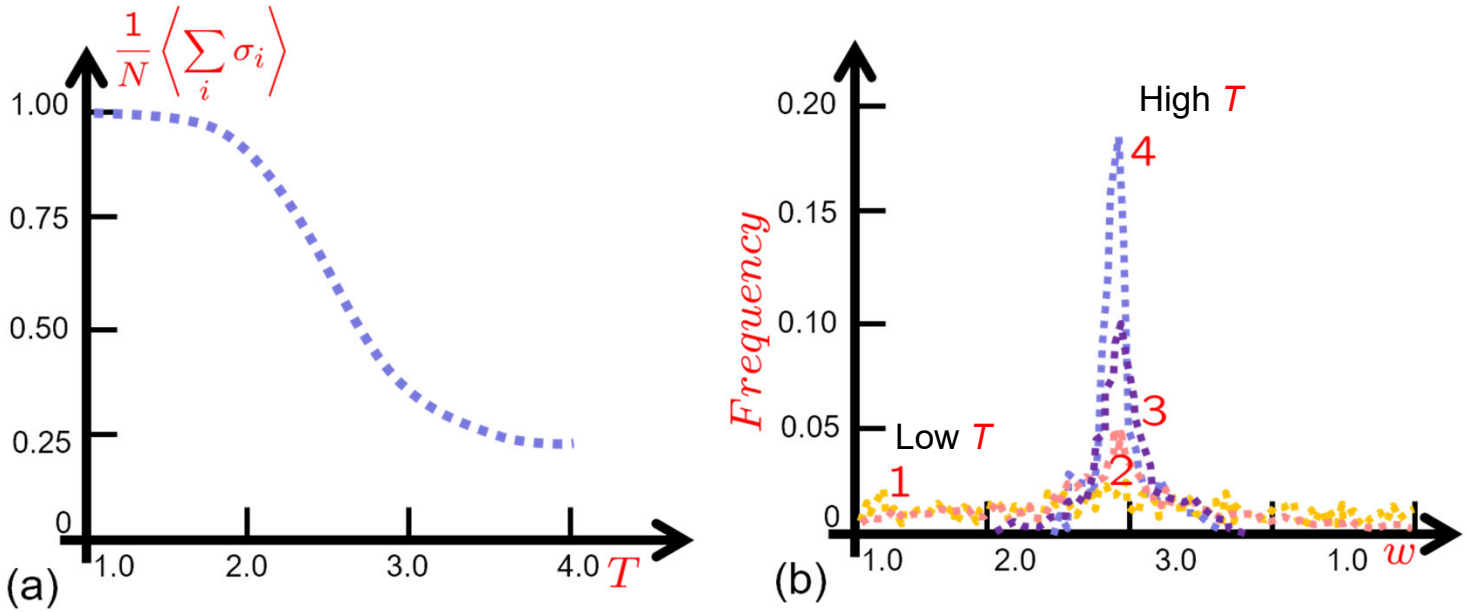
Marginalize the joint distribution

$$\mathfrak{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}) = \sum_h \mathfrak{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}, h) = \frac{1}{Z_{\boldsymbol{\lambda}}} \exp[-E'_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})]$$

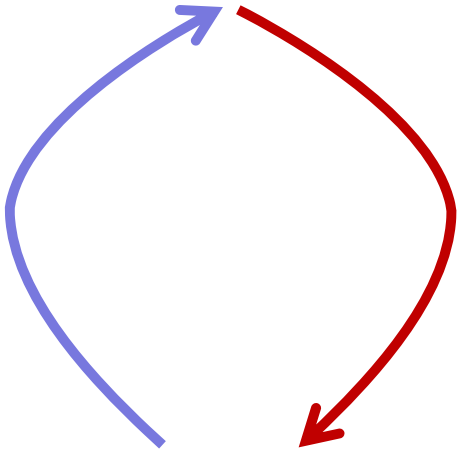
Restricted Boltzmann hidden layer implies posteriors are just products and probabilities follow from Bayes.

Now apply gradient descent.

Boltzmann machine



Physics-based minimization, together with mathematical loss minimization and regularization
Lagrangians, Hamiltonians, Transformations to other state spaces (Simplectic, Hilbert, ...)
+cross-entropy, mean square, reglurization



Data

This approach can also be fitted together with autoencoders and extract physical parameters guiding dynamics.

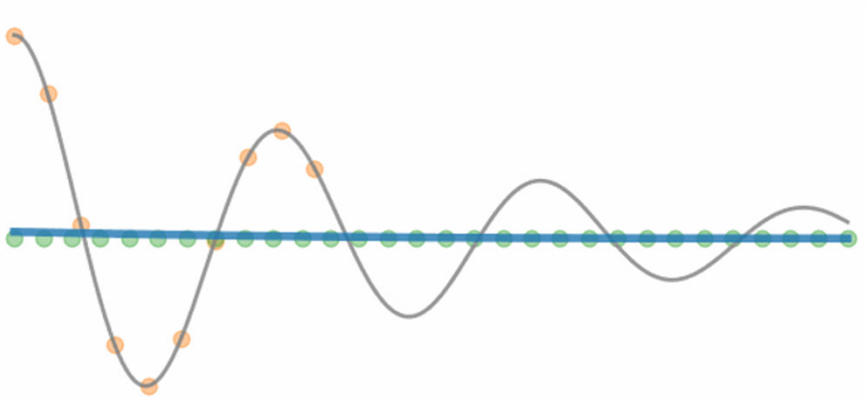
A trivial model example: Damped HO (equation known)

$$
\begin{aligned}
m d_{t^2}^2 z + \mu d_t z - k_s z &= 0 \\
z(t=0) &= 0 \\
d_t z|_{t=0} &= 0 \\
\left[\delta = \frac{\mu}{2m}\right] &< \omega_0 \text{ underdamped} \\
\omega_0 &= \left(\frac{k_s}{m}\right)^{1/2}
\end{aligned}
$$

$$
\begin{aligned}
z(t) &= 2A \exp(-\delta t) \cos(\phi + \omega t) \\
\omega &= (\omega_0^2 - \delta^2)^{1/2}
\end{aligned}
$$

Train neural network to interpolate part of the solution from a training point.

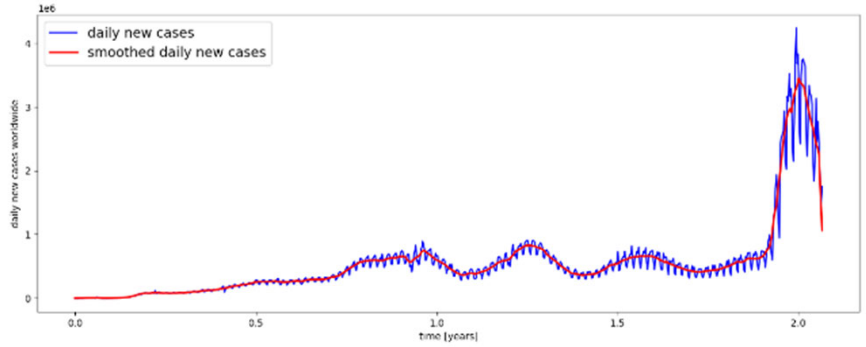
Force to extrapolate by penalising the underlying differential equation in its loss function. ($\| \hat{z} - z \|^2$)



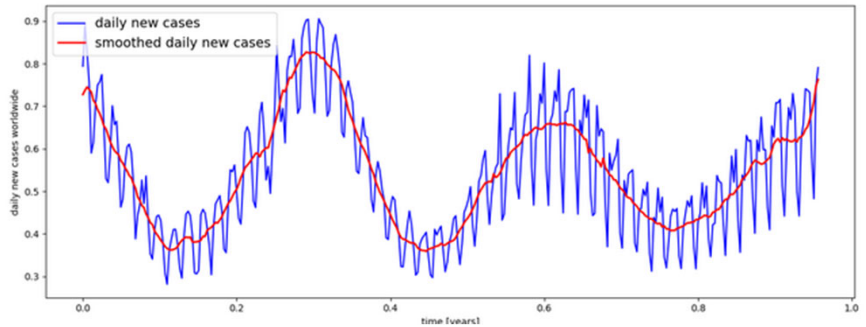
Training from past data

Covid damped oscillator (find the parameters)

```
In [4]: # plot data
plt.figure(figsize=(1100/72,400/72))
plt.plot(years,covid_world[:,1],color='blue',label='daily new cases')
plt.plot(years,covid_world_smooth,linewidth=2.0,color='red', label='smoothed daily new cases')
plt.xlabel('time [years]')
plt.ylabel('daily new cases worldwide')
plt.legend(loc='upper left',fontsize=14)

Out[4]: <matplotlib.legend.Legend at 0x2accf2c1240>
```



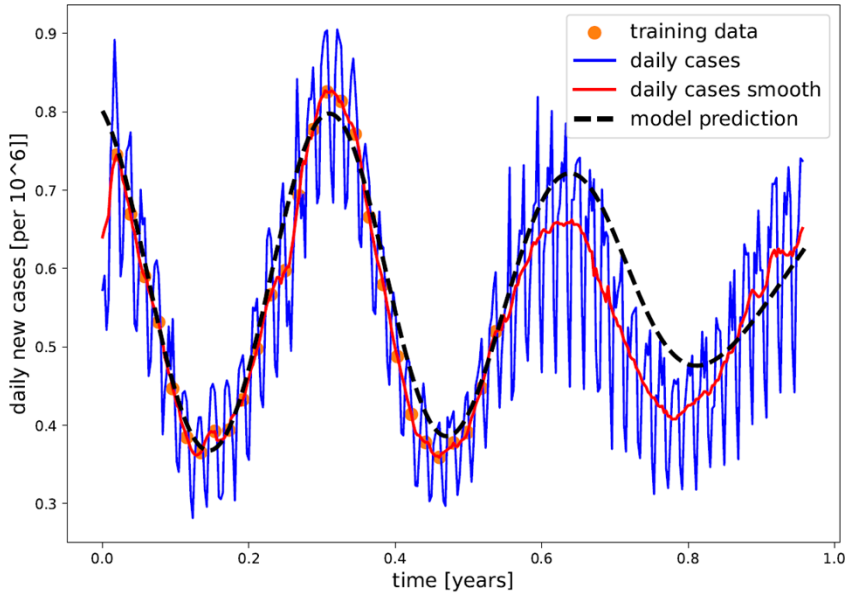

```
In [5]: # select time window of interest
d1 = 350
d2 = 700

# plot data in window of interest (normalize case numbers by 10^6)
plt.figure(figsize=(1100/72,400/72))
plt.plot((days[d1:d2]-d1)/365,covid_world[d1:d2,1]/1e06,color='blue',label='daily new cases')
plt.plot((days[d1:d2]-d1)/365,covid_world_smooth[d1:d2]/1e06,linewidth=2.0,color='red', label='smoothed daily new cases')
plt.xlabel('time [years]')
plt.ylabel('daily new cases worldwide')
plt.legend(loc='upper left',fontsize=14)

Out[5]: <matplotlib.legend.Legend at 0x2accebd29e0>
```



$$z(t) = 2A\exp(-\delta t)\cos(\phi + \omega t) + c$$

Learn $k_s, \mu, c, (m = 1)$ by training from data.

Science-constrained-PDE prediction



Lagrangians (non-canonical)

$$d_t \nabla_{\dot{q}} \mathscr{L} = \nabla_q \mathscr{L}$$

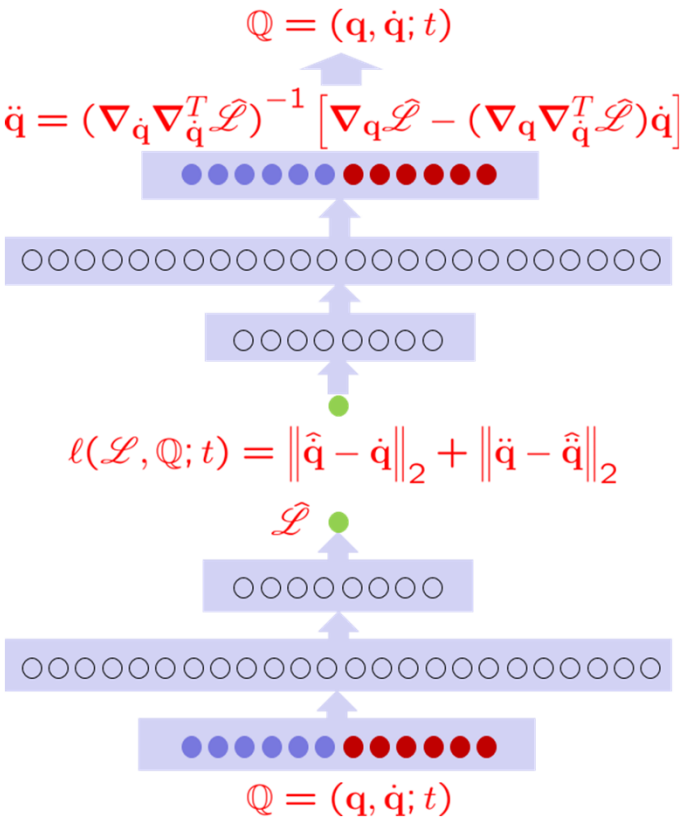
$$(\nabla_{\dot{\mathbf{q}}} \nabla_{\dot{\mathbf{q}}}^T \mathscr{L}) \ddot{\mathbf{q}} + (\nabla_{\mathbf{q}} \nabla_{\dot{\mathbf{q}}}^T \mathscr{L}) \dot{\mathbf{q}} = \nabla_{\mathbf{q}} \mathscr{L}$$

$$\therefore \quad \ddot{\mathbf{q}} = (\nabla_{\dot{\mathbf{q}}} \nabla_{\dot{\mathbf{q}}}^T \mathscr{L})^{-1} \left[\nabla_{\mathbf{q}} \mathscr{L} - (\nabla_{\mathbf{q}} \nabla_{\dot{\mathbf{q}}}^T \mathscr{L}) \dot{\mathbf{q}} \right].$$

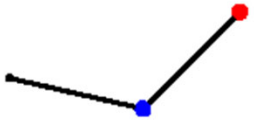
If $\dot{\mathbf{q}}$ is known, then one has determined $(\mathbf{q}, \dot{\mathbf{q}})$

1. Generate training and test data using an analytic solution of the Lagrangian formulation. Incorporate noise.
2. Use the test data, employ the loss function on the target versus predicted $(\dot{\mathbf{q}}, \ddot{\mathbf{q}})$ to optimize the network, as also three different derivatives: $\partial/\partial \mathbf{q}$, $\partial^2_{\mathbf{q}\dot{\mathbf{q}}}$ and $\partial^2_{\mathbf{q}^2}$.
3. Minimize loss function to obtain maximum accuracy.
4. For a general problem, employ the network to now obtain $\ddot{\mathbf{q}}$.
5. From $\ddot{\mathbf{q}}$, obtain the dynamic trajectory.

The network finds the Lagrangian implicitly.



M.Crammer et al., ICLR 2020 Workshop on Deep Differential Equations



Symplectic Hamiltonian dynamics

$$\partial_t p = -\partial_q \mathscr{H}, \quad \partial_t q = +\partial_p \mathscr{H}$$

Phase space:

$\mathbf{x} = (p, q)$ → $\mathbf{z} = (P, Q)$

Symplectic:

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Symplectic condition: $(\nabla_{\mathbf{x}}\mathbf{z}) J (\nabla_{\mathbf{x}}\mathbf{z})^T = J$

$$\therefore \quad \partial_t \mathbf{z} = \nabla_{\mathbf{z}} \mathscr{K}(\mathbf{z}) J \quad , \quad \mathscr{K}(\mathbf{z}) = \mathscr{H} \cdot \mathbf{x}(\mathbf{z})$$

Symplectic flow:

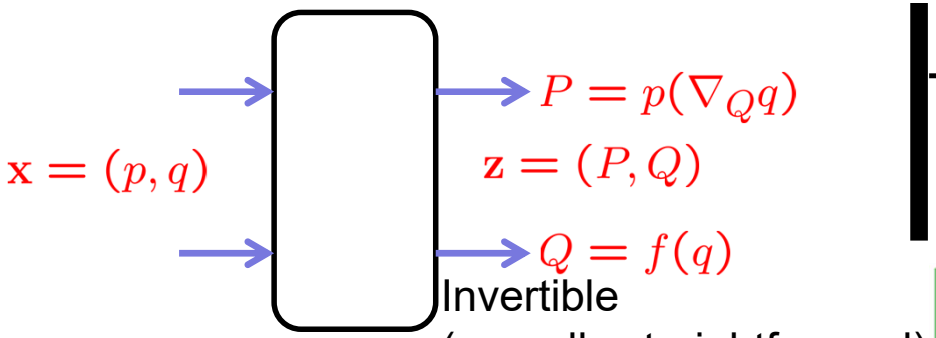
$$\partial_t \mathbf{x} = \nabla_{\mathbf{x}} \mathscr{H}(\mathbf{x}) J$$

$\mathbf{z}$ is a latent phase space preserving Hamiltonian dynamics.

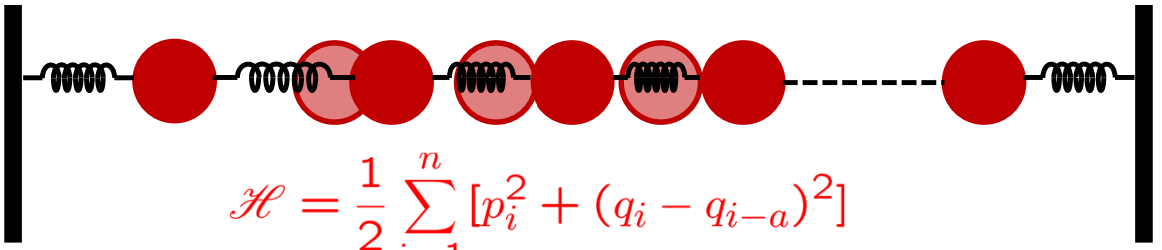
$\mathfrak{p}(\mathbf{x}) = \exp[-\beta \mathscr{H}(\mathbf{x})]$ preserved

Making dynamics much simpler to solve.

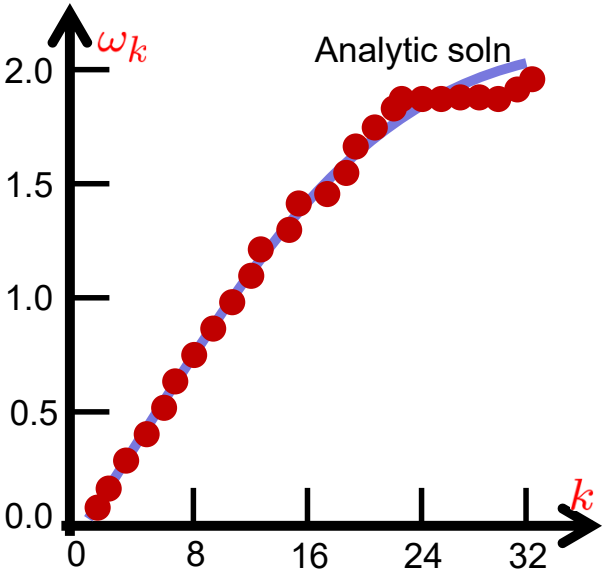
Symplectic neural transformation

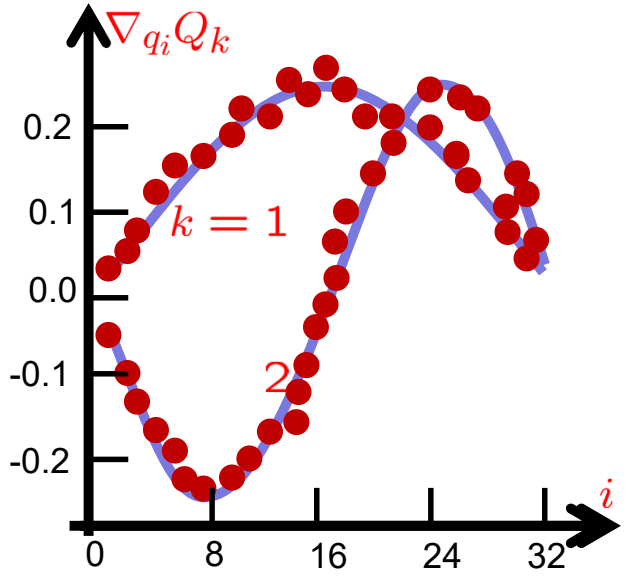


Invertible
(neurally straightforward)




```
In [7]:  Q = forward(params,q)
         _, vjp=jax.vjp(lambda Q: inverse(params, Q), Q)
         P= vjp(p)(0)
```





Summary

NNs, if carefully designed and used, are a new useful tool for complexity.

Noise, fluctuations and randomness are useful so long as we maintain information flow.
Need to keep information together during aggregation and dimensionality reduction.

NNs and probabilistic approaches do come closer to the fundamental principles by accepting the statistical variations as a given and a more realistic physical mapping of reality so long as we keep science in it.

Nonlinearities and dimensionality reduction are subject to laceration by Occam's razor. This is Mencken's rule that draws on absence of sufficient statistics.

Incorporating science-constrained rules into the problem definition is a major help.

There is still much unexplored and unlearned from what short- and long-range interactions is captured in nature's dynamics which is not captured in the NNs and Bayesian probabilism.