# 生成对抗网络

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio

信息与研​​究部

蒙特利尔大学

蒙特利尔，QC H3C 3J7

## 摘要

我们提出了一个新的框架，用于通过对抗过程估算生成模型，我们同时训练两个模型：生成模型G捕获数据分布，以及判别模型D估算来自训练数据而不是G的样本概率。G的处理过程是使D犯错的可能性最大化。这个框架对应于一个极小极大两人游戏。在任意函数G和D的空间，存在唯一的解决方案，其中G恢复训练数据分布和D等于 无处不在。在通过多层感知器定义了G和D的情况下，可以通过反向传播训练整个系统。在训练或生成样本期间，不需要任何马尔可夫链或展开的近似推理网络。通过定性和定量评估生成的样本，实验证明了框架的潜力。

## 1 引言

深度学习的希望是发现丰富的，层次化的模型，这些模型可以表示在人工智能应用中遇到的各种数据的概率分布，例如自然图像，包含语音的音频波形以及自然语言语料库中的符号。到目前为止，深度学习中最引人注目的成功涉及判别模型，通常是那些将高维，丰富的感觉输入映射到类别标签。这些惊人的成功，主要基于反向传播和辍学算法，使用具有特别好的行为梯度的分段线性单位。深度生成模型有更少的影响，这是因为难以近似许多难以解决的在最大似然估计和相关策略中出现的概率计算，还因为难以利用在生成上下文中分段线性单元的好处。我们提出了一种新的生成模型估算程序避免这些困难。

在提出的对抗网框架中，生成模型与对手形成了对抗：判别模型，该模型学习确定样本是来自模型分布还是来自数据分布。生成模型可以被视为类似于造假团队，试图生产假货币并在未经检测的情况下使用它，而判别模型是类似于警察，试图发现假币。此游戏中的竞争推动两支球队都改进自己的方法，直到假冒品与真品无法区分为止

该框架可以针对多种模型和优化算法产生特定的训练算法。在本文中，我们探讨了生成模型通过使随机噪声穿过多层感知器生成样本时的特殊情况，判别模型也是多层感知器。我们将此特殊情况称为对抗网。在这种情况下，我们可以训练两种模型都仅使用非常成功的反向传播和辍学算法，仅使用从正向传播生成模型中获取样本。没有近似推断或必要的马尔可夫链。

## 2 相关工作

具有潜在变量的有向图形模型的替代方法是具有潜在变量的无向图形模型，例如受限的玻尔兹曼机（RBMs），深波耳兹曼机器（DBMs）及其众多变体。这些模型之间的相互作用是表示为未归一化的潜在函数的乘积，这些函数被所有随机变量状态的全局总和/积分归一化。尽管可以通过马尔可夫链蒙特卡罗（MCMC）方法估算，但对于大多数平凡的实例而言， 此数量（分区函数）和其梯度是很难处理的。混合姿势是学习依赖MCMC的算法的重要问题。

深度信念网络（DBNs）是包含单个无向层和若干个定向层的轮空方法。尽管存在快速近似的分层训练准则，但DBNs会导致与无向和有向模型相关的计算困难。

还提出了不近似或限制对数似然性的替代标准，例如得分匹配和噪声对比估计（NCE）。这两个都需要可以分析地指定学习到的概率密度，直到归一化常数为止。注意在具有几层潜在变量的许多有趣的生成模型中（例如DBN和DBM），甚至不可能得出可处理的未归一化概率密度。一些模型例如降噪自动编码器和压缩自动编码器具有非常相似的学习规则，应用于RBM的得分匹配。在NCE中，就像在这项工作中一样，区别培训标准是用于拟合生成模型。但是，与其采用单独的判别模型，不如生成模型本身用于从样本中以固定的噪声分布区分生成的数据。由于NCE使用固定的噪声分布，因此模型学习后，学习速度将大大降低，甚至在一小部分观察变量上的近似正确分布。

最后，某些技术不涉及显式定义概率分布，而是训练用于从所需分布中提取样本的生成机。这种方法的优点是这样的机器可以设计为通过反向传播进行训练。最近在这方面的杰出工作包括生成随机网络（GSN）框架，该框架扩展了广义去噪自动编码器：两者都可以看作是定义了一个参数化的马尔可夫链，即一个了解执行生成马尔可夫链的一个步骤的机器的参数。与GSN相比较，对抗网络框架不需要马尔可夫链进行采样。因为对抗网络在生成过程中不需要反馈循环，它们更能够利用分段线性单元，可提高反向传播的性能，但具有在反馈回路中使用时无限激活的问题。最新的通过向后传播的方式训练生成机器的示例，包括有关自动编码变化的贝叶斯和随机反向传播最新工作。

## 3 对抗网

当两个模型都是多层感知器时，对抗建模框架最容易应用。为了了解生成器在数据x上的分布p g，我们定义一个先验输入噪声变量p z（z），则表示到数据空间的映射为G（z; θg），其中G是参数为θg的多层感知器表示的微分函数。我们还定义了输出单个标量的第二个多层感知器D（x; θd）。D（x）表示x来自数据而不是p g的概率。我们训练D以最大程度地分配为训练示例和来自G的样本提供正确的标签。我们同时训练G以最大程度地减少log（1-D（G（z）））：

换句话说，D和G玩以下具有值函数V（G，D）的两人极小极大游戏：

min max V（D，G）= E x〜p data（x） [log D（x）] + E z〜p z （z） [log（1- D（G（z）））） （1）

G D

在下一部分中，我们将对对抗网进行理论分析，从本质上表明，训练准则允许恢复数据生成分布作为G和D的是给定足够的容量，即在非参数限制内。参见图1的较不正式的，更具教学性的方法的解释。在实践中，我们必须使用迭代的数值方法来实现游戏。在训练的内部循环中将D优化到完成是计算上的障碍，在有限的数据集上会导致过度拟合。相反，我们在优化D的k个步骤和优化G的1步之间交替。只要G变化足够缓慢，这导致D保持在最佳解附近。该策略类似于SML / PCD训练会从一个学习步骤到下一个学习步骤，维护来自马尔可夫链的样本，以避免在是学习内在循环的一部分的马尔可夫链中燃烧。该程序在算法1正式提出。

在实践中，等式1可能无法为G提供足够的梯度来学习。在学习初期，当G差时，D可以高置信度拒绝样本，因为它们明显不同于训练数据。在这种情况下，log（1- D（G（z）））饱和。而不是训练G以使log（1- D（G（z）））最小化，我们可以训练G以使log D（G（z））最大化。该目标函数导致G和D动力学的相同固定点，但在学习早期提供了更强的梯度。

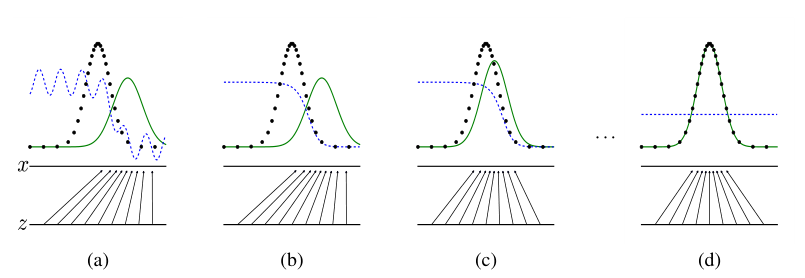


图1：通过同时更新判别分布来训练的生成对抗网络（D，蓝色，虚线），以便区分数据生成分布的样本（黑色，虚线）p x和来自生成分布p g（G）的那些（绿色，实线）。下水平线是z的采样域，在这种情况下是均匀采样。上面的水平线是x域的一部分。向上的箭头表示映射x = G（z）如何将非均匀分布p g施加到转换后的样本。G在p g的高密度区域收缩，并在pg的低密度区域扩展。（a）考虑一个接近收敛的对抗对：p g类似于p data，D是部分准确的分类器。（b）在算法的内部循环中，对D进行训练，以从数据中区分出样本，收敛到D ∗（x）=pdata(x)/(pdata(x)+p g(x))。（c）在更新G之后，D的梯度已引导G（z）流向更可能被归类为数据的区域。（d）经过几步训练，如果G和D有足够的能力，他们将达到因为p g = p data而两者都无法改善的点。辨别器无法区分两个分布，即D（x）= 1/2

## 4 理论结果

生成器G隐式定义概率分布p g作为样本G（z）的获得的分布，当z〜p z。因此，我们希望算法1收敛到一个好的p data的估计量，如果有足够的容量和训练时间。本节的结果以非参数设置完成，例如，我们通过研究在概率密度函数空间中的收敛来表示具有无限容量的模型。

我们将在第4.1节中说明，对于p g = p data，此极小极大游戏具有全局最优值。我们会然后在第4.2节中展示算法1优化了公式1，从而获得了所需的结果。

**算法1**生成对抗网络的Minibatch随机梯度下降训练。应用于鉴别符k的步骤的数量是超参数。我们使用k = 1，这是我们实验中最廉价的选择。

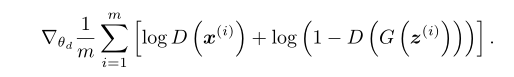
for训练迭代次数

for k步

• 采样小批量来自噪声先验概率p g（z）的m个噪声样本{z （1），...，z （m） }。

• 采样小批量来自数据生成分布pdata（x）的m个示例{x （1），...，x （m） }。

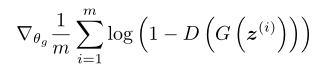
• 通过提高其随机梯度来更新鉴别器：



end for

• 采样小批量来自噪声先验概率p g（z）的m个噪声样本{z （1），...，z （m） }。

•通过降低其随机梯度来更新生成器：



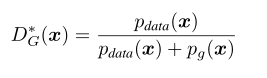
end for

基于梯度的更新可以使用任何标准的基于梯度的学习规则。在我们的实验中我们用了动量。

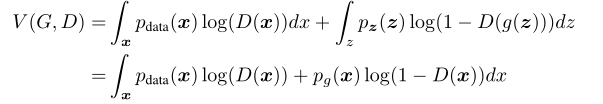
### 4.1 p g = p data全局最优性

我们首先考虑对于任何给定的生成器G的最佳鉴别符D。

**命题1：**对于固定的G，最优判别器D为

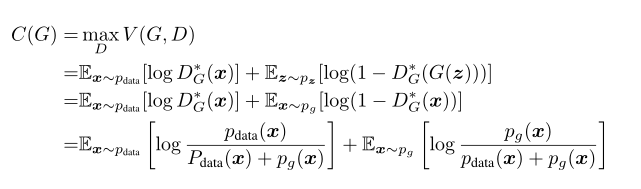
（2）

**证明：**给定任何生成器G，鉴别器D的训练准则是使数量V（G，D）最大化

（3）

对于任何（a，b）∈R2\ {0，0}，函数y→a log（y）+ b log（1- y）在下式中达到最大值，a/(a+b) 在[0，1]。不需要在Supp（p data ）∪Supp（p g ）之外定义鉴别符，结论证明。

请注意，D的训练目标可以解释为最大化的对数似然来估计条件概率P（Y = y | x），其中Y表示x是来自p data（y = 1）还是p g（y = 0）。等式1中的极大极小游戏现在可以重新表示为：

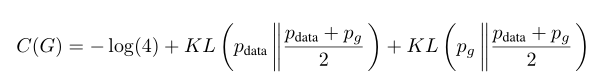
（4）

**定理1：**只有当且仅当p g = pdata时，虚拟训练准则C（G）的全局最小值是可达到的。此时，C（G）达到值− log 4。

**证明：**对于p g = p data，D ∗G （x）= 1/2，（考虑式2）。因此，通过检查等式4在D ∗G （x）= 1/2，我们发现C（G）= log（1/2）+ log（1/2）= - log4。要看到这是达到的最佳C（G）值，仅对于p g = p data，观察到

E x〜p data [-log 2] + E x〜p g [-log 2] = − log 4

并通过从C（G）= V（D \*G ，G）减去这个表达式，我们得到：

（5）

KL是Kullback-Leibler散度。我们在前面的表达式中认识到模型的分布与数据生成过程之间的詹森-香农差异：

C（G）= − log（4）+ 2·JSD（p data || p g） （6）

由于两个分布之间的詹森-香农散度总是非负和零仅当它们相等时，我们证明C ∗ = − log（4）是C（G）的全局最小值，唯一的解决方案是p g = p data，即生成模型完美地复制了数据生成过程。

### 4.2 算法1的收敛

**命题2：**如果G和D具有足够的容量，则在算法1的每个步骤中，允许给定G使判别器达到其最佳值，并更新p g以改进准则

E x〜p data [log D ∗G （x）] + E x〜p g [log（1 − D ∗G （x））]

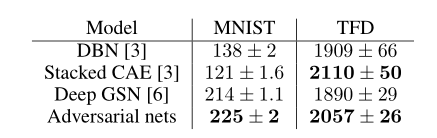
然后p g收敛到p data

**证明：**根据上述准则，将V（G，D）= U（p g，D）视为p g的函数。注意U（p g，D）在p g中是凸的。凸函数的最大次导数包括函数在达到最大值时的导数。换句话说，如果f（x）= sup α∈A f α（x）和f α（x）是凸的在x中对每个α，然后如果β=arg sup α∈A f α（x）则∂f β（x）∈∂f的。这等效于在给定相应G的情况下，在最佳D下计算p g的梯度下降更新。sup D U（p g，D）在p g中是凸的，具有唯一的全局最优解，如Thm 1中所证明的那样，因此，只要p g的更新足够小，p g就会收敛到p x，从而得出证明。

实际上，对抗网络通过函数G（z; θg）表示有限的p g分布族，我们优化了θg而不是p g本身。使用多层感知器定义G引入参数空间中的多个临界点。但是，实践中的多层感知器的优异性能表明，尽管缺乏理论性保证，它们还是可以使用的合理模型。

## 5 实验

我们训练了对抗网络，包括MNIST [23]，多伦多人脸数据库（TFD）[28] 和CIFAR-10 [21]等一系列数据集。生成网络使用了整流器线性激活和S型激活的混合物，而鉴别器网使用maxout [10]激活。dropout [17]被用于训练鉴别器网。虽然我们的理论框架允许使用dropout和发生器中间层的其他噪声，但我们仅将噪声用作输入生成网络的最底层。



**表1：**基于Parzen窗口的对数似然估计。MNIST上报告的数字是测试集上样本的平均对数似然率，并通过示例计算平均值的标准误差。在TFD上，我们使用每折的验证集选择不同的σ来计算数据集折叠的标准误差。在TFD上，对每个折叠进行σ交叉验证，并计算每个折叠的平均对数似然率。对于MNIST，我们将其与数据集的实值（而不是二进制）版本的其他模型进行比较。

我们通过将高斯Parzen窗口拟合到G生成的样本并报告此分布下的对数似然率，来估计p g下测试集数据的概率。通过对验证集进行交叉验证来获得高斯分布的σ参数。此程序产生于Breuleux等，并用于各种生成模型，其确切的可能性难以处理。结果报告在表1中。这种估计可能性的方法具有较高的方差并且在高维空间中表现不佳，但这是我们所掌握的方法中最好的。优势在于生成模型可以采样但无法估计似然率直接激发了对如何评估此类模型的进一步研究。

在图2和3中，我们展示了训练后从生成网络中抽取的样本。虽然我们不做声明称这些样本比通过现有方法生成的样本更好，但我们认为这些样本至少能与文献中较好的生成模型竞争，并强调这种对抗框架的潜力。

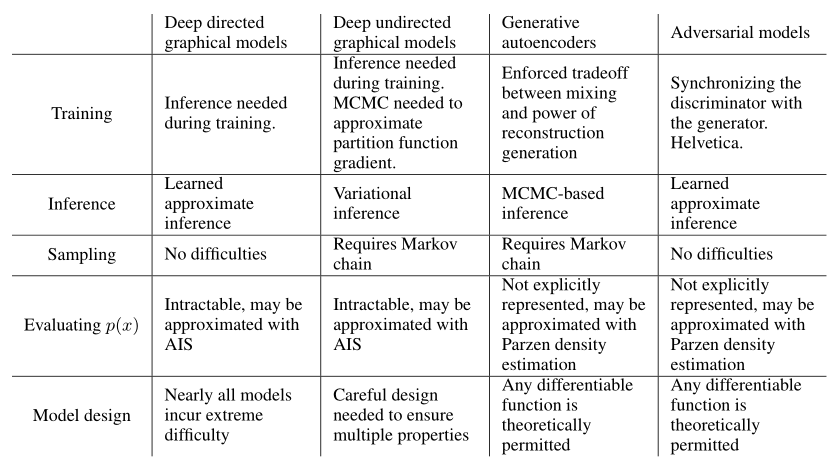


**图2：**模型样本的可视化。最右边的列显示了最近的相邻样本训练示例，为了证明该模型没有记忆训练集。样品是公平的随机抽取，并非精选。与其他大多数深度生成模型的可视化不同，这些图片显示的是模型分布中的实际样本，而不是给定隐藏单元样本的有条件平均值。而且，这些采样是不相关的，因为采样过程不依赖于马尔可夫链，混合了a）MNIST b）TFD c）CIFAR-10（全连接型号）d）CIFAR-10（卷积鉴别器）

和“反卷积”生成器）



**图3：**通过在完整模型的z空间中的坐标之间进行线性插值获得的数字。

**表2：**生成建模中的挑战：涉及模型的每个主要操作的深入生成建模的不同方法遇到的困难的摘要。

## 6 优缺点

与以前的建模框架相比，此新框架具有优缺点。缺点主要是没有明确表示p g（x），而D必须在训练过程中与G保持良好的同步（特别是，不更新D不能对G进行过多训练，以避免G折叠过多z值等于x的相同值的“ Helvetica场景”，以具有足够的多样性来建模p data），与Boltzmann机的负链一样，在学习步骤之间必须保持的最新状态。优点是马尔可夫链永远不需要，仅使用backprop来获得梯度，在学习过程中不需要推理，以及各种各样的功能可以纳入模型。表2总结生成对抗网络与其他生成建模方法的比较。

前述优点主要是计算上的。对抗模型也可能获得生成网络的未通过数据检查直接更新的某些统计优势，但仅在梯度流过鉴别器的情况下。这意味着输入的组件不会直接复制到生成器的参数中。对抗网络的另一个优势是，它们可以表示非常尖锐的甚至简并的分布，尽管基于马尔可夫链条要求分布有些模糊，以便链条能够混合模式。

## 7 结论与未来工作

该框架允许许多直接扩展：

1.可以通过将c作为输入添加到G和D来获得条件生成模型p（x | c）。

2.可以通过训练辅助网络来给定x预测z，学习到的近似推断。这类似于唤醒睡眠算法训练的推理网络，但具有可以在生成网络完成训练之后，为固定的生成器网络训练推理网络的优点。

3.可以近似建模所有条件p（x S | x S），其中S是x索引的子集，通过训练共享参数的一系列条件模型来获得x。本质上，一个人可以使用对抗网络来实现确定性MP-DBM的随机扩展。

4.半监督学习：当有限的标签数据可用时，鉴别器或推理网的功能可以提高分类器表现的性能。

5.提高效率：可以通过采用更好的方法来大大提高培训速度。例如，在训练过程中协调G和D或确定样本z的更好分布。

本文证明了对抗建模框架的可行性，这表明这些研究方向可能被证明是有用的。

## 致谢

我们要感谢Patrice Marcotte，Olivier Delalleau，Kyunghyun Cho，Guillaume Alain和Jason Yosinski进行了有益的讨论。Yann Dauphin和我们分享了他的Parzen窗口评估代码。我们要感谢Pylearn2 [12]和Theano [7，1]的开发人员，特别是FrédéricBastien，他急忙推出Theano功能以使该项目受益。Arnaud Bergeron提供了LATEX排版急需的支持。我们还要感谢CIFAR和加拿大研究主席负责资助，Compute Canada和CalculQuébec负责提供计算资源。Ian Goodfellow获得了2013年Google奖学金的支持在深度学习方面。最后，我们要感谢Les Trois Brasseurs激发了我们的创造力。

## 参考文献

[1] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., and Bengio, Y. (2012). Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

[2] Bengio, Y. (2009). Learning deep architectures for AI. Now Publishers.

[3] Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. (2013a). Better mixing via deep representations. In ICML’13.

[4] Bengio, Y., Yao, L., Alain, G., and Vincent, P. (2013b). Generalized denoising auto-encoders as generative models. In NIPS26. Nips Foundation.

[5] Bengio, Y., Thibodeau-Laufer, E., and Yosinski, J. (2014a). Deep generative stochastic networks trainable by backprop. In ICML’14.

[6] Bengio, Y., Thibodeau-Laufer, E., Alain, G., and Yosinski, J. (2014b). Deep generative stochastic networks trainable by backprop. In Proceedings of the 30th International Conference on Machine Learning (ICML’14).

[7] Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. In Proceedings of the Python for Scientific Computing Conference (SciPy). Oral Presentation.

[8] Breuleux, O., Bengio, Y., and Vincent, P. (2011). Quickly generating representative samples from an RBM-derived process. Neural Computation, 23(8), 2053–2073.

[9] Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In AISTATS’2011.

[10] Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013a). Maxout networks. In ICML’2013.

[11] Goodfellow, I. J., Mirza, M., Courville, A., and Bengio, Y. (2013b). Multi-prediction deep Boltzmann machines. In NIPS’2013.

[12] Goodfellow, I. J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F., and Bengio, Y. (2013c). Pylearn2: a machine learning research library. arXiv preprint arXiv:1308.4214.

[13] Gutmann, M. and Hyvarinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In AISTATS’2010.

[14] Hinton, G., Deng, L., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012a). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 29(6), 82–97.

[15] Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. Science, 268, 1558–1161. 8

[16] Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18, 1527–1554.

[17] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012b). Improving neural networks by preventing co-adaptation of feature detectors. Technical report, arXiv:1207.0580.

[18] Hyvärinen, A. (2005). Estimation of non-normalized statistical models using score matching. J. Machine Learning Res., 6.

[19] Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture forobjectrecognition? InProc.InternationalConferenceonComputerVision(ICCV’09), pages2146–2153. IEEE.

[20] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In Proceedings of the International Conference on Learning Representations (ICLR).

[21] Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

[22] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In NIPS’2012.

[23] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

[24] Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. Technical report, arXiv:1401.4082.

[25] Rifai, S., Bengio, Y., Dauphin, Y., and Vincent, P. (2012). A generative process for sampling contractive auto-encoders. In ICML’12.

[26] Salakhutdinov, R. and Hinton, G. E. (2009). Deep Boltzmann machines. In AISTATS’2009, pages 448–455.

[27] Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing, volume 1, chapter 6, pages 194–281. MIT Press, Cambridge.

[28] Susskind, J., Anderson, A., and Hinton, G. E. (2010). The Toronto face dataset. Technical Report UTML TR 2010-001, U. Toronto.

[29] Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, ICML 2008, pages 1064–1071. ACM.

[30] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In ICML 2008.

[31] Younes, L. (1999). On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. Stochastics and Stochastic Reports, 65(3), 177–228