GAUSSIAN-BERNOULLI RBMS WITHOUT TEARS

Renjie Liao*, Simon Kornblith, Mengye Ren, David J. Fleet, Geoffrey Hinton
University of British Columbia, Google Research, Brain Team, New York University, University of Toronto, Vector Institute
rjliao@ece.ubc.ca, mengye@cs.nyu.edu
{skornblith, davidfleet, geoffhinton}@google.com

ABSTRACT

We revisit the challenging problem of training Gaussian-Bernoulli restricted Boltzmann machines (GRBMs), introducing two innovations. We propose a novel Gibbs-Langevin sampling algorithm that outperforms existing methods like Gibbs sampling. We propose a modified contrastive divergence (CD) algorithm so that one can generate images with GRBMs starting from noise. This enables direct comparison of GRBMs with deep generative models, improving evaluation protocols in the RBM literature. Moreover, we show that modified CD and gradient clipping are enough to robustly train GRBMs with large learning rates, thus removing the necessity of various tricks in the literature. Experiments on Gaussian Mixtures, MNIST, FashionMNIST, and CelebA show GRBMs can generate good samples, despite their single-hidden-layer architecture. Our code is released at: https://github.com/lrjconan/GRBM
The neural sampling hypothesis: **exploit intrinsic variability to sample from probability distributions.**

\[ P(x) = \frac{1}{Z} e^{-E(x)/T} \]

**Langevin dynamics:**
\[ \tau \dot{x} = -\nabla_x E(x) + \sqrt{2T\tau} n \Rightarrow x \sim P(x) \]

- **Build analog electronics to implement** \( \tau \dot{x} = -\nabla_x E(x) \).
- **Add noise** \( n \sim \mathcal{N}(0, I) \).

\[ \Rightarrow \]
- Noisy (low power) electronics can be exploited to do probabilistic inference.
- Deterministic, high-precision circuits are wasted on most problems of AI which involve reasoning under uncertainty.
We found that training with modified CD alone occasionally diverges, necessitating careful tuning of the learning rate. However, adding gradient clipping (e.g., clip gradient norm to 10) enables stable training with all aforementioned sampling methods. We therefore set learning rate to $0\Delta 0.1$ for all experiments. Such a large learning rate almost never works in the literature. Melchior et al. (2017) used gradient clipping and similarly large learning rates, but they had to set the learning rate for the variances 100 times smaller than that for the weights and biases during CD training. But thanks to the modified CD and gradient clipping, we found this special treatment of variances is unnecessary. We do not use momentum, weight decay, PCD, or other tricks.

4.1 Modeling Gaussian Mixtures

We first evaluate density modelling by GRBMs when the data density is known, i.e., Gaussian mixture models (GMMs) in our case. This is challenging for GRBMs as the marginal distribution of visible units of GRBMs is essentially a constrained Gaussian mixture, i.e., the weights of mixture components depend on one another (Melchior et al., 2017). As such, the mixture components in GRBMs can not be freely placed in the visible domain so one actually needs more hidden units than the log of the number of mixture components to fit GMMs well. We consider the 2D case for simplicity and better visibility. We generate 1,000 samples from two types (isotropic and anisotropic variances) of GMMs with 3 components as shown in Fig. 1, and learn GRBMs using our modified CD with different sampling algorithms, from which we can draw samples. Here all samplers run for 100 steps during both CD training and testing (see Appendix B.1 for more detail). Density plots and samples are shown in Fig. 1. Notice that Gibbs manages to recover the three modes in the isotropic case but fails in the anisotropic case. Both Langevin and Gibbs-Langevin sampling collapse when the adjustment is absent. We believe the cosine step size schedule contributes to the collapse as it removes more stochasticity of Langevin dynamics with small step sizes, thus making sampling more similar to gradient descent. But as we will see later, in image modelling, this may not be so severe; there are more modes so that the sampling may collapse to different modes, and the diversity of images remains acceptable. Finally, both Langevin and Gibbs-Langevin do recover all three modes with the adjustment, which shows the adjustment helps the mixing in this synthetic case.

4.2 Image Generation

We learn GRBMs to fit image datasets including MNIST, FashionMNIST, and CelebA. To the best of our knowledge, this is the first time that GRBMs have been shown to (unconditionally) generate good images. We provide the ablation study in Appendix B.2 and more results in Appendix B.3.

Table 1: Results on MNIST dataset.

| Methods                        | FID    |
|--------------------------------|--------|
| VAE                            | 16.13  |
| 2sVAE (Dai & Wipf, 2019)       | 12.60  |
| PixelCNN++ (Salimans et al.)    | 11.38  |
| WGAN (Arjovsky et al., 2017)    | 10.28  |
| NVAE (Vahdat & Kautz, 2020)    | 7.93   |
| GRBMsis                        |        |
| Gibbs                          | 47.53  |
| Langevin wo. Adjust             | 43.80  |
| Langevin w. Adjust              | 41.24  |
| Gibbs-Langevin wo. Adjust       | 17.49  |
| Gibbs-Langevin w. Adjust        | 19.27  |

Table 1: Results on MNIST dataset.
Figure 3: (a) Learning curve of (natural) log variances, (b) learned filters, and (c) samples on MNIST.

Figure 4: Samples from GRBMs on (a) FashionMNIST, (b) CelebA-32, and (c) CelebA-2K-64.

theless, GRBMs trained with Gibbs-Langevin without adjustment achieve FID scores comparable to other deep generative models, which is impressive given the single-hidden-layer architecture. The learning curve of (natural) log variance is shown in Fig. 3a. The learned variance converges to around $10^{-5}$ which is significantly smaller than those reported in the literature. The learned filters are shown in Fig. 3b. Although some point-like filters still exist, stroke-like filters are common, thus indicating GRBMs indeed learn meaningful features. We show samples drawn from the best GRBM in Fig. 3c. The intermediate samples from Gibbs-Langevin are shown in Fig. 2. Since Gibbs-Langevin without adjustment works the best, we use it for remaining experiments.

FashionMNIST

We then train GRBMs on FashionMNIST which is more challenging than MNIST. We set hidden size to 10,000 and the sampling step to 100. Samples drawn from learned GRBMs are shown in Fig. 4a. GRBMs successfully learn the shapes of clothes, shoes, bags, and so on. However, they fail to capture fine textures. Since many images in this dataset look similar in shape but differ in texture, the resulting samples look similar to each other.

CelebA

Last, we consider the even more challenging CelebA dataset. In particular, we explore two versions of this dataset: 1) CelebA-32 where we center-crop (140 × 140) and downsample images to 32 × 32; 2) CelebA-2K-64 where randomly select 2,000 images from the original CelebA and apply the same center crop and downsampling to 64 × 64. We set hidden size to 10,000 and explore the number of 100 and 200 sampling steps. Generated samples are shown in Fig. 4b and 4c. From the figure, we can see that GRBMs can learn to generate reasonably good face images.

5C

ONCLUSION

In this paper, we revisit learning Gaussian-Bernoulli restricted Boltzmann machines. We investigate Langevin Monte Carlo and propose a novel Gibbs-Langevin sampling method. Furthermore, we modify the contrastive divergence (CD) algorithm so that one can sample data from learned GRBMs starting from noise. Modified CD along with gradient clipping enables robust training of GRBMs with large learning rates. Finally, we show that GRBMs can unconditionally generate images with good qualities, despite its single-hidden-layer architecture. In the future, it would be beneficial to

9
Learning and Inference in Sparse Coding Models With Langevin Dynamics

Michael Y.-S. Fang
michaelfang@berkeley.edu
Department of Physics, University of California, Berkeley, Berkeley, CA 94720, U.S.A., and Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA 94720, U.S.A.

Mayur Mudigonda
mudigonda@berkeley.edu
Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA 94720, U.S.A.

Ryan Zarcone
ryanzarcone@gmail.com
Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA 94720, U.S.A., and Biophysics Graduate Group, University of California, Berkeley, Berkeley, CA 94720, U.S.A.

Amir Khosrowshahi
khosra@gmail.com
Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA 94720, U.S.A., and Intel Corporation, Santa Clara, CA 95054, U.S.A.

Bruno A. Olshausen
baolshausen@berkeley.edu
Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA, U.S.A., and Helen Wills Neuroscience Institute and School of Optometry, University of California, Berkeley, Berkeley, CA 94720, U.S.A.