On Autoencoders and Score Matching for Energy Based Models

Kevin Swersky[1], Marc'Aurelio Ranzato[2], David Buchman[1], Benjamin M. Marlin[1] and Nando de Freitas[1]

1. Department of Computer Science, University of British Columbia **2.** Department of Computer Science, University of Toronto

Introduction

- We consider two classes of models over continuous data: energy based models (EBMs) and feed-forward autoencoders.
- EBMs define a full probabilistic model, while autoencoders admit fast inference and tractable learning.
- Score matching is a statistical estimator designed for fully visible EBMs with intractable partition functions, but applying it to different EBMs often yields learning objectives that are difficult to interpret.
- We generalize score matching to the class of EBMs with latent variables. We further show that these lead to novel autoencoder architectures with comparable empirical performance on a recently proposed EBM.

Models

A latent variable energy based model defines a probability distribution $\boldsymbol{P_\theta(v,h)} = \frac{\exp(-\boldsymbol{E_\theta(v,h)})}{\boldsymbol{Z(\theta)}}$ over real data vectors $\boldsymbol{v}$ with latent variables $\boldsymbol{h}$, parameterized by θ. $\boldsymbol{E_\theta(v,h)}$ is called the *energy function* and the distribution is normalized by the *partition function* $\boldsymbol{Z(\theta)}$. We also define the marginal distribution $\boldsymbol{P_\theta(v)} = \frac{\exp(-\boldsymbol{F_\theta(v)})}{\boldsymbol{Z(\theta)}}$ in terms of the *free energy* $\boldsymbol{F_\theta(v)} = -\log(\int_{\boldsymbol{h}} \exp(-\boldsymbol{E_\theta(v,h)})\mathbf{d}\boldsymbol{h})$. Some examples include:

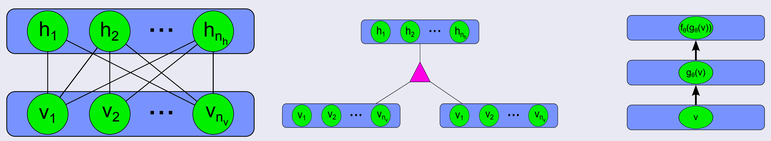
- Gaussian-binary RBMs: $\theta = (\boldsymbol{W}, \sigma, \boldsymbol{b}, \boldsymbol{c})$ and $\boldsymbol{h_j} \in \{0,1\}$,

$$\boldsymbol{E_\theta(v,h)} = -\sum_{i=1}^{n_v}\sum_{j=1}^{n_h}\frac{\boldsymbol{v_i}}{\sigma_i}\boldsymbol{W_{ij}h_j} - \sum_{j=1}^{n_h}\boldsymbol{b_j h_j} + \frac{1}{2}\sum_{i=1}^{n_v}\frac{(\boldsymbol{c_i}-\boldsymbol{v_i})^2}{\sigma_i^2}$$

- mPoT: $\theta = (\gamma, \boldsymbol{W}, \boldsymbol{C}, \boldsymbol{b^v}, \boldsymbol{b^m})$, $\boldsymbol{h_j^m} \in \{0,1\}$ and $\boldsymbol{h_k^c} \in \mathbb{R} > \mathbf{0}$,

$$\boldsymbol{E_\theta(v,h)} = \sum_{k=1}^{n_{hc}}\left[\boldsymbol{h_k^c}(1+\frac{1}{2}(\sum_{i=1}^{n_v}\boldsymbol{C_{ik}v_i})^2) + (1-\gamma)\log(\boldsymbol{h_k^c})\right] + \frac{1}{2}\sum_{i=1}^{n_v}\boldsymbol{v_i^2} - \sum_i \boldsymbol{b_i^v v_i} - \sum_{i=1}^{n_v}\sum_{j=1}^{n_{hm}}\boldsymbol{h_j^m W_{ij} v_i} - \sum_{j=1}^{n_{hm}}\boldsymbol{b_j^m h_j^m}$$

An autoencoder learns to encode its input using an *activation function* $\boldsymbol{g_\theta(v)}$ and produces reconstructions by a decoder $\boldsymbol{f_\theta(g_\theta(v))}$. Learning consists of setting θ to minimize the reconstruction error $||\boldsymbol{f_\theta(g_\theta(v))} - \boldsymbol{v}||^2$.



(a) RBM (b) Factored EBM (c) Autoencoder

Score Matching

We assume the data is generated from some unknown distribution $\tilde{\boldsymbol{p}}(\boldsymbol{v})$. The score matching (SM) estimator is defined as:

$$\boldsymbol{J(\theta)} = \mathbb{E}_{\tilde{\boldsymbol{p}}(\boldsymbol{v})}\left[\sum_{i=1}^{n_v}\frac{1}{2}(\psi_i(\boldsymbol{p_\theta(v)}))^2 + \frac{\partial\psi_i(\boldsymbol{p_\theta(v)})}{\partial\boldsymbol{v_i}}\right].$$

where $\psi_i(\boldsymbol{p(v)}) = \frac{\partial \log \boldsymbol{p(v)}}{\partial \boldsymbol{v_i}}$ is called the *score function*. Since $\frac{\partial \boldsymbol{Z(\theta)}}{\partial \boldsymbol{v_i}} = \mathbf{0}$, we avoid having to compute the partition function.

Applying and Generalizing Score Matching

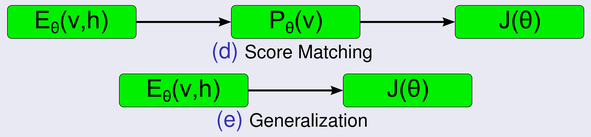
- Score Matching for Gaussian-binary RBMs

$$\boldsymbol{F_\theta(v)} = \frac{1}{2}\sum_{i=1}^{n_v}\frac{(\boldsymbol{c_i}-\boldsymbol{v_i})^2}{\sigma_i^2} - \sum_{j=1}^{n_h}\log\left(1+\exp\left(\sum_{i=1}^{n_v}\frac{\boldsymbol{v_i}}{\sigma_i}\boldsymbol{W_{ij}} + \boldsymbol{b_j}\right)\right)$$

$$\boldsymbol{J(\theta)} = \mathbb{E}_{\tilde{\boldsymbol{p}}(\boldsymbol{v})}\left[\sum_{i=1}^{n_v}\left[\frac{1}{2}\left(\frac{\boldsymbol{v_i}}{\sigma_i^2} - \frac{\boldsymbol{c_i}}{\sigma_i^2} - \sum_{j=1}^{n_h}\frac{\boldsymbol{W_{ij}}}{\sigma_i}\hat{\boldsymbol{h}}_j\right)^2 - \frac{1}{\sigma_i^2} + \sum_{j=1}^{n_h}\frac{\boldsymbol{W_{ij}^2}}{\sigma_i^2}\hat{\boldsymbol{h}}_j\left(1-\hat{\boldsymbol{h}}_j\right)\right]\right]$$

where $\hat{\boldsymbol{h}}_j = \text{sigm}\left(\sum_{i=1}^{n_v}\frac{\boldsymbol{v_i}}{\sigma_i}\boldsymbol{W_{ij}} + \boldsymbol{b_j}\right)$ and $\text{sigm}(\boldsymbol{x}) := \frac{1}{1+\exp(-\boldsymbol{x})}$

  - When $\sigma = \mathbf{1}$, this objective is equivalent to a regularized autoencoder.
- We can also apply score matching to factored models like mPoT.
- When applied to the free energy of a general EBM, the resulting objectives are often difficult to interpret. Instead, we apply score matching without directly marginalizing over the hidden variables.



(d) Score Matching

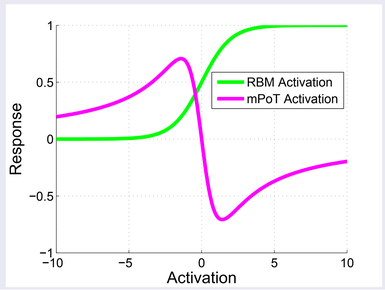
(e) Generalization

Theorem

The score matching objective for a latent EBM can be expressed succinctly in terms of expectations of the energy with respect to the conditional distribution $\boldsymbol{P_\theta(h|v)}$.

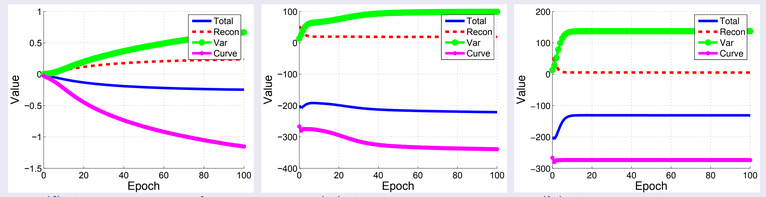
$$\begin{aligned}\boldsymbol{J(\theta)} &= \mathbb{E}_{\tilde{\boldsymbol{p}}(\boldsymbol{v})}\left[\sum_{i=1}^{n_v}\frac{1}{2}\left(-\frac{\partial\boldsymbol{F_\theta(v)}}{\partial\boldsymbol{v_i}}\right)^2 - \frac{\partial^2\boldsymbol{F_\theta(v)}}{\partial\boldsymbol{v_i^2}}\right] \\ &= \mathbb{E}_{\tilde{\boldsymbol{p}}(\boldsymbol{v})}\left[\sum_{i=1}^{n_v}\frac{1}{2}\left(\mathbb{E}_{\boldsymbol{p_\theta(h|v)}}\left[\frac{\partial\boldsymbol{E_\theta(v,h)}}{\partial\boldsymbol{v_i}}\right]\right)^2\right. \\ &\quad \left. + \boldsymbol{var}_{\boldsymbol{p_\theta(h|v)}}\left[\frac{\partial\boldsymbol{E_\theta(v,h)}}{\partial\boldsymbol{v_i}}\right] - \mathbb{E}_{\boldsymbol{p_\theta(h|v)}}\left[\frac{\partial^2\boldsymbol{E_\theta(v,h)}}{\partial\boldsymbol{v_i^2}}\right]\right].\end{aligned}$$

- Score matching will always minimize expected error and variance, and maximize expected energy curvature.
- When the conditional distribution $\boldsymbol{P_\theta(v|h)}$ over the visible units is Gaussian, the error will be a scaled reconstruction term.
- Score matching can be used to derive new autoencoder models from EBMs that account for covariance structure in $\boldsymbol{v}$.
- Applying score matching to factored models leads to autoencoders with peculiar activation functions.



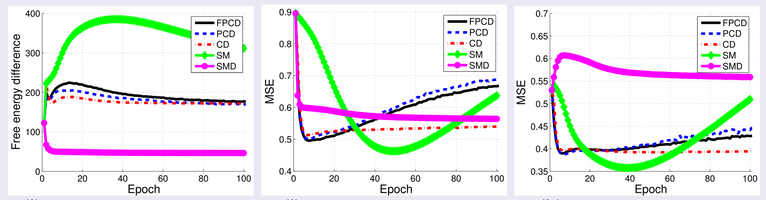
Experiments using mPoT

- We analyze the empirical properties of score matching and compare against several other training methods.
  - CD: Contrastive Divergence
  - PCD: Persistent Contrastive Divergence
  - FPCD: Fast Persistent Contrastive Divergence
  - SM: Score Matching
  - SMD: Denoising Score Matching (Vincent, 2011)
  - AE: mPoT-based autoencoder without regularization
- Objective function analysis



(f) SM terms ($\times\mathbf{10^4}$) (g) SMD terms (h) Autoencoder terms

- Denoising



(i) Free energy difference (j) Mean-field denoising (k) Bayesian denoising

- Learned features

(l) Mean filters (m) Covariance filters

- Classification accuracy on CIFAR 10

CD	PCD	FPCD	SM	SMD	AE
64.6%	64.7%	65.5%	65.0%	64.7%	57.6%

Conclusion

- Applying score matching to a general EBM in terms of the energy function leads to an interpretable objective.
- Score matching for an RBM yields a standard autoencoder, while factored models specify new architectures.
- Regularization, particularly on the curvature of the energy seems to be important for learning good features with autoencoders.
- Methods based on score matching are comparable in performance to stochastic estimators and give tractable objectives.

References

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 2011.