Energy-Based Models with Applications to Speech and Language Processing

Other titles in Foundations and Trends® in Signal Processing

Model-Based Deep Learning
Nir Shlezinger and Yonina C. Eldar
ISBN: 978-1-63828-264-8

Generalizing Graph Signal Processing: High Dimensional Spaces, Models and Structures
Xingchao Jian, Feng Ji and Wee Peng Tay
ISBN: 978-1-63828-150-4

Learning with Limited Samples: Meta-Learning and Applications to Communication Systems
Lisha Chen, Sharu Theresa Jose, Ivana Nikoloska, Sangwoo Park, Tianyi Chen and Osvaldo Simeone
ISBN: 978-1-63828-136-8

Signal Decomposition Using Masked Proximal Operators
Bennet E. Meyers and Stephen P. Boyd
ISBN: 978-1-63828-102-3

Energy-Based Models with Applications to Speech and Language Processing

Zhijian Ou
Tsinghua University
ozj@tsinghua.edu.cn

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

Z. Ou. *Energy-Based Models with Applications to Speech and Language Processing.* Foundations and Trends® in Signal Processing, vol. 18, no. 1-2, pp. 1–199, 2024.

ISBN: 978-1-63828-307-2

Foundations and Trends® in Signal Processing
Volume 18, Issue 1-2, 2024

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
  - Classification and detection
  - Estimation and regression
  - Tree-structured methods

Information for Librarians

Foundations and Trends® in Signal Processing, 2024, Volume 18, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Contents

List of Algorithms

List of Figures

List of Tables

Energy-Based Models with Applications to Speech and Language Processing

Zhijian Ou

Tsinghua University, Beijing, China; ozj@tsinghua.edu.cn

ABSTRACT

Energy-Based Models (EBMs) are an important class of probabilistic models, also known as random fields and undirected graphical models. EBMs are un-normalized and thus radically different from other popular self-normalized probabilistic models such as hidden Markov models (HMMs), autoregressive models, generative adversarial nets (GANs) and variational auto-encoders (VAEs). During these years, EBMs have attracted increasing interest not only from core machine learning but also from application domains such as speech, vision, natural language processing (NLP) and so on, with significant theoretical and algorithmic progress. To the best of our knowledge, there are no review papers about EBMs with applications to speech and language processing. The sequential nature of speech and language also presents special challenges and needs treatment different from processing fix-dimensional data (e.g., images).

The purpose of this monograph is to present a systematic introduction to energy-based models, including both algorithmic progress and applications in speech and language processing, which is organized into four main sections. First,

Zhijian Ou (2024), "Energy-Based Models with Applications to Speech and Language Processing", Foundations and Trends® in Signal Processing: Vol. 18, No. 1-2, pp 1–199. DOI: 10.1561/2000000117.

we will introduce basics for EBMs, including classic models, recent models parameterized by neural networks, sampling methods, and various learning methods from the classic learning algorithms to the most advanced ones. The next three sections will present how to apply EBMs in three different scenarios, i.e., for modeling marginal, conditional and joint distributions, respectively. 1) EBMs for sequential data with applications in language modeling, where we are mainly concerned with the marginal distribution of a sequence itself; 2) EBMs for modeling conditional distributions of target sequences given observation sequences, with applications in speech recognition, sequence labeling and text generation; 3) EBMs for modeling joint distributions of both sequences of observations and targets, and their applications in semi-supervised learning and calibrated natural language understanding. In addition, we will introduce some open-source toolkits to help the readers to get familiar with the techniques for developing and applying energy-based models.

1

Introduction

1.1 The Probabilistic Approach

As a community we seem to have embraced the fact that dealing with *uncertainty* is crucial for machine intelligence tasks such as speech recognition and understanding, speech synthesis, natural language labeling, machine translation, text generation, computer vision, signal denoising, decision making, and so on. Uncertainty arises because of limitations in our ability to observe the world, limitations in our ability to model it, and possibly even because of innate nondeterminism [77]. In the face of such uncertainty, we use probabilistic models to describe the random phenomena. Indeed, many tasks in intelligent signal processing and machine learning are solved in the *probabilistic approach*, which generally involves probabilistic modeling, inference and learning, as shown in Figure 1.1. Such probabilistic approach has been introduced in textbooks with sufficient details [14], [59], [77], [110], and thus in this paper we only give a brief overview as the background material.

A *probabilistic model* is, in mathematical terms, a distribution over a set of random variables, which are assumed to characterise the random phenomena in the specific task. The set of variables can generally be divided into observations x and (optionally) hidden variables h,

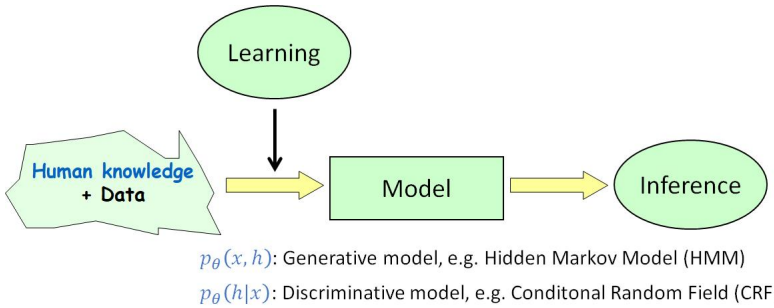

Figure 1.1: The probabilistic approach

according to their roles in the task. Hidden variables, or called latent variables, are variables that are part of the model, but which we do not observe, and are therefore not part of the data. Remarkably, the observability of some variables may change, depending on what phase (training or testing) the model is used. A most common example is the *target variable* in prediction tasks, such as the class label in classification or the response variable in regression, which is observed in training but becomes unknown in testing. To avoid clutter in this paper, such variable is viewed as part of the hidden variables and usually denoted by y.

We will typically denote a variable by a lower case letter such as x, h and y. Whether x denotes the value that the variable takes or represents the variable itself would be clear from the context. Further, for notational simplicity, we also use lower case letter (e.g., x) to denote a set of random variables, i.e., flattened and concatenated such that the set is represented as a single vector. So if x is a vector or a sequence, its components can be accessed by subscripts x_i. Here, we are using the terminology *distribution* or *density* loosely, typically denoted by p. Our notation p should be understood as a mass function (density with respect to counting measure) in the discrete case, and a density function with respect to Lebesgue measure in the continuous case. See Appendix A for more on notations.

Given the form of the probabilistic model, namely the distribution $p_\theta(x, h)$ with parameters θ, there are two crucial problems that must be solved in applying the model in real-world tasks:

- Inference: how to reason in the presence of uncertainty;
- Learning: how to learn from experience.

The former problem is often referred to *probabilistic inference* with a fully-specified model, or inference for short; and the later problem sometimes referred to *statistical inference* (or more often to say, *learning* in machine learning terminology) for model parameters [112].

Put in a more straightforward way, *learning* is to find the most appropriate model with parameters, using both data and human knowledge. Human knowledge is implicitly employed to specify the family of parametric distributions, and data are used to estimate the parameters. Given a fully-specified model, i.e., fully-determined with fixed parameters, *inference* is to infer the unknown from the observation x. There are several typical classes of inference problems:

- Computing conditional probabilities, e.g., $p_\theta(h|x)$. This amounts to computing the posterior probability of some variables given the values of other variables (i.e., given evidence on others).
- Computing marginal probabilities, including the likelihood $p_\theta(x)$.
- Computing modes, e.g., $\arg\max_h p_\theta(h|x)$.
- Sampling from the model [87], [112].

We provide two more points for readers to appreciate the importance of the inference problems. First, the inference problems themselves are often taken as the means to use the model. For example, speech recognition is generally to find the mode of the posterior distribution on state sequences given observed speech. Second, learning algorithms often make use of some inference problem as a subroutine. For example, algorithms that maximize the likelihood for learning latent variable models, e.g., the *expectation-maximization* (EM) algorithm [38], call the calculation of $p_\theta(h|x)$ as a subroutine. Seeking computational efficient algorithms to solve these inference problems for increasingly complex models has been an enduring challenge for our research community.

1.1.1 Generative models and discriminative models

One major division in the probabilistic approach is generative versus discriminative modeling. In generative modeling, one aims to learn the joint distribution $p_\theta(x, h)$ over all the variables. In discriminative modeling, one only models the conditional distribution $p_\theta(h|x)$ over the target variable (denoted by h for convenience) given the observation x. In discriminative modeling, the observation and the target variable are also called the input and output, respectively.

The generative-discriminative distinction has received much attention in machine learning [84], [116]. When a discriminative model follows the induced form of the conditional distribution $p_\theta(h|x)$ from a generative model $p_\theta(x, h)$, the two models are called a *generative-discriminative pair* (i.e., under the same parametric family of models) [116]. For example, naive Bayes classifier and logistic regression, *hidden Markov model* (HMM) [133] and *conditional random field* (CRF) [79], [170], form Generative-Discriminative pairs, respectively. To compare generative and discriminative learning, it seems natural to focus on such pairs. Basically, there are different regimes of performance as the training set size is increased. Taking naive Bayes and logistic regression as a case study, it is shown in [116] that "while discriminative learning has lower asymptotic error, a generative classifier may also approach its (higher) asymptotic error much faster". The comparison of HMM and CRF is further studied in [84], and it is found that generative modeling (modeling more of the data) tends to reduce asymptotic variance, but at the cost of being more sensitive to model misspecification. These previous results, including [84], [116], to name a few, strengthen our basic intuitions about generative-discriminative distinction.

Given that the generative and discriminative estimators are complementary, one natural question is how to interpolate between the two to get the benefits of both. There have been studies on *hybrid generative-discriminative* methods (see [15] and the references therein). Notably, those hybrid models have been applied for *semi-supervised learning* (SSL), where one may have few labeled examples and many more unlabeled examples, but mostly based on traditional generative models like naive Bayes.

In recent years, generative modeling techniques have been greatly advanced by inventing new models with new learning algorithms under the umbrella of *deep generative models* (DGMs), which are characterized by using *multiple layers of stochastic or deterministic variables* in modeling and are much more expressive than classic generative models such as naive Bayes and HMM. See [121] for a systematic introduction to DGMs from perspective of graphical modeling. The generative-discriminative discussion continues with new points, when more types of generative models have constantly emerged and become studied. Here we provide two examples with the new points.

- A type of DGMs, *variational autoencoders* (VAEs) [75], has been successfully applied in the setting of semi-supervised learning.
- It is concurrently shown in [49], [162] that a standard discriminative classifier $p_\theta(y|x)$ can be used to directly define an *energy-based model* (EBM) for the joint distribution $p_\theta(x, y)$. It is shown in [162] that energy-based semi-supervised training of the joint distribution produces strong classification results on par with state-of-art DGM-based semi-supervised methods. It is demonstrated in [49] that energy based training of the joint distribution improves calibration, robustness, and out-of-distribution detection while also generating samples rivaling the quality of recent *generative adversarial network* (GAN) [45] approaches.

1.1.2 Conditional models

Discriminative models are a kind of conditional models for discriminative tasks. However, conditional modeling is a more general modeling concept than discriminative modeling. Basically, a *conditional model* is, in probability terms, a conditional distribution of a random variable of interest, when another variable c is known to take a particular value. In this case, c is often called the *input* of the model. The variable of interest generally can still consist of observable and (optionally) hidden components, denoted by x and h respectively. Thus, a conditional model can generally be denoted by $p_\theta(x, h|c)$.

Many real-world applications are solved by conditional modeling. Some examples from discriminative tasks are as follows.

- First, by abuse of notation, discriminative modeling of image classification involves the conditional model $p_\theta(y|x)$, where x is the input image and y is the images's class.
- A more complicated example is the *recurrent neural network transducer* (RNN-T) model [50] for speech recognition. Let x denote the input speech, y the label sequence (e.g., word transcription), and π the hidden state sequence (or say, a path) which realizes the alignment of x and y. Then the RNN-T model involves the conditional model $p_\theta(y, \pi|x)$. See Section 4.3.1 for more details on RNN-T.

Apart from discriminative tasks, conditional models can also be used for *conditional generation* tasks. One example is the reverse of the image classification problem: prediction of a distribution over images, conditioned on the class label.

Importantly, one should keep in mind that the learning and inference methods introduced in unconditional modeling are in theory equally applicable to conditional models. So the basics introduced in Section 2 lay the foundation for both (unconditional) EBMs in Section 3 and conditional EBMs in Section 4. On the other hand, the unconditional and conditional settings have their own characteristics, and thus needs different treatments, as we will detail in Section 3 and 4 respectively.

1.2 Features of EBMs

In the probabilistic approach, the family of models chosen in real-world applications clearly plays a crucial role. In terms of graphical modeling terminology [77], probabilistic models can be broadly classified into two classes - directed and undirected.

- In *directed graphical models* (DGMs), also known as (a.k.a.) *Bayesian networks* (BNs) or called *locally-normalized models*, the distribution is factorized into a product of local conditional density functions.

- In contrast, in *undirected graphical models* (UGMs), also known as *Markov random fields* (MRFs) or *energy-based models* (EBMs) or called *globally-normalized models*, the distribution is defined to be proportional to the product of local potential functions. The three terms, UGMs, MRFs and EBMs, are exchangeable in this monograph.

Simply speaking, an easy way to tell an undirected model from a directed model is that an undirected model is un-normalized and involves the normalizing constant (also called the partition function in physics), while the directed model is self-normalized.

In general, directed models and undirected models make different assertions of conditional independence. Thus, there are families of probability distributions that are captured by a directed model and are not captured by any undirected model, and vice versa [126]. Therefore, undirected models, though less explored, provide an important complementary choice to directed models for various real-world applications.

During these years, EBMs have attracted increasing interest not only from core machine learning but also from application domains such as speech, vision, natural language processing and so on, with significant theoretical and algorithmic progress. There has emerged a dedicated workshop at ICLR 2021, which is a broad forum about EBM research, and a tutorial at CVPR 2021, which focuses on computer vision tasks.

- ICLR2021 Workshop - Energy Based Models: Current Perspectives, Challenges, and Opportunities, https://sites.google.com/view/ebm-workshop-iclr2021

- CVPR 2021 Tutorial: Theory and Application of Energy-Based Generative Models, https://energy-based-models.github.io/

To the best of our knowledge, there are no review papers about EBMs with applications to speech and language processing. The sequential nature of speech and language also presents special challenges and needs treatment different from processing fix-dimensional images that was described in the CVPR 2021 tutorial. The aim of this monograph is to present a systematic introduction to energy-based models, including

both algorithmic progress and applications in speech and language processing. We hope it will also be of general interest to the artificial intelligence and signal processing communities.

Before delving into the specific content, we first point out *five key features of EBMs*, which may motivate you to pursue the study and application of EBMs.

- Flexibility in modeling. Compared to modeling a self-normalized density function, learning EBMs relaxes the normalization constraint and thus allows much greater flexibility in the parameterization of the energy function. Moreover, undirected modeling is more natural for certain domains, where fixing the directions of edges is awkward in a graphical model.
- Computation efficiency in likelihood evaluation, since the negative log likelihood of an EBM (by ignoring an additive constant) can be easily evaluated, without incurring any calculation for normalization.
- Naturally overcoming label bias and exposure bias suffered by locally-normalized models (Section 4.1.2).
- Superiority for hybrid generative-discriminative and semi-supervised learning (Section 5).
- Challenge in model training. Both computation of the exact likelihood and exact sampling from EBMs are generally intractable, which makes training especially difficult.

1.3 Organization of This Monograph

The rest of the monograph is organized as follows.

In Section 2, we present the basics for EBMs. We start with a brief introduction to probabilistic graphical models (PGMs), because we introduce EBMs as undirected graphical models (UGMs). Then, we present EBM model examples, including both classic ones (such as Ising model and restricted Boltzmann machines) and modern ones parameterized by neural networks. Next, basic algorithms for learning

EBMs are described, which covers the two most widely used classes of methods - Monte Carlo based maximum likelihood methods and noise-contrastive estimation (NCE) methods. Finally, we present a dedicated section to introduce how to sample/generate from EBMs, since sampling is not only a critical step in maximum likelihood learning of EBMs, but also itself forms an important class of applications in speech and language processing.

The basics for inference and learning with EBMs are general for both discrete and continuous data modeling. Remarkably, most applications covered in this monograph are discrete data modeling (text in natural language processing, discrete labels in speech recognition), but in some places, we also present examples and applications in images. For example, Ising model is introduced for readers to get the abstract concepts conveyed by EBMs. EBM based joint-training for semi-supervised image classification is a fixed-dimensional counterpart of the more complicated sequence setting, which is for semi-supervised natural language labeling.

The next three sections introduce how to develop EBMs in three different scenarios respectively.

- Note that the sequential nature of speech and language presents special challenges and needs treatment different from processing fix-dimensional data (e.g., images). In Section 3, we introduce EBMs for sequential data with applications in language modeling. In this scenario, we are mainly concerned with learning the (*marginal*) distribution of an observation sequence x itself, e.g., a natural language sentence as in language modeling.

- In Section 4, we introduce EBMs for modeling *conditional* distributions of target sequences given observation sequences. Conditional EBMs have been successfully applied in speech recognition, sequence labeling in natural language processing (NLP), and various forms of conditional text generation (e.g., controlled text generation, factual error correction).

- In Section 5, we introduce EBMs for modeling *joint* distributions of both sequences of observations and targets. We first introduce the fixed-dimensional case, then move on to the sequential case,

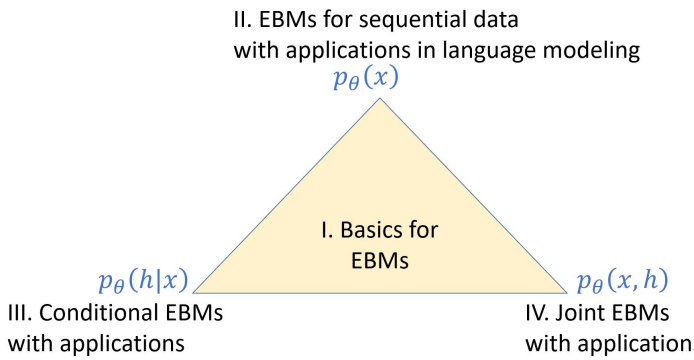

Figure 1.2: Outline of this monograph

and finally present the applications in semi-supervised natural language labeling and calibrated natural language understanding.

Finally, conclusions are given in Section 6 to summarize the monograph and to discuss future challenges and directions.

We visualize the content of this monograph in Figure 1.2. At the center is the basic knowledge for EBM modeling and learning. The basic theory can be applied to model different types of distributions – the distribution of the observation itself, the conditional distribution, and the joint distribution. In different applications or scenarios, we are concerned with different types of distributions. In Sections 3, 4, and 5, we in fact show how to develop EBMs for the three different types of distributions in three different scenarios, respectively, as described above.

This monograph contains the material expanded from the tutorial that the author gave at ICASSP 2022 in May 2022. Substantial updates have been made to incorporate more recent work and cover wider areas of research activities.

Acknowledgements

Thanks to collaborators and students: Zhiqiang Tan, Bin Wang, Hongyu Xiang, Yunfu Song, Kai Hu, Keyu An, Huahuan Zheng, Silin Gao, Hong Liu, Junlan Feng, and Yi Huang.
Thanks for funding support from:

- NSFC (National Science Foundation of China) through No. 60402029, No. 61075020, No. 61473168, and No. 61976122;
- Ministry of Education and China Mobile joint funding through No. MCM20170301;
- Joint Institute of Tsinghua University - China Mobile Communications Group Co. Ltd.;
- Beijing National Research Center for Information Science and Technology;
- Tsinghua Initiative through No. 20121088069 and No. 20141081250;
- Toshiba Corporation;
- Apple Corporation.

Appendices

A

Notations and Definitions

A.1 Notations

Example	Description
$z_{i:j}$	For any generic sequence $\{z_n\}$, we shall use $z_{i:j}$ to denote $z_i, z_{i+1}, \cdots z_j$. Similarly, wherever a collection of indices appears in the subscript, we refer to the corresponding collection of indexed variables, e.g., $c_{l,1:H} \triangleq \{c_{l,1}, c_{l,2}, \cdots c_{l,H}\}$.
x	x generally denotes a random variable, which can either be scalar- or vector-valued, and often denotes the observation variable. For simplicity, we also use the same notation x to denote the values taken by the random variable x, e.g., in the argument of its density function, which should be clear from the context.
h	The hidden variable.
y	The class label, or the output variable.
$\|\mathcal{B}\|$	The cardinality/size of a set $\mathcal{B}$

x^T, A^T	A superscript T denotes the transpose of a vector x or matrix A
Δ^K	The K-dimensional probability simplex.
$\sum_x f(x)$	The summation over x is a shorthand, which should be an appropriate combination of summation and integration, depending on the components of x being discrete variables, continuous variables, or a combination of the two.
$p_{\text{ora}}(\cdot)$	The (unknown) oracle density, sometimes also known as the data distribution and denoted as $p_{\text{data}}(\cdot)$.
$p_{\text{emp}}(\cdot)$	The empirical density. For a training dataset consisting of n independent and identically distributed (IID) data points $\{x_1, \cdots, x_N\}$, we have $$p_{\text{emp}}(x) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i)$$
$p_\theta(\cdot), p(\cdot; \theta)$	The (target) model density, parameterized by θ.
$q_\phi(\cdot), q(\cdot; \phi)$	The auxiliary density introduced in training, parameterized by ϕ.
$\text{Uni}[a, b]$	Uniform distribution for a continuous variable over interval $[a, b]$, or for a discrete variable over integers from a to b.

A.2 Definitions

Term	Description
$\sigma(v)$	The sigmoid function, $\sigma(v) \triangleq \frac{1}{1+e^{-v}}$, also called the logistic sigmoid function. It is also known as a squashing function, since it maps the whole real line to [0, 1], which is necessary for the output to be interpreted as a probability.

$\text{logit}(\sigma)$	The logit function, $\text{logit}(\sigma) \triangleq \log(\frac{\sigma}{1-\sigma})$ for $0 < \sigma < 1$, also known as the inverse of the sigmoid function. It represents the log of the ratio of probabilities for two classes, also known as the log odds.
$\text{softmax}(z_{1:K})$	The softmax function, $\text{softmax}(z_{1:K})_k \triangleq \frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)}$, which realizes normalization from $\mathbb{R}^K$ to Δ^K (the K-dimensional probability simplex). It is also known as the normalized exponential and can be regarded as a multiclass generalization of the logistic sigmoid.
$\delta(x = a)$	An indicator function of x which takes the value 1 when $x = a$ and 0 otherwise.
$H[q]$	The entropy is defined as $H[q] \triangleq -\int qlogq$.
$KL[p\|\|q]$	The inclusive KL-divergence between two distributions $p(\cdot)$ and $q(\cdot)$ is defined as $KL[p\|\|q] \triangleq \int plog\left(\frac{p}{q}\right)$, which by default is called the KL-divergence, and is sometimes referred to as the forward KL-divergence, relative entropy.
$KL[q\|\|p]$	The exclusive KL-divergence is defined as $KL[q\|\|p] \triangleq \int qlog\left(\frac{q}{p}\right)$, which is sometimes also referred to as the reverse KL-divergence.

B

Background Material

B.1 Maximum entropy models

Theorem B.1. When confronted by a probability distribution $p(x)$ about which only a few facts are known, the maximum entropy principle (*maxent*) offers a rule for choosing a distribution that satisfies those constraints [31], [94]. According to maxent, one should select the $p(x)$ that maximizes the entropy

$$H(p) = -\sum_x p(x) \log p(x) \tag{B.1}$$

subject to the constraints. When there is a reference distribution $q(x)$, one should select the $p(x)$ that minimizes the relative entropy or Kullback-Leibler divergence[1]

$$KL(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{B.2}$$

Assuming the constraints assert that the averages of certain functions $f_k(x)$ are known, i.e.,

$$E_{p(x)}\left[f_k(x)\right] = F_k, k = 1, 2, \cdots \tag{B.3}$$

Then, it can be shown that by introducing Lagrange multipliers (one for each constraint, including normalization),

- The distribution that maximizes the entropy has the following form

$$p^*(x) = \frac{1}{Z} \exp\left(\sum_k w_k f_k(x)\right) \tag{B.4}$$

[1] When $q(x)$ is uniform, this is the same as maximizing the entropy.

- The distribution that minimizing relative entropy relative to $q(x)$, has the following form

$$p^*(x) = \frac{1}{Z} q(x) \exp\left(\sum_k w_k f_k(x)\right) \tag{B.5}$$

where $\{w_k\}$ are set such that the constraints Eq. (B.3) are satisfied, and Z is the normalizing constant. The two forms in Eq. (B.4) and Eq. (B.5) are often collectively referred to as *maximum entropy distributions*.

Theorem B.1 gives the form of maximum entropy distributions that satisfy certain moment constraints. In an opposite way, when given that a distribution satisfies the form of Eq. (B.4) or Eq. (B.5), the following theorem establish the connection between the maximum entropy distribution and the maximum likelihood distribution.

Theorem B.2. Assume that a variable x comes from a probability distribution of the form in Eq. (B.4) or Eq. (B.5), where the functions $f_k(x)$ are given, and the parameters $\{w_k\}$ are not known. A dataset $\{x^{(n)}\}$ is supplied. Then, it can be shown that by differentiating the log likelihood, the maximum-likelihood (ML) parameters w_{ML} satisfy

$$\begin{aligned} E_{p(x)}\left[f_k(x)\right] &= \frac{1}{N}\sum_n f_k(x^{(n)}), k = 1, 2, \cdots \\ &= E_{p_{\text{emp}}(x)}\left[f_k(x)\right] \end{aligned} \tag{B.6}$$

where the left-hand is the model average under the fitted model, the right-hand the empirical average over the training data points, and $p_{\text{emp}}(\cdot)$ denotes the empirical density over the training data points.

Combining the above two theorems, we can easily see that maximum entropy fitting with F_k's being set as the empirical averages is equivalent to maximum likelihood fitting of a log-linear distribution [94], [129].

B.2 Fisher equality

Formally, for any density function $p_\theta(x)$, the partial derivative w.r.t. θ of the log density function, $\frac{\partial}{\partial\theta} log p_\theta(x)$, is called the "score". Under certain regularity conditions, the expectation of the score w.r.t. the

density itself is 0. This formula is often referred in presenting Fisher information[2], so we call it *Fisher equality*, which, is frequently used in this monograph.

$$E_{p_\theta(x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x)\right] = 0. \tag{B.7}$$

Further, based on the above basic Fisher equality, we have the following very useful theorem.

Theorem B.3. Consider any latent-variable model $p_\theta(x,h)$, which consisting of observation x and latent variable h, then we have

$$\frac{\partial}{\partial\theta}\log p_\theta(x) = E_{p_\theta(h|x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x,h)\right] \tag{B.8}$$

which means that the gradient of the log marginal likelihood is equal to the expected log joint likelihood, where the expectation is taken over the posteriori distribution.

Proof.

$$\begin{aligned}\frac{\partial}{\partial\theta}\log p_\theta(x) &= E_{p_\theta(h|x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x)\right]\\ &= E_{p_\theta(h|x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x,h) - \frac{\partial}{\partial\theta}\log p_\theta(h|x)\right]\\ &= E_{p_\theta(h|x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x,h)\right]\end{aligned}$$

where in the second line, according to Fisher equality, we have

$$E_{p_\theta(h|x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(h|x)\right] = 0,$$

and thus we obtain the final line. For simplicity, Eq. (B.8) is also referred to as Fisher equality. ■

[2] https://en.wikipedia.org/wiki/Fisher_information

C

Open-source Toolkits Related to EBMs

- Trans-dimensional random field (TRF) LMs: https://github.com/thu-spmi/SPMILM
- Energy-based cloze models for representation learning over text (Electric): https://github.com/google-research/electra
- CRF-based ASR Toolkit (CAT): https://github.com/thu-spmi/CAT
- Neural CRF Transducers for Sequence Labeling: https://github.com/thu-spmi/SPMISeq
- Controlled text generation from pre-trained language models (mix-and-match): https://github.com/mireshghallah/mixmatch
- Learning neural random fields with inclusive auxiliary generators: https://github.com/thu-spmi/Inclusive-NRF
- JEMs and JRFs for semi-supervised learning: https://github.com/thu-spmi/semi-EBM

References

[1] F. Amaya and J. M. Benedi, "Improvement of a whole sentence maximum entropy language model using grammatical features," in *Proc. Ann. Meeting of the Association for Computational Linguistics (ACL)*, 2001.

[2] K. An, H. Xiang, and Z. Ou, "CAT: A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency," in *INTERSPEECH*, 2020.

[3] K. An, H. Zheng, Z. Ou, H. Xiang, K. Ding, and G. Wan, "Cuside: Chunking, simulating future context and decoding for streaming ASR," in *INTERSPEECH*, 2022.

[4] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally normalized transition-based neural networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[5] C. Andrieu, É. Moulines, and P. Priouret, "Stability of stochastic approximation under verifiable conditions," *SIAM Journal on control and optimization*, vol. 44, no. 1, 2005, pp. 283–312.

[6] C. Andrieu and J. Thoms, "A tutorial on adaptive mcmc," *Statistics and computing*, vol. 18, no. 4, 2008, pp. 343–373.

[7] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *NIPS*, 2007.

[8] T. Artieres *et al.*, "Neural conditional random fields," in *AISTATS*, 2010.

[9] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato, and A. Szlam, "Real or fake? learning to discriminate machine from human generated text," *arXiv preprint arXiv:1906.03351*, 2019.

[10] D. Belanger and A. McCallum, "Structured Prediction Energy Networks," in *ICML*, 2016.

[11] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *Advances in neural information processing systems*, 2015.

[12] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive algorithms and stochastic approximations*. New York: Springer, 1990.

[13] J. E. Besag, "Comments on "Representations of knowledge in complex systems" by U. Grenander and M.I. Miller," *Journal of the Royal Statistical Society: Series B*, vol. 56, 1994, pp. 549–581.

[14] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[15] G. Bouchard, "Bias-variance tradeoff in hybrid generative-discriminative models," in *International Conference on Machine Learning and Applications (ICMLA)*, 2007.

[16] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017.

[17] S. P. Chatzis and Y. Demiris, "The Infinite-Order Conditional Random Field Model for Sequential Data Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[18] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," in *INTERSPEECH*, 2014.

[19] H. Chen, *Stochastic approximation and its applications*. Springer Science & Business Media, 2002.

[20] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, 1999, pp. 359–394.

[21] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *ICML*, 2014.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv:2002.05709*, 2020.

[23] X. Chen, X. Liu, Y. Wang, A. Ragni, J. H. Wong, and M. J. Gales, "Exploiting future word contexts in neural network language models for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, 2019, pp. 1444–1454.

[24] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP*, 2018.

[25] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[26] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," in *International Conference on Learning Representations (ICLR)*, 2020.

[27] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Pre-training transformers as energy-based cloze models," *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[28] K. Clark, M.-T. Luong, C. D. Manning, and Q. Le, "Semi-supervised sequence modeling with cross-view training," in *EMNLP*, 2018.

[29] M. Collins and B. Roark, "Incremental Parsing with the Perceptron Algorithm," in *ACL*, 2004.

[30] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. Aug, 2011, pp. 2493–2537.

[31] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[32] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.

[33] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *CVPR*, 2020.

[34] X. Cui, B. Kingsbury, G. Saon, D. Haws, and Z. Tuske, "Reducing exposure bias in training recurrent neural network transducers," in *INTERSPEECH*, 2021.

[35] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, 2012, pp. 30–42.

[36] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov, "Good semi-supervised learning that requires a bad GAN," in *NIPS*, 2017.

[37] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The helmholtz machine," *Neural computation*, vol. 7, no. 5, 1995, pp. 889–904.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, 1977.

[39] Y. Deng, A. Bakhtin, M. Ott, A. Szlam, and M. Ranzato, "Residual energy-based models for text generation," in *ICLR*, 2020.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018, pp. 4171–4186.

[41] G. Durrett and D. Klein, "Neural CRF Parsing," in *ACL*, 2015.

[42] B. J. Frey and N. Jojic, "A comparison of algorithms for inference and learning in probabilistic graphical models," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, no. 9, 2005, pp. 1392–1416.

[43] S. Gao, Z. Ou, W. Yang, and H. Xu, "Integrating discrete and neural features via mixed-feature trans-dimensional random field language models," in *ICASSP*, 2020.

[44] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models," *arXiv preprint arXiv:1904.09324*, 2019.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[46] J. Goodman, "A bit of progress in language modeling," *Computer Speech & Language*, vol. 15, 2001, pp. 403–434.

[47] K. Goyal, C. Dyer, and T. Berg-Kirkpatrick, "Exposing the implicit energy networks behind masked language models via metropolis–hastings," in *International conference on learning representations*, 2022.

[48] W. Grathwohl, K. Swersky, M. Hashemi, D. Duvenaud, and C. Maddison, "Oops I took a gradient: Scalable sampling for discrete distributions," in *International Conference on Machine Learning*, 2021.

[49] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," in *ICLR*, 2020.

[50] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[51] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[52] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Ninth European Conference on Speech Communication and Technology (EUROSPEECH)*, 2005.

[53] C. E. Guo, S. C. Zhu, and Y. N. Wu, "Modeling visual patterns by integrating descriptive and generative methods.," *International Journal of Computer Vision*, vol. 53, no. 1, 2003, pp. 5–29.

[54] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[55] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *AISTATS*, 2010.

[56] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[57] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "Flat-start single-stage discriminatively trained HMM-based models for ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, 2018, pp. 1949–1961.

[58] T. Han, E. Nijkamp, X. Fang, M. Hill, S.-C. Zhu, and Y. N. Wu, "Divergence triangle for joint training of generator model, energy-based model, and inferential model," in *CVPR*, 2019.

[59] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[60] T. He, B. McCann, C. Xiong, and E. Hosseini-Asl, "Joint energy-based model training for better calibrated natural language understanding models," *preprint arXiv:2101.06829*, 2021.

[61] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, 2002, pp. 1771–1800.

[62] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, "The wake-sleep algorithm for unsupervised neural networks.," *Science*, vol. 268, no. 5214, 1995, pp. 1158–1161.

[63] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, 2006, pp. 1527–1554.

[64] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations (ICLR)*, 2019.

[65] K. Hu, Z. Ou, M. Hu, and J. Feng, "Neural CRF transducers for sequence labeling," in *ICASSP*, 2019.

[66] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv:1508.01991*, 2015.

[67] P. Huembeli, J. M. Arrazola, N. Killoran, M. Mohseni, and P. Wittek, "The physics of energy-based models," *Quantum Machine Intelligence*, vol. 4, no. 1, 2022, p. 1.

[68] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[69] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, 1976, pp. 532–556.

[70] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, 1999, pp. 183–233.

[71] M. I. Jordan, "Graphical models," *Statistical science*, vol. 19, no. 1, 2004, pp. 140–155.

[72] M. Khalifa, H. Elsahar, and M. Dymetman, "A distributional approach to controlled text generation," in *International conference on learning representations*, 2021.

[73] K. Kim, J. Oh, J. Gardner, A. B. Dieng, and H. Kim, "Markov chain score ascent: A unifying framework of variational inference with markovian gradients," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[74] T. Kim and Y. Bengio, "Deep directed generative models with energy-based probability estimation," in *ICLR Workshop*, 2016.

[75] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, 2019, pp. 307–392.

[76] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *NIPS*, 2014.

[77] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[78] V. Kuleshov and S. Ermon, "Neural variational inference and learning in undirected graphical models," in *NIPS*, 2017.

[79] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International conference on Machine learning (ICML)*, 2001.

[80] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, 2017.

[81] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in *NAACL-HLT*, 2016.

[82] H. Larochelle, M. I. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted Boltzmann machine," *Journal of Machine Learning Research*, vol. 13, no. 1, 2012, pp. 643–669.

[83] F. Liang, C. Liu, and R. J. Carroll, "Stochastic approximation in monte carlo computation," *Journal of the American Statistical Association*, vol. 102, no. 477, 2007, pp. 305–320.

[84] P. Liang and M. I. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," in *International conference on Machine learning (ICML)*, pp. 584–591, 2008.

[85] W. Ling, C. Dyer, A. W. Black, I. Trancoso, R. Fermandez, S. Amir, L. Marujo, and T. Luis, "Finding function in form: Compositional character models for open vocabulary word representation," in *EMNLP*, 2015.

[86] H. Liu and Z. Ou, "Exploring energy-based language models with different architectures and training methods for speech recognition," in *INTERSPEECH*, 2023.

[87] J. S. Liu, *Monte Carlo strategies in scientific computing*, vol. 10. Springer, 2001.

[88] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[89] L. Lu, L. Kong, C. Dyer, N. A. Smith, and S. Renals, "Segmental recurrent neural networks for end-to-end speech recognition," in *INTERSPEECH*, 2016.

[90] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "RWTH ASR systems for librispeech: Hybrid vs attention," in *INTERSPEECH*, 2019.

[91] Y.-A. Ma, T. Chen, and E. Fox, "A complete recipe for stochastic gradient mcmc," in *NIPS*, 2015.

[92] X. Ma and E. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," in *ACL*, 2016.

[93] Z. Ma and M. Collins, "Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency," *EMNLP*, 2018.

[94] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[95] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," 1993.

[96] S. Martin, J. Liermann, and H. Ney, "Algorithms for bigram and trigram word clustering," *Speech Communication*, vol. 24, 1998, pp. 19–37.

[97] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation.," in *ICML*, 2000.

[98] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Z. Hakkani-Tür, X. He, L. P. Heck, G. Tür, D. Yu, and G. Zweig, "Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, 2015, pp. 530–539.

[99] N. Miao, H. Zhou, L. Mou, R. Yan, and L. Li, "CGMH: Constrained sentence generation by metropolis-hastings sampling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[100] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *ASRU*, 2015.

[101] T. Mikolov, S. Kombrink, L. Burget, J. H. Cernocky, and S. Khudanpur, "Extensions of recurrent neural network language model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[102] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[103] B. Millidge, Y. Song, T. Salvatori, T. Lukasiewicz, and R. Bogacz, "Backpropagation at the infinitesimal inference limit of energy-based models: Unifying predictive coding, equilibrium propagation, and contrastive hebbian learning," in *International Conference on Machine Learning*, 2023.

[104] T. Minka, "Divergence measures and message passing," *Microsoft Research Technical Report*, 2005.

[105] F. Mireshghallah, K. Goyal, and T. Berg-Kirkpatrick, "Mix and match: Learning-free controllable text generation using energy language models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

[106] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, 2018, pp. 1979–1993.

[107] M. Mohri, F. Pereira, and M. Riley, "Speech recognition with weighted finite-state transducers," in *Springer Handbook of Speech Processing*, Springer, 2008, pp. 559–584.

[108] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-Dynamic Discriminative Models for Continuous Gesture Recognition," in *CVPR*, 2007.

[109] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng, "Sobolev GAN," in *ICLR*, 2018.

[110] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[111] C. Naesseth, F. Lindsten, and D. Blei, "Markovian score climbing: Variational inference with kl (p|| q)," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[112] R. M. Neal, *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto, Canada, 1993.

[113] R. M. Neal, "MCMC using Hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, 2011.

[114] R. M. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*, Springer, 1998, pp. 355–368.

[115] R. M. Neal, "Connectionist learning of belief networks," *Artificial Intelligence*, vol. 56, 1992, pp. 71–113.

[116] A. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, 2001.

[117] J. Ngiam, Z. Chen, P. W. Koh, and A. Y. Ng, "Learning deep energy models," in *International conference on machine learning (ICML)*, 2011.

[118] S. Nowozin, "Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference," in *International conference on learning representations*, 2018.

[119] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of semi-supervised learning algorithms," in *ICLR*, 2018.

[120] M. Ostendorf, "Continuous-space language processing: Beyond word embeddings," in *International Conference on Statistical Language and Speech Processing*, 2016.

[121] Z. Ou, "A review of learning with deep generative models from perspective of graphical modeling," *arXiv preprint arXiv:1808.01630*, 2018.

[122] Z. Ou and Y. Song, "Joint stochastic approximation and its application to learning discrete latent variable models," in *Conference on Uncertainty in Artificial Intelligence*, PMLR, pp. 929–938, 2020.

[123] Z. Ou and J. Xiao, "A study of large vocabulary speech recognition decoding using finite-state graphs," in *The 7th International Symposium on Chinese Spoken Language Processing*, 2010.

[124] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015.

[125] T. Parshakova, J.-M. Andreoli, and M. Dymetman, "Global autoregressive models for data-efficient sequence learning," *arXiv preprint arXiv:1909.07063*, 2019.

[126] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[127] J. Peng, L. Bo, and J. Xu, "Conditional Neural Fields," in *NIPS*, 2009.

[128] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[129] S. D. Pietra, V. D. Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 19, 1997, pp. 380–393.

[130] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*, PMLR, pp. 8599–8608, 2021.

[131] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *INTERSPEECH*, 2016.

[132] L. Qin, S. Welleck, D. Khashabi, and Y. Choi, "Cold decoding: Energy-based constrained text generation with langevin dynamics," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[133] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, 1989, pp. 257–286.

[134] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[135] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, 2019, p. 9.

[136] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, 2020, pp. 1–67.

[137] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *EMNLP*, 2016.

[138] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *International Conference on Learning Representations (ICLR)*, 2016.

[139] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko, "Semi-supervised learning with ladder networks," in *NIPS*, 2015.

[140] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[141] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, 1951, pp. 400–407.

[142] G. O. Roberts and J. S. Rosenthal, "Examples of adaptive mcmc," *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, 2009, pp. 349–367.

[143] G. O. Roberts and R. L. Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, 1996, pp. 341–363.

[144] R. Rosenfeld, S. F. Chen, and X. Zhu, "Whole-sentence exponential language models: A vehicle for linguistic-statistical integration," *Computer Speech & Language*, vol. 15, 2001, pp. 55–73.

[145] T. Ruokolainen, T. Alumae, and M. Dobrinkat, "Using dependency grammar features in whole sentence maximum entropy language model for speech recognition.," in *Baltic HLT*, 2010.

[146] S. Russell and P. Norvig, *Artificial intelligence: a modern approach (3rd)*. Upper Saddle River, Prentice-Hall, 2010.

[147] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," *Journal of Machine Learning Research*, vol. 5, no. 2, 2009, pp. 1967–2006.

[148] R. Salakhutdinov, "Learning deep generative models," *Ph.D. thesis, University of Toronto*, 2009.

[149] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *NIPS*, 2016.

[150] S. Sarawagi and W. W. Cohen, "Semi-Markov Conditional Random Fields for Information Extraction," in *NIPS*, 2004.

[151] R. Sarikaya, S. F. Chen, A. Sethy, and B. Ramabhadran, "Impact of word classing on shrinkage-based language models," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[152] I. Sato and H. Nakagawa, "Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process," in *ICML*, 2014.

[153] K. Sato and Y. Sakakibara, "RNA secondary structural alignment with conditional random fields," *Bioinformatics*, vol. 21, 2005, pp. 237–42.

[154] L. K. Saul, T. Jaakkola, and M. I. Jordan, "Mean field theory for sigmoid belief networks," *Journal of artificial intelligence research*, vol. 4, no. 1, 1996, pp. 61–76.

[155] B. Scellier and Y. Bengio, "Equilibrium propagation: Bridging the gap between energy-based models and backpropagation," *Frontiers in computational neuroscience*, vol. 11, 2017, p. 24.

[156] H. Schwenk, "Continuous space language models," *Computer Speech & Language*, vol. 21, 2007, pp. 492–518.

[157] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11, no. 3, 1965, pp. 363–371.

[158] N. Shazeer, J. Pelemans, and C. Chelba, "Sparse non-negative matrix language modeling for skip-grams," in *INTERSPEECH*, 2015.

[159] A. Søgaard and Y. Goldberg, "Deep multi-task learning with low level tasks supervised at lower layers," in *ACL*, pp. 231–235, 2016.

[160] K. Sohn, D. Berthelot, C.-L. Li, and *et al*, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv:2001.07685*, 2020.

[161] Q. Song, M. Wu, and F. Liang, "Weak convergence rates of population versus single-chain stochastic approximation mcmc algorithms," *Advances in Applied Probability*, vol. 46, no. 4, 2014, pp. 1059–1083.

[162] Y. Song and Z. Ou, "Learning neural random fields with inclusive auxiliary generators," *arXiv preprint arXiv:1806.00271*, 2018.

[163] Y. Song, Z. Ou, Z. Liu, and S. Yang, "Upgrading CRFs to JRFs and its benefits to sequence modeling and labeling," in *ICASSP*, 2020.

[164] Y. Song, H. Zheng, and Z. Ou, "An empirical comparison of joint-training and pre-training for domain-agnostic semi-supervised learning via energy-based models," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021.

[165] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," in *ICML*, 2016.

[166] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, 2014.

[167] W. Sun, Z. Tu, and A. Ragni, "Energy-based models for speech synthesis," *arXiv preprint arXiv:2310.12765*, 2023.

[168] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling.," in *INTERSPEECH*, pp. 194–197, 2012.

[169] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[170] C. Sutton, A. McCallum, *et al.*, "An introduction to conditional random fields," *Foundations and Trends® in Machine Learning*, vol. 4, no. 4, 2012, pp. 267–373.

[171] Z. Tan, "Optimally adjusted mixture sampling and locally weighted histogram analysis," *Journal of Computational and Graphical Statistics*, vol. 26, 2017, pp. 54–65.

[172] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, 2017.

[173] L. Theis, A. V. Den Oord, and M. Bethge, "A note on the evaluation of generative models," in *ICLR*, 2016.

[174] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *ICML*, 2008.

[175] S. Toshniwal, A. Kannan, and *et al*, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in *SLT*, 2018.

[176] L. Tu and K. Gimpel, "Learning Approximate Inference Networks for Structured Prediction," in *ICLR*, 2018.

[177] Z. Tüske, K. Audhkhasi, and G. Saon, "Advancing sequence-to-sequence based speech recognition," in *INTERSPEECH*, 2019.

[178] E. Variani, K. Wu, M. D. Riley, D. Rybach, M. Shannon, and C. Allauzen, "Global normalization for streaming speech recognition in a modular framework," *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 4257–4269.

[179] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[180] M. J. Wainwright, M. I. Jordan, *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, 2008, pp. 1–305.

[181] A. Wang and K. Cho, "BERT has a mouth, and it must speak: BERT as a Markov random field language model," in *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 2019.

[182] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *International Conference on Learning Representations (ICLR)*, 2019.

[183] B. Wang, "Statistical language models based on trans-dimensional random fields," *Ph.D. thesis, Tsinghua University*, 2018.

[184] B. Wang and Z. Ou, "Language modeling with neural trans-dimensional random fields," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017.

[185] B. Wang and Z. Ou, "Improved training of neural trans-dimensional random field language models with dynamic noise-contrastive estimation," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[186] B. Wang and Z. Ou, "Learning neural trans-dimensional random field language models with noise-contrastive estimation," in *ICASSP*, 2018.

[187] B. Wang, Z. Ou, Y. He, and A. Kawamura, "Model interpolation with trans-dimensional random field language models for speech recognition," *arXiv preprint arXiv:1603.09170*, 2016.

[188] B. Wang, Z. Ou, and Z. Tan, "Trans-dimensional random fields for language modeling," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 785–794, 2015.

[189] B. Wang, Z. Ou, and Z. Tan, "Learning trans-dimensional random fields with applications to language modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, 2018, pp. 876–890.

[190] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *ICML*, 2011.

[191] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, 1989, pp. 270–280.

[192] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," in *EMNLP*, 2016.

[193] H. Xiang and Z. Ou, "CRF-based single-stage acoustic modeling with CTC topology," in *ICASSP*, pp. 5676–5680, 2019.

[194] J. Xie, Y. Lu, R. Gao, S.-C. Zhu, and Y. N. Wu, "Cooperative training of descriptor and generator networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 1, 2018, pp. 27–45.

[195] J. Xie, Y. Lu, S.-C. Zhu, and Y. Wu, "A theory of generative convnet," in *ICML*, 2016.

[196] H. Xu and Z. Ou, "Joint stochastic approximation learning of helmholtz machines," in *ICLR Workshop Track*, 2016.

[197] L. Younes, "Parametric inference for imperfectly observed gibbsian fields," *Probability Theory and Related Fields*, vol. 82, 1989, pp. 625–645.

[198] F. Yu, Z. Yao, X. Wang, K. An, L. Xie, Z. Ou, B. Liu, X. Li, and G. Miao, "The slt 2021 children speech recognition challenge: Open datasets, rules and baselines," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[199] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv:1409.2329*, 2014.

[200] A. Zeyer, E. Beck, R. Schlüter, and H. Ney, "CTC in the context of generalized full-sum HMM training," in *INTERSPEECH*, 2017.

[201] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, and Z. Peng, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.

[202] L. Zhang, D. M. Blei, and C. A. Naesseth, "Transport score climbing: Variational inference using forward kl and adaptive neural transport," *arXiv preprint arXiv:2202.01841*, 2022.

[203] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[204] X. Zhang, Z. Tan, and Z. Ou, "Persistently trained, diffusion-assisted energy-based models," *Stat*, 2023.

[205] Y. Zhang, X. Sun, S. Ma, Y. Yang, and X. Ren, "Does Higher Order LSTM Have Better Accuracy for Segmenting and Labeling Sequence Data?" In *COLING*, 2018.

[206] Y. Zhang, Z. Ou, M. Hu, and J. Feng, "A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[207] S. Zhao, J.-H. Jacobsen, and W. Grathwohl, "Joint energy-based models for semi-supervised classification," in *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[208] H. Zheng, K. An, and Z. Ou, "Efficient neural architecture search for end-to-end speech recognition via straight-through gradients," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[209] H. Zheng, K. An, Z. Ou, C. Huang, K. Ding, and G. Wan, "An empirical study of language model integration for transducer based speech recognition," in *INTERSPEECH*, 2022.

[210] H. Zheng, W. Peng, Z. Ou, and J. Zhang, "Advancing ctc-crf based end-to-end speech recognition with wordpieces and conformers," *arXiv preprint arXiv:2107.03007*, 2021.

[211] C. Zhu, K. An, H. Zheng, and Z. Ou, "Multilingual and crosslingual speech recognition using phonological-vector based phone embeddings," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

[212] X. Zhu, "Semi-supervised learning literature survey," *Technical report, University of Wisconsin-Madison*, 2006.

[213] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.

Index