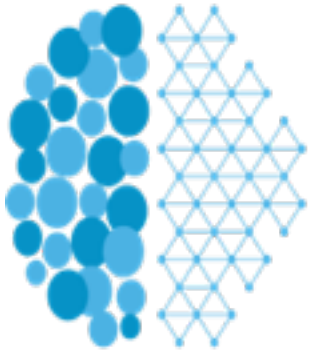
Capturing Dependencies Implicitly

Yoshua Bengio

July 14th, 2018

ICML'2018 Workshop on Deep Generative Models



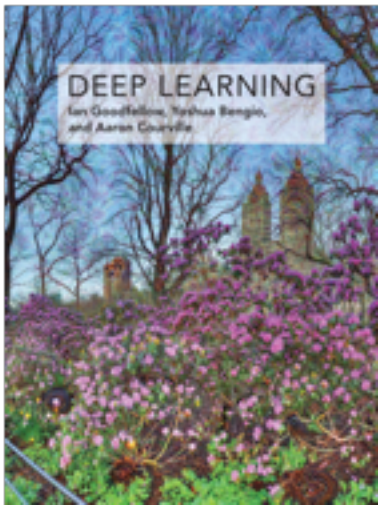
MILA

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

Université de Montréal

IVADO

PLUG: **Deep Learning**, MIT Press book is out, chapters will remain online



Still Far from Human-Level AI

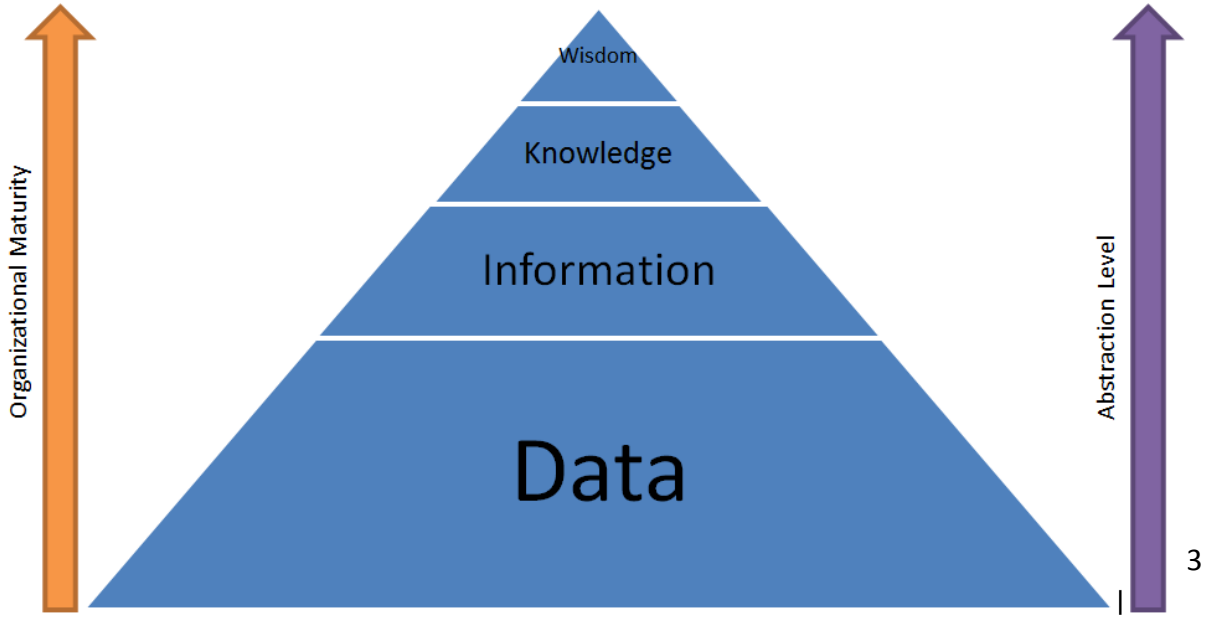
- Industrial successes mostly based on **supervised** learning

- Learning superficial clues, not generalizing well enough outside of training contexts, easy to fool trained networks:
  - Current models cheat by picking on surface regularities

Learning Multiple Levels of Abstraction

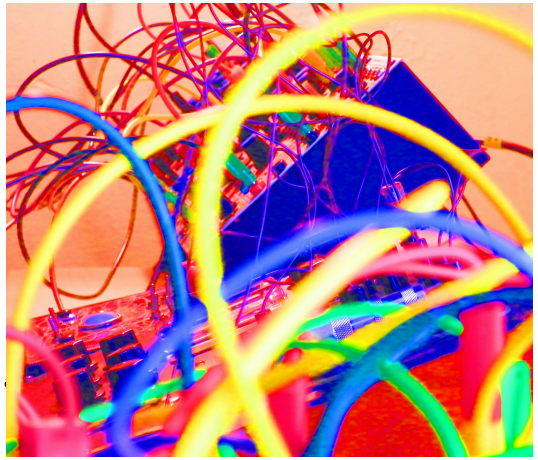
(Bengio & LeCun 2007)

- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions **disentangle the factors of variation**, which allows much easier generalization and transfer



Invariance and Disentangling

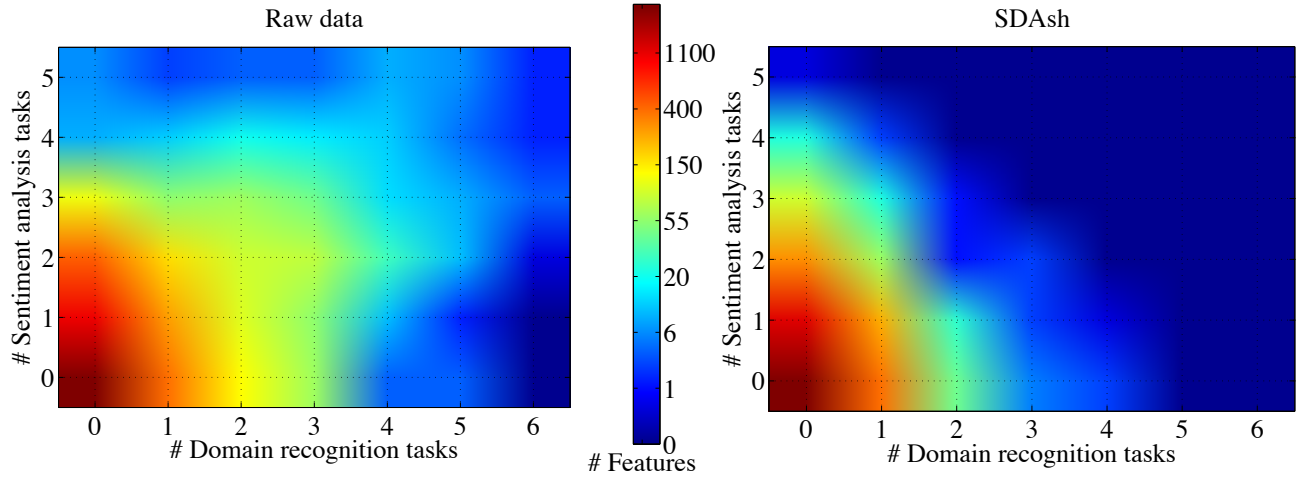
- Invariant features
- Which invariances?
- Alternative: learning to disentangle
- Good disentangling →
  avoid the curse of dimensionality:
  Dependencies are "simple" when the data is projected in the right abstract space



Disentangling from denoising objective

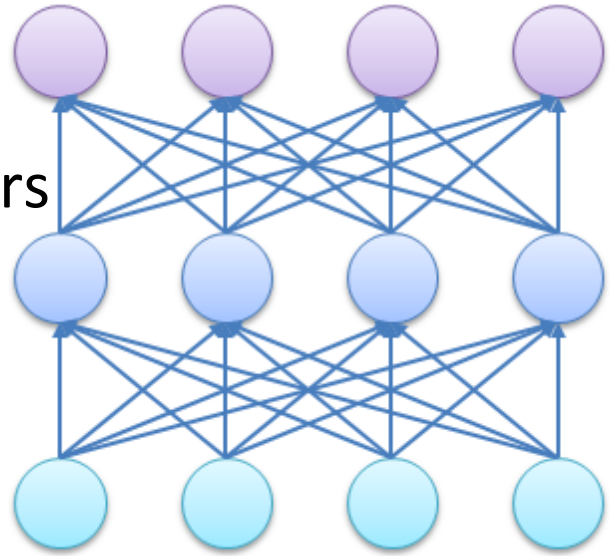
(Glorot, Bordes & Bengio ICML 2011)

- Early deep learning research already is looking for possible disentangling arising from unsupervised learning of representations
- Experiments on stacked denoising auto-encoders with ReLUs, on BoW text classification
- Features tend to specialize to either sentiment or domain



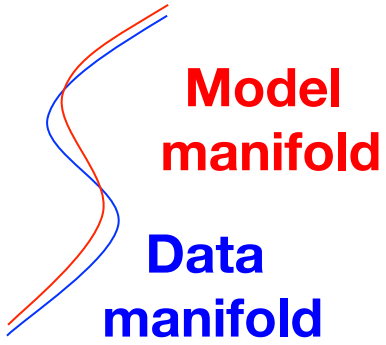
How to Discover Good Disentangled Representations

- How to discover abstractions?
- What is a good representation? ***(Bengio et al 2013)***
- Need clues (= priors) to help **disentangle** the underlying factors, such as
  - Spatial & temporal scales
  - Marginal independence
  - Simple dependencies between factors
    - *Consciousness prior*
  - Causal / mechanism independence
    - *Controllable factors*



What's wrong with standard maximum likelihood?

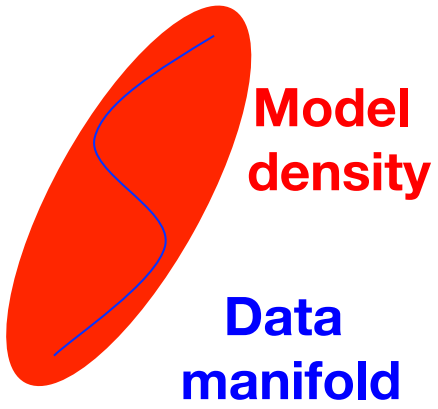
- Pay a huge price for not putting probability mass at even a single training example, even if the data manifold and model manifold are very close.



What's wrong with standard maximum likelihood?

1. Pay a huge price for not putting probability mass at even a single training example, even if the data manifold and model manifold are very close.

   - So MLE makes the model distribution very fat and conservative



2. Often requires an explicit and marginalizable formulation of the density, precludes powerful estimation of mutual information

3. Another problem is that MLE measures error bits in pixel space whereas humans really care about errors in abstract space, so we would like loss measured in learned latent space

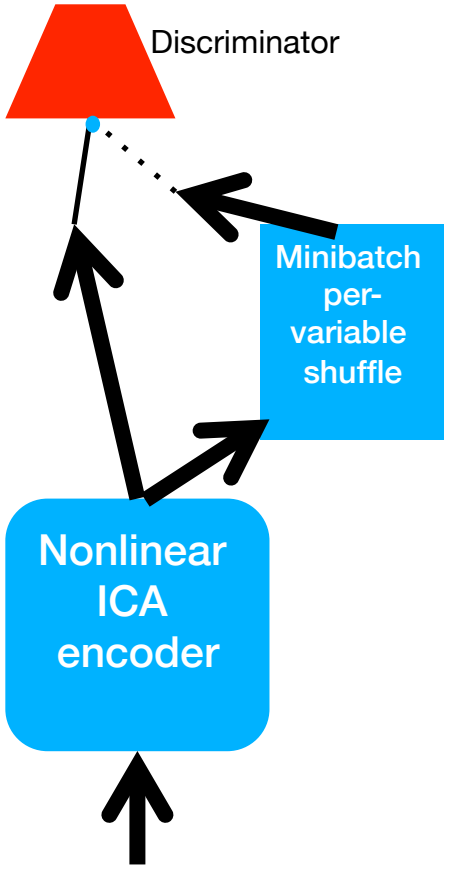
Using a discriminator to optimize **independence**, mutual information or entropy

Brakel & Bengio ArXiv:1710.05050

- Train a discriminator to separate between pairs (A,B) coming from P(A,B) and pairs coming from P(A) P(B)

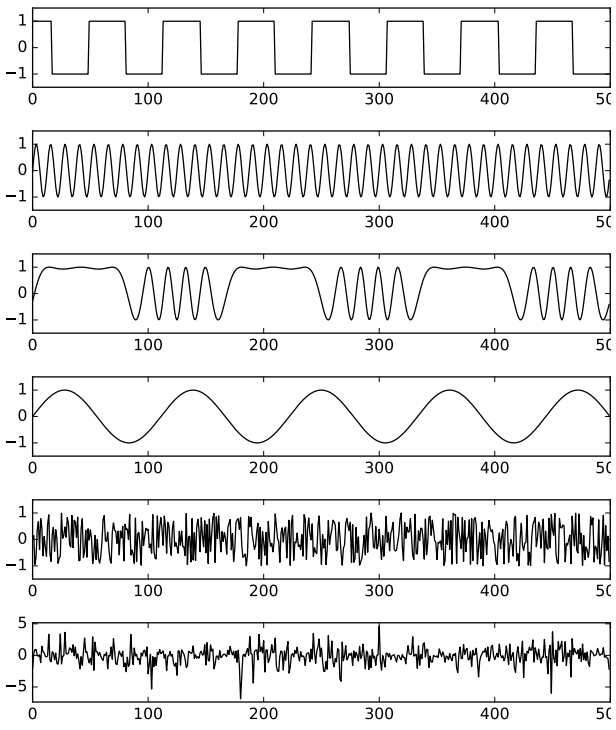
- Generalize this to measuring **independence** of all the outputs of a representation function (encoder). Maximize independence by backpropagating the independence score into the encoder

→ NON-LINEAR ICA.

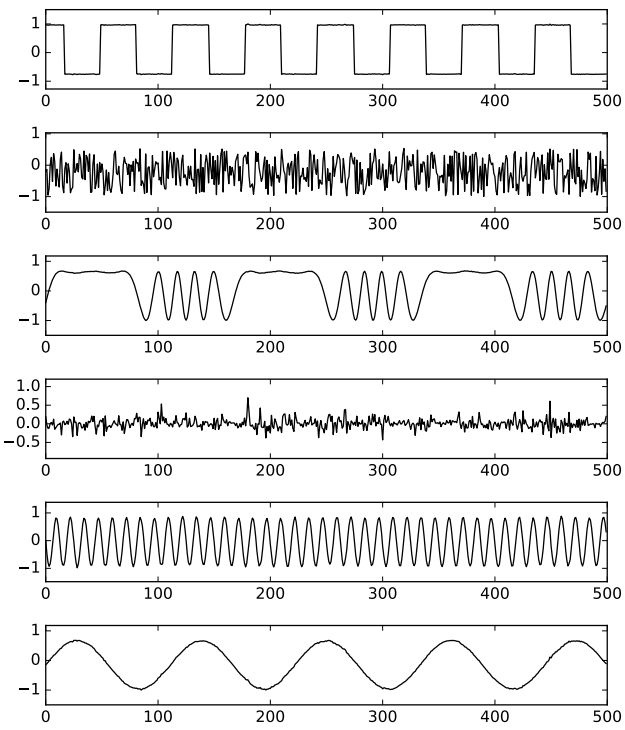


Non-Linear Independent Component Analysis Results

- Sources were either mixed linearly or non-linearly, independent components recovered in both cases

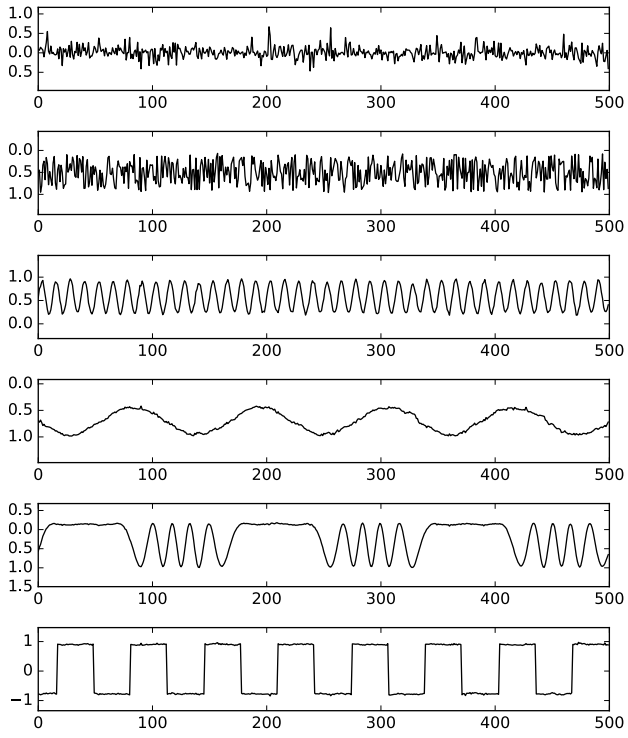


(a) Source signals.



(b) Anica reconstructions $\rho_{\max} = .997$.

Linearly mixed



(a) Anica PNL reconstructions $\rho_{\max} = .997$.

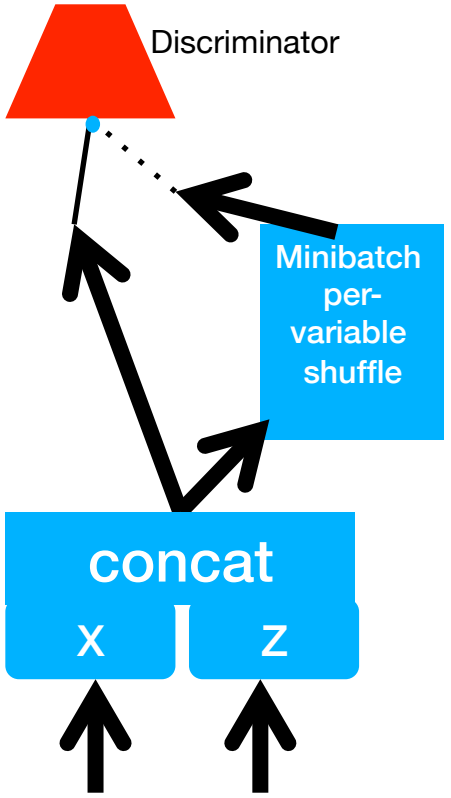
Nonlinearly mixed

Using a discriminator to optimize independence, mutual information or entropy

MINE: Mutual Information Neural Estimator

Belghazi et al ArXiv:1801.04062

Same architecture, but with a twist in the training objective which provides an asymptotically consistent estimator of mutual independence



Mutual information, KL divergence and Donsker-Varadhan Representation

[Belghazi et. al., 2018]

Mutual information: measure of dependence btwn 2 variables

$$I(X;Z) = \mathcal{D}_{KL}(\mathbb{P}_{X,Z} || \mathbb{P}_X \otimes \mathbb{P}_Z) = \mathbb{E}_{\mathbb{P}_{X,Z}} \left[\log \left(\frac{p(x,z)}{p(x)p(z)} \right) \right]$$

$$I(X;Z) = H(X) + H(Z) - H(X,Z) = D_{KL}(\mathbb{P}_{XZ} \,||\, \mathbb{P}_X \otimes \mathbb{P}_Z)$$

(Donsker & Varadhan, 1983):

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{T:\Omega \to \mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])$$

Optimal T:

$$T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}} + C$$

With suboptimal T:

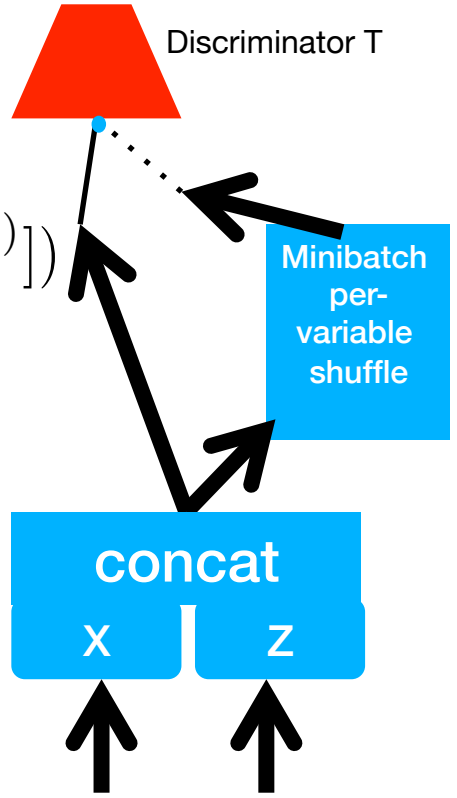
$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) \geq \sup_{T \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])$$

MINE: Estimator of MI

Given two r.v. X & Z and samples of their joint & marginals:

$$\widehat{I(X;Z)}_n = \mathbb{E}_{\hat{\mathbb{P}}^{(n)}_{XZ}}[T_{\hat{\theta}_n}(x,z)] - \log(\mathbb{E}_{\hat{\mathbb{P}}^{(n)}_{X} \otimes \hat{\mathbb{P}}^{(n)}_{Z}}[e^{T_{\hat{\theta}_n}(x,z)}])$$

where discriminator T is optimized to maximize the rhs

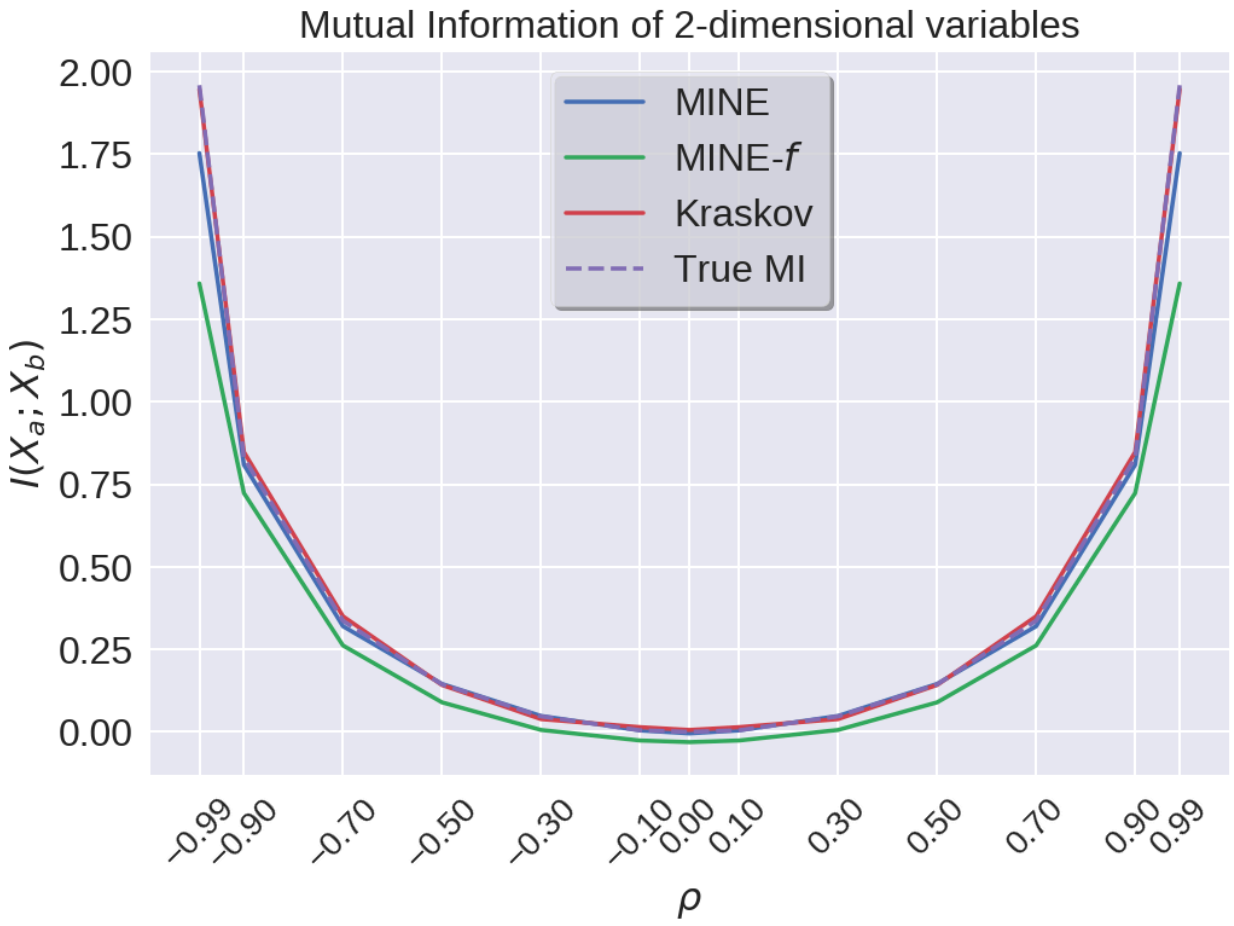


MINE: Consistency

Theorem: there exists a neural net architecture such that for all $\epsilon > 0$ there exists an integer *N* s.t.

$$\forall n \geq N, \quad |I(X,Z) - \widehat{I(X;Z)}_n| \leq \epsilon \text{ with probability one}$$

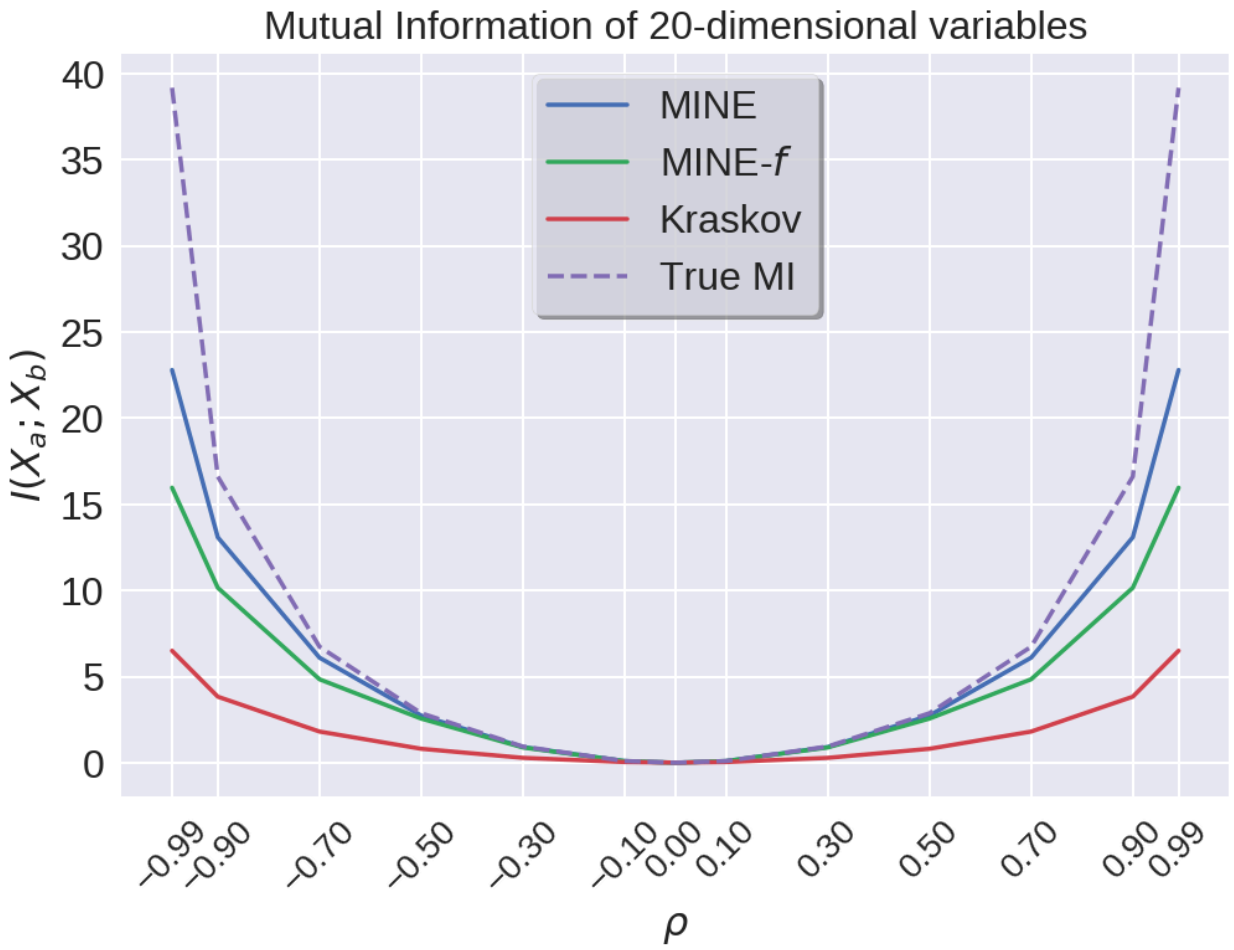
Demonstration of estimation



MI between 2 Gaussians

[Belghazi et. al., 2018]

Demonstration of estimation



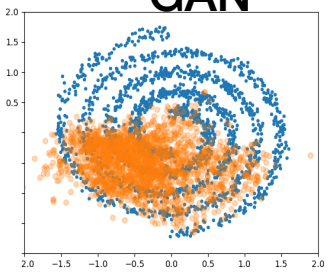
MI between 2 Gaussians

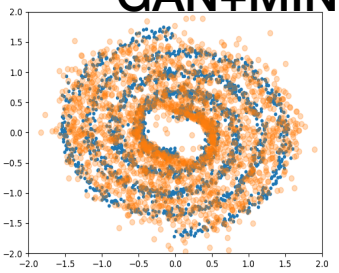
[Belghazi et. al., 2018]

Maximizing ENTROPY: avoid GAN mode dropping by max MI(X,Z)

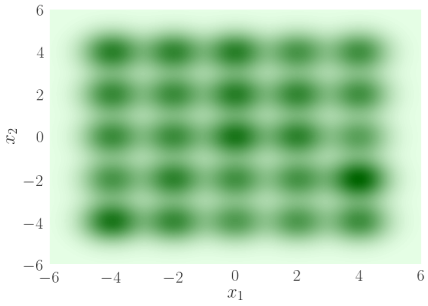
GAN

GAN+MINE





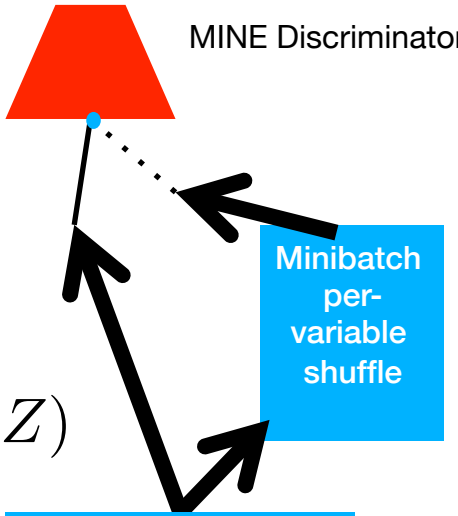
Ground Truth



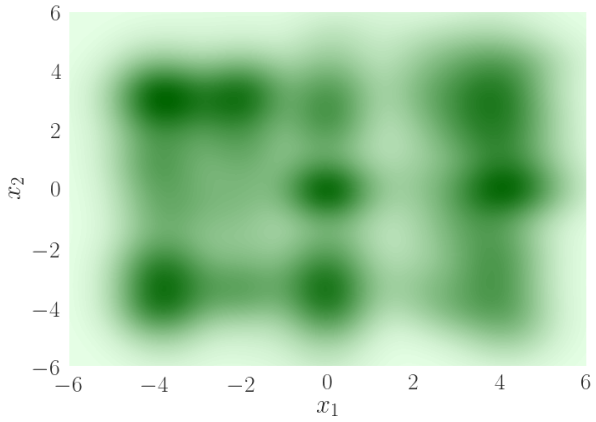
since *H(X|Z)=0*

I(X;Z)=H(X)

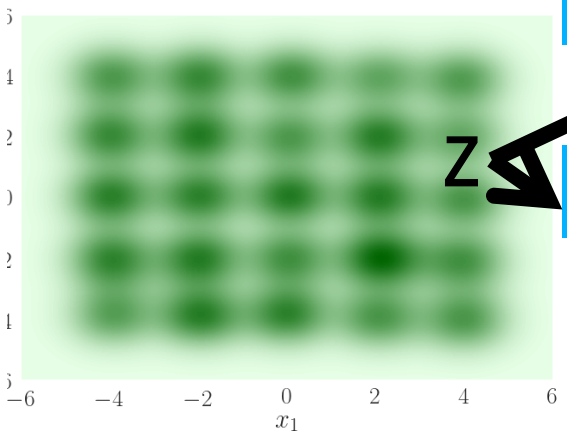
$$I(X;Z) := H(X) - H(X \mid Z)$$

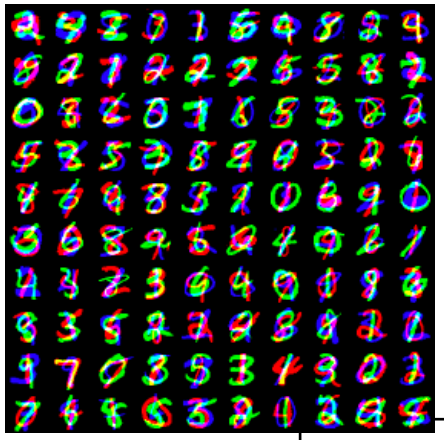


GAN



GAN+MINE





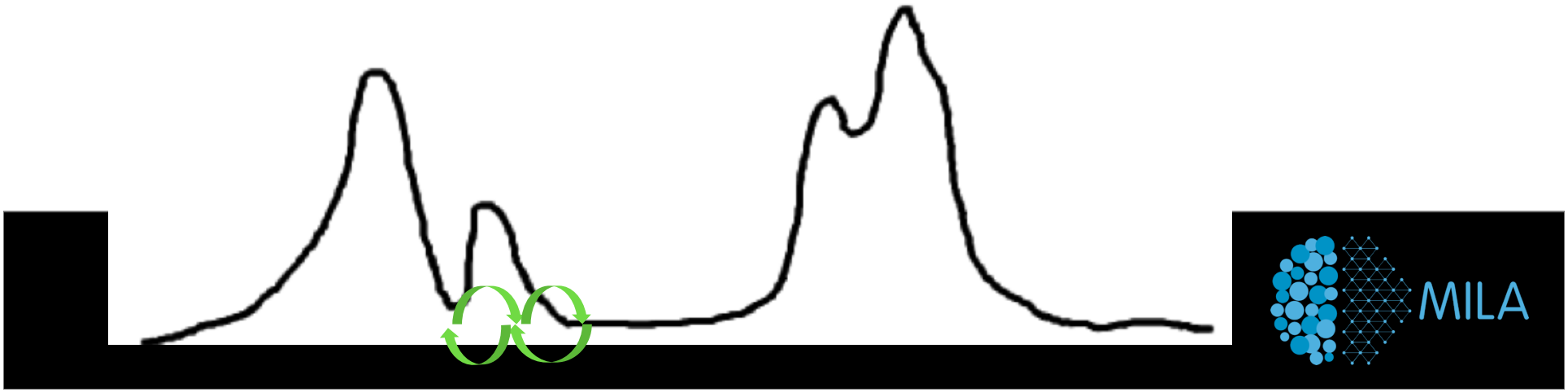
Maximizing entropy at the output of a neural net (stacked MNIST)

	Modes (max 1000)	$\mathcal{D}_{KL}(\mathbb{P}_Y \| \mathbb{Q}_Y)$
DCGAN	99	3,4
ALI	16	5,4
Unrolled GAN	48,7	4,32
VEEGAN	150	2,96
PacGAN	1000	0,6
DCGAN+MINE	1000	0,5

[Belghazi et. al., 2018]

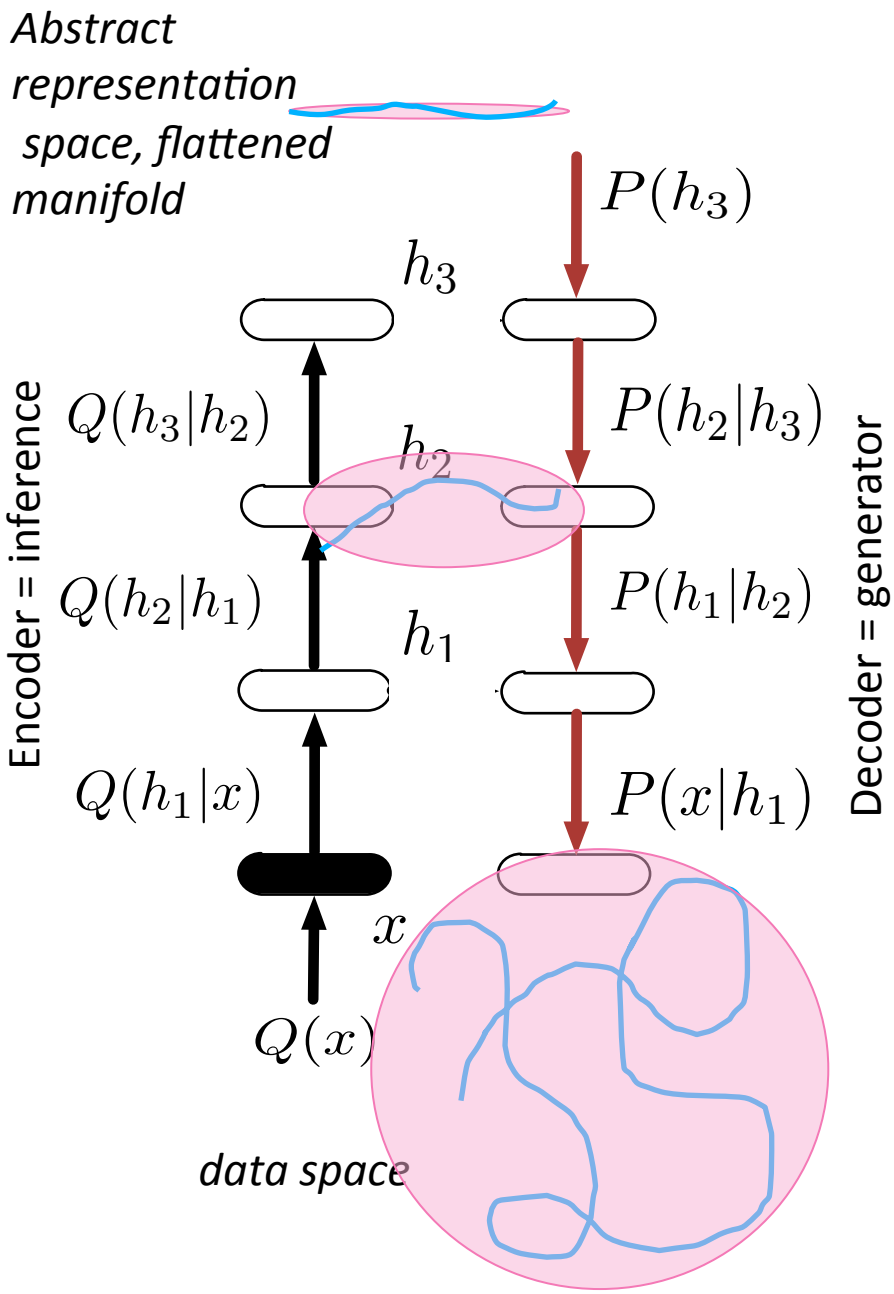
Undirected Graphical Models

- Learning these models involves two fundamental goals
  - The model must place probability mass (i.e. lower the energy function) where the data is located.
  - Remove probability mass (i.e. raise the energy function) elsewhere.
- Probability modes where there is no data are known as **spurious modes**.
- Fundamental goal of learning is to hunt down these spurious modes and remove them.



Encoders and generators as iterated transformations between distributions

Can we share the same mechanism at each step?

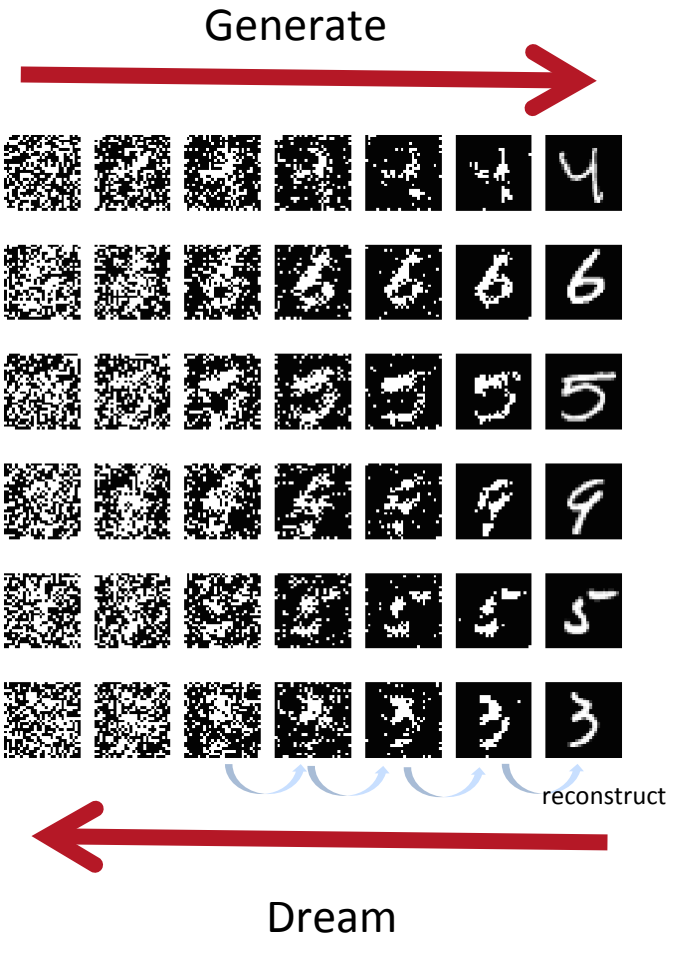


Variational Walkback

Goyal, Ke. Ganguli & Bengio, NIPS 2017



- Sample a data point (dream seed)
- Start running the free-runnning Markov Chain of the brain's transition operator
- Gradually increase temperature (noise)
- At each step, update parameters to make previous state more likely than next state (similar to denoising objective)
- This makes the model FORGET the states it visits in this noisy dream-like simulation (reverse-STDP)
- Carves dynamics to move towards data



Learning a transition operator

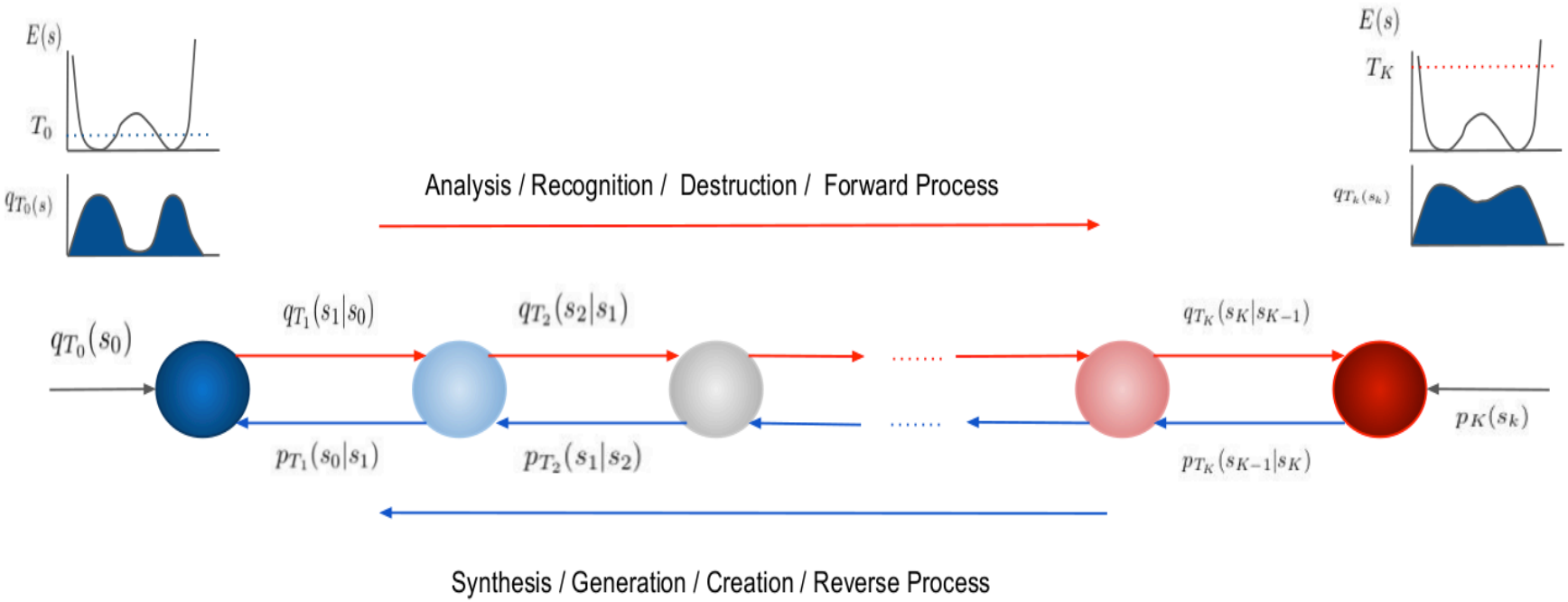
- Instead of learning P(x) directly, learn Markov chain operator $P(x_t \mid x_{t-1})$ as P(x) could potentially have MANY MODES
- More efficient parameterization for a given amount of non-linearity.
- Being able to clamp an arbitrary subset.
- Looks like what brain does in many ways (stochastic, recurrent and being able to handle missing inputs)
- Problem of finding good deep unsupervised generative models is still very much open.
  - **IMPORTANT** to explore new approaches

Variational Walkback
Goyal et al NIPS 2017

- Method to directly parameterize transition operator.
  - Providing an **empirical** method to control the stationary distribution of non-equilibrium stochastic process that does not obey detail balance.
- Modification of variational method
- Potentially asymptotically infinite generative sampling process corresponds to non-equilibrium generalizations of energy based undirected models.
- Radical departure from both directed and undirected graphical models.

Training Process

Learn a stochastic transition operator whose repeated application yields a sample from data distribution.



MILA

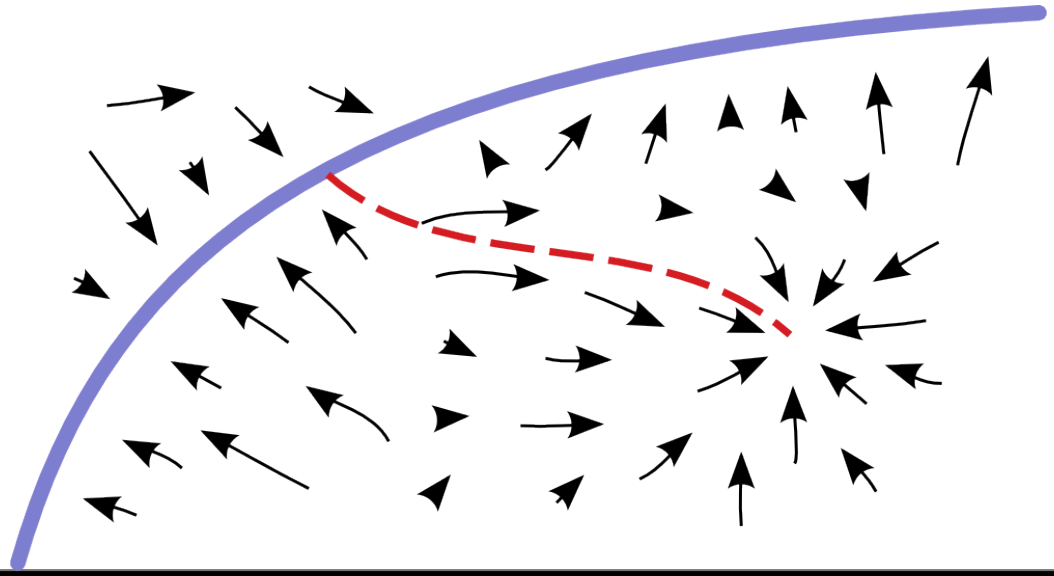
Training Process

- Index operators by temperature, gradually increase for Q_t, gradually decrease for P_t
- Repeated annealed application of P_t operator → data distr.
- Data → repeated de-annealed application of Q_t → Gaussian
- Training:
  - Sample an example, apply Q_1, Q_2, etc.
  - Make reverse trajectory more likely:

learn to walk back heated trajectories starting at data points

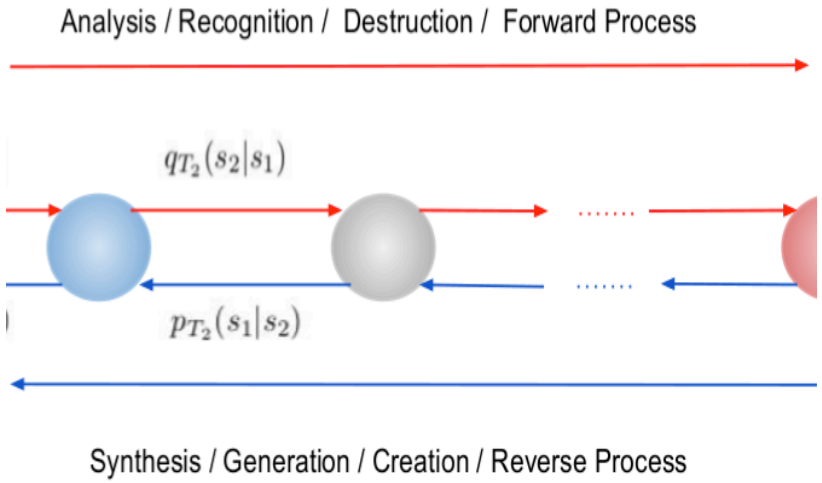
Spurious Modes

- Making the destructive process identical to the transition operator to be learned is motivated by the idea that the destructive process should efficiently explore the spurious modes of the current transition operator.
- The walkback training will then destroy these modes.



Circumvents Credit Assignment Issue

- Providing targets at each time step!
- Each past time step of the heated trajectory
  - Act as a training target for the future output of the generative operator.



Variational Derivation of Walkback

- Marginal probability of the data point at the end of generative process.

$$p(\boldsymbol{s}_0) = \sum_{\boldsymbol{s}_1^K} d\boldsymbol{s}_1^K p_{T_0}(\boldsymbol{s}_0|\boldsymbol{s}_1) \left(\prod_{t=2}^{K} p_{T_t}(\boldsymbol{s}_{t-1}|\boldsymbol{s}_t) \right) p^*(\boldsymbol{s}_K)$$

$$\ln p(v) \equiv \ln \sum_h p(v|h)p(h) = \underbrace{\sum_h q(h|v) \ln \frac{p(v,h)}{q(h|v)}}_{\mathcal{L}} + D_{KL}[q(h|v)||p(h|v)].$$

Tightness of Variational bound

- Tight when the distribution of the heated trajectory starting from a point s_0, matches the posterior distribution of the cooled trajectory ending at s_0.

$$D_{KL} = \sum_{s_1^k} q(s_1^k|s_0) \ln \frac{p(s_0)}{p^*(s_K)} \prod_{t=1}^{K} \frac{q_{T_t}(s_t|s_{t-1})}{p_{T_t}(s_{t-1}|s_t)}.$$

Connection to Dreams

- STDP - Corresponds to increasing the probability of configurations towards which the network goes (i.e remembering observed configurations)
- Reverse-STDP has opposite sign, and corresponds to forgetting the states towards which the model goes.
  - Consistent with the observation that dreams are forgotten quickly.
  - Awake states could be remember for as long as possible(more like STDP)
- Dreams are often incoherent and this could correspond to some form of high temperature version of normal(awake) brain dynamics (again matching VW)

Lower Bound on CIFAR

Method	Lower Bound
NET	5 bits/pixel
Deep VAE	4.54 bits/pixel
VW (5 steps)	8.1 bits/pixel
VW(20 steps)	5.2 bits/pixel
VW (30 steps)	4.23 bits/pixel
DRAW	4.13 bits/pixel

Montreal Institute for Learning Algorithms

MILA

Université de Montréal