CeMM

SCIENCE IS OUR MEDICINE

CENTER FOR MEDICAL DATA SCIENCE
MEDICAL UNIVERSITY OF VIENNA
Institute of Artificial Intelligence

ÖAW

GSNs: generative stochastic networks

Animesh Awasthi
Predoctoral Fellow
GenAI Reading Club | 25th September 2024

BOCK LAB

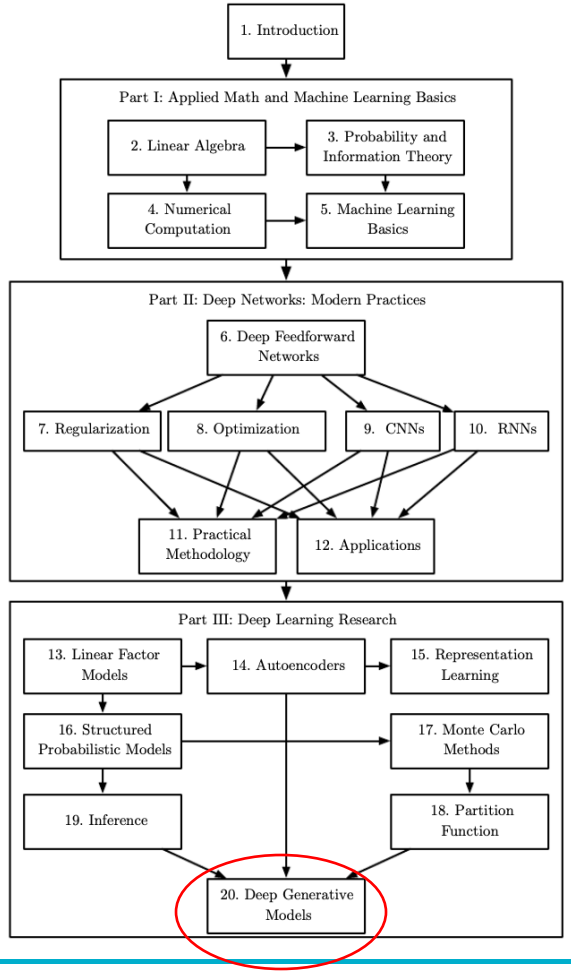
Research Center for Molecular Medicine
of the Austrian Academy of Science

www.cemm.at

Introduction

Deep Learning

Ian Goodfellow
Yoshua Bengio
Aaron Courville



 https://www.deeplearningbook.org/ (pdf shared on GH)

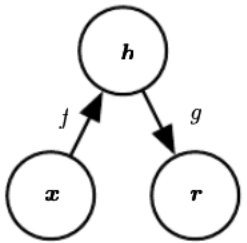
CeMM

Introduction

Generative stochastic networks or **GSNs** are generalizations of **denoising autoencoders** that include latent variables h in the generative Markov chain, in addition to the visible variables (usually denoted x).

What is an Autoencoder?

An autoencoder is a neural network that is trained to attempt to copy its input to its output. Internally, it has a hidden layer h that describes a code used to represent the input.



Stochastic mappings: $p_{encoder}(h \mid x)$, $p_{decoder}(x \mid h)$

Undercomplete Autoencoder

- h is constrained to have smaller dimension than x
- L(x, g(f (x)))

Introduction - Autoencoders

Overcomplete Autoencoders

An autoencoder fails to learn anything useful if the hidden code is allowed to have equal to or greater dimensions than the input

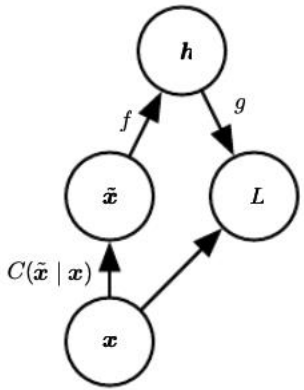
Regularized Autoencoders

Rather than limiting the model capacity, we can use a loss function to encourage the model to have other properties besides copying the input to its output

Denoising Autoencoders

$L(\boldsymbol{x}, g(f(\boldsymbol{x})))$ $\quad$ $L(\boldsymbol{x}, g(f(\tilde{\boldsymbol{x}}))),$

Where $\tilde{x}$ is corrupted data obtained through a corruption process $C(\tilde{\mathbf{x}} \mid \mathbf{x})$



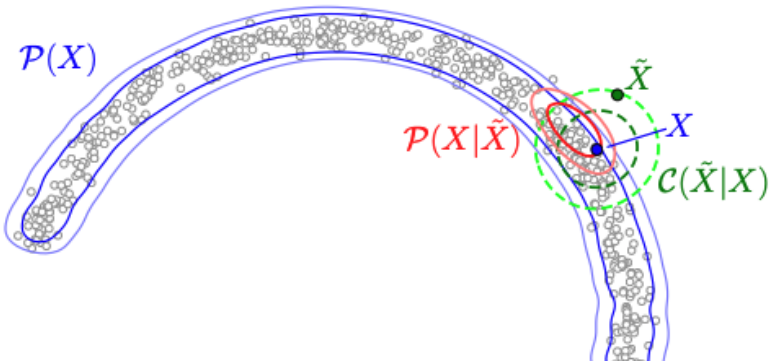
x

x̃

CeMM

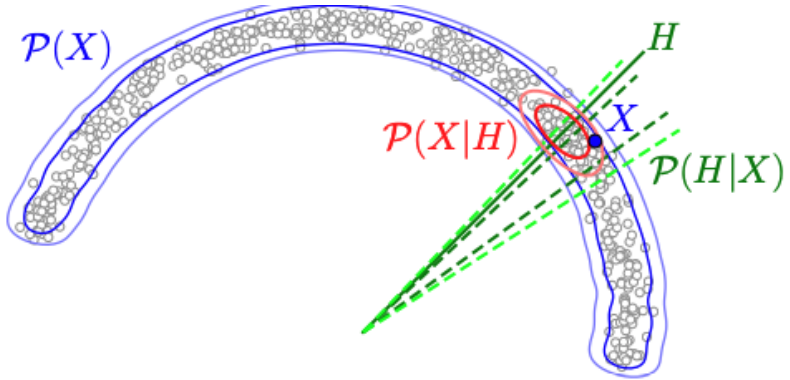
Introduction - GSNs

Generative stochastic networks or **GSNs** are generalizations of **denoising autoencoders** that include latent variables h in the generative Markov chain, in addition to the visible variables (usually denoted x).



A denoising auto-encoder defines an estimated Markov chain

- first samples a corrupted X˜ from C(X˜|X)
- samples a reconstruction from Pθ(X|X˜), trained to estimate the ground truth P(X|X˜)



More generally, a GSN allows the use of arbitrary latent variable H, where H is the angle about the origin.

Contributions - GSNs

1. Intuition: Alternate ways of training unsupervised models, avoiding intractable sums or maximization

2. Training Framework: Generalized denoising autoencoders by introducing latent variables in the framework. Estimates the data-generating distribution by parameterizing the transition operator of a Markov chain

3. General Theory: Proof that if our estimate P(X|h) is consistent, then the stationary distribution of the resulting chain is a consistent estimator of the data generating density P(X)

4. Consequences of theory: GSNs allow for the modeling of structured output, i.e. modeling the conditional distribution of high-dimensional output with complex multi-modal joint distribution.

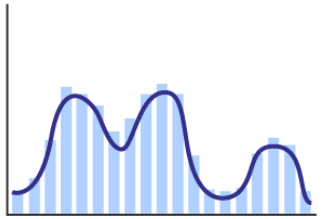
5. Example application: Create a trainable deep GSN whose graph resembles the one followed by Gibbs sampling in DBMs.

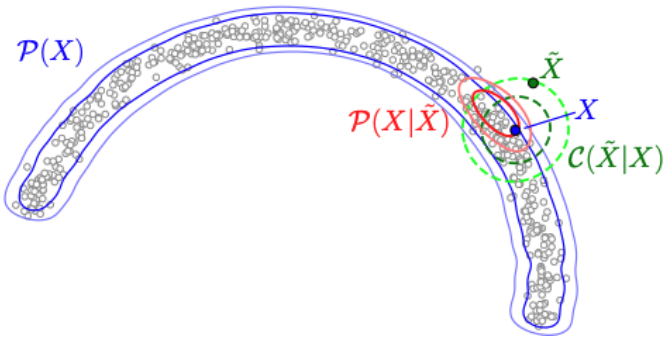
6. Dependency Networks: It allows us to provide a novel justification for dependency networks and a proper joint distribution between all visible variables

CeMM

1. Intuition

- Summing over too many major modes (in P(x) or P(x|h) is intractable for many graphical models
- Similarly for sampling from P(y,h|x) where y and h are high-dimensional
- MCMC works with a small number of non-negligible models
- The assumption for GSNs: the effectiveness of the function approximation
- Estimate the transition operator of the Markov Chain $P(x_t|\bar{x}_{t-1})$ or $P(x_t, h_t|\bar{x}_{t-1}, h_{t-1})$
- Since each step is local, the transition distribution will often include only a small number of modes





CeMM

2./3. Training Framework and General Theory

$P(X)$

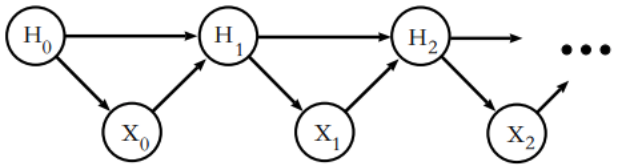
$\mathcal{C}(\tilde{X}|X)$

Reconstruction distribution using Bayes rule:

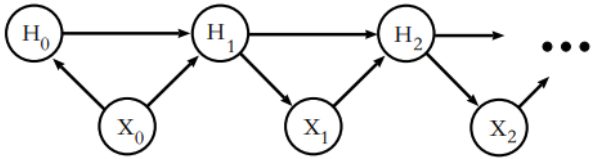
$$P(X|\tilde{X}) = \frac{1}{z}\mathcal{C}(\tilde{X}|X)P(X)$$

Generalizing and parameterizing the denoising autoencoder to GSNs

$$H_{t+1} \sim P_{\theta_1}(H|H_t, X_t)$$
$$X_{t+1} \sim P_{\theta_2}(X|H_{t+1}).$$



Theorem 2. *Let $(H_t, X_t)_{t=0}^{\infty}$ be the Markov chain defined by the following graphical model.*



If we assume that the chain has a stationary distribution $\pi_{H,X}$, and that for every value of (x, h) we have that

- *all the $P(X_t = x|H_t = h) = g(x, h)$ share the same density for $t \geq 1$*
- *all the $P(H_{t+1} = h|H_t = h', X_t = x) = f(h, h', x)$ share the same density for $t \geq 0$*
- *$P(H_0 = h|X_0 = x) = P(H_1 = h|X_0 = x)$*
- *$P(X_1 = x|H_1 = h) = P(X_0 = x|H_1 = h)$*

then for every value of (x, h) we get that

- *$P(X_0 = x|H_0 = h) = g(x, h)$ holds, which is something that was assumed only for $t \geq 1$*
- *$P(X_t = x, H_t = h) = P(X_0 = x, H_0 = h)$ for all $t \geq 0$*
- *the stationary distribution $\pi_{H,X}$ has a marginal distribution π_X such that $\pi(x) = P(X_0 = x)$.*

Those conclusions show that our Markov chain has the property that its samples in X are drawn from the same distribution as X_0.

CeMM

4. Consequence of theory

Handling missing inputs or structured output: **clamp** the observed inputs and then apply the Markov chain with the constraint that the observed inputs are fixed and **not resampled** at each time step, whereas the unobserved inputs are **resampled** each time, **conditioned on the clamped inputs**

Proposition 1. **If** *a subset* $x^{(s)}$ *of the elements of* X *is kept fixed (not resampled) while the remainder* $X^{(-s)}$ *is updated stochastically during the Markov chain of Theorem 2, but using* $P(X_t|H_t, X_t^{(s)} = x^{(s)})$, **then** *the asymptotic distribution* π_n *of the Markov chain produces samples of* $X^{(-s)}$ *from the conditional distribution* $\pi_n(X^{(-s)}|X^{(s)} = x^{(s)})$.

Extended to structured output

This method of dealing with missing inputs can be immediately applied to structured outputs. If $X^{(s)}$ is viewed as an "input" and $X^{(-s)}$ as an "output", then sampling from $X_{t+1}^{(-s)} \sim P(X^{(-s)}|f((X^{(s)}, X_t^{(-s)}), Z_t, H_t), X^{(s)})$ will converge to estimators of $P(X^{(-s)}|X^{(s)})$. This still

CeMM

5. Experimental Example of GSN

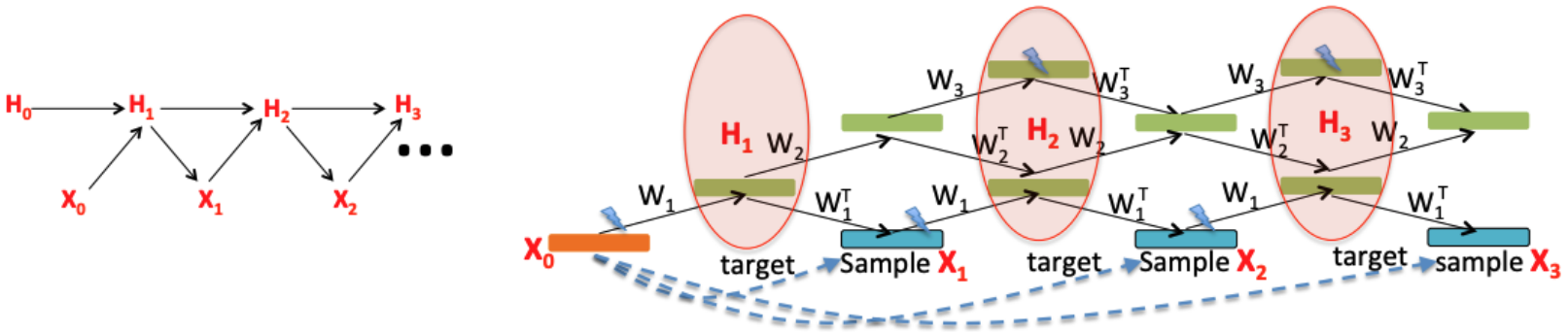
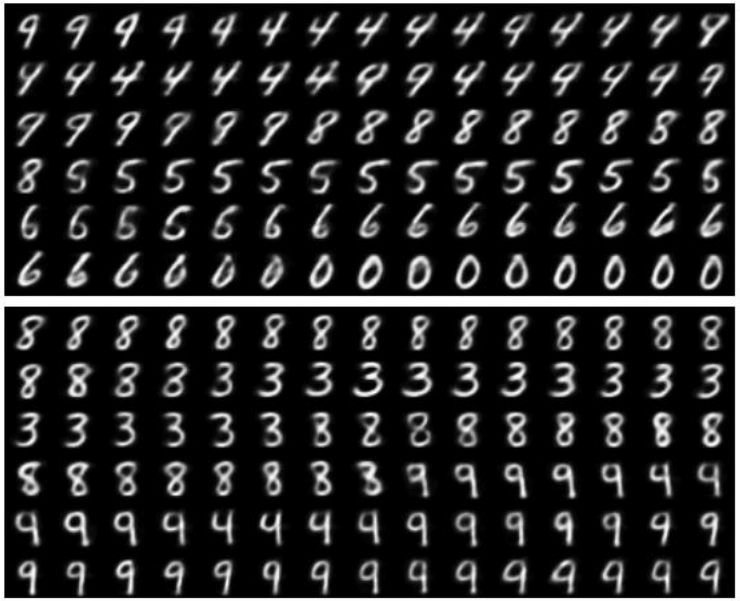


Figure 2. Left: Generic GSN Markov chain with state variables X_t and H_t. *Right:* GSN Markov chain inspired by the unfolded computational graph of the Deep Boltzmann Machine Gibbs sampling process, but with backprop-able stochastic units at each layer. The training example $X = x_0$ starts the chain. Either odd or even layers are stochastically updated at each step. All x_t's are corrupted by salt-and-pepper noise before entering the graph (lightning symbol). Each x_t for $t > 0$ is obtained by sampling from the reconstruction distribution for that step, $P_{\theta_2}(X_t|H_t)$. The walkback training objective is the sum over all steps of log-likelihoods of target $X = x_0$ under the reconstruction distribution. In the special case of a unimodal Gaussian reconstruction distribution, maximizing the likelihood is equivalent to minimizing reconstruction error; in general one trains to maximum likelihood, not simply minimum reconstruction error.

CeMM

Conclusion



MNIST Dataset

Toronto Face Database

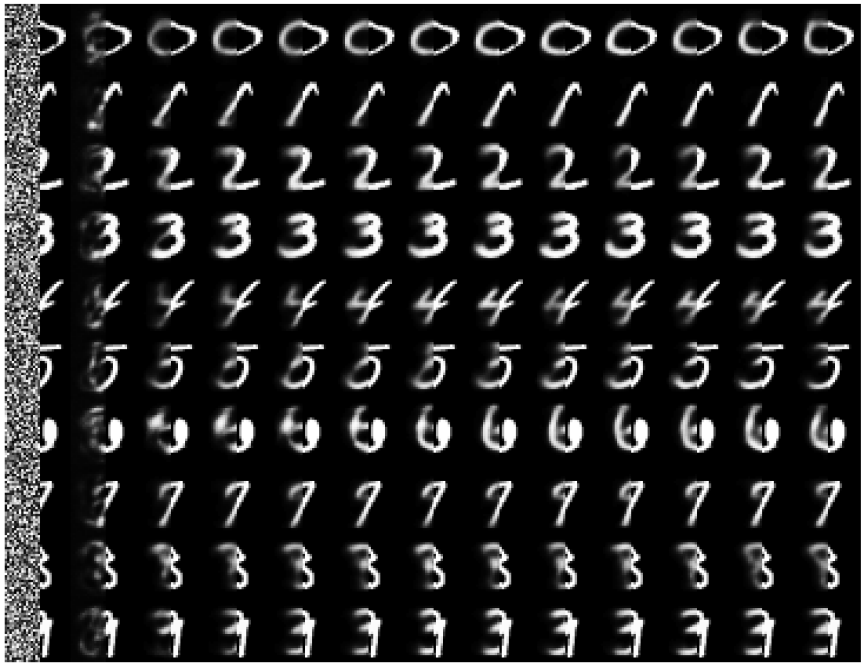
6. Dependency Networks

Dependency Networks as GSNs

Proposition 2. *If the above GSN Markov chain has a stationary distribution, then the dependency network defines a joint distribution (which is that stationary distribution), which does not have to be known in closed form. Furthermore, if the conditionals are consistent estimators of the ground truth conditionals, then that stationary distribution is a consistent estimator of the ground truth joint distribution.*

Resources

Paper: https://proceedings.mlr.press/v32/bengio14.html
Full paper: https://arxiv.org/abs/1503.05571
DL Book: https://www.deeplearningbook.org/



CeMM