Discrete Latent Structure in Neural Networks

Other titles in Foundations and Trends® in Signal Processing

Sound Field Estimation: Theories and Applications
Natsuki Ueno and Shoichi Koyama
ISBN: 978-1-63828-524-3

Min-Max Framework for Majorization-Minimization Algorithms in Signal Processing Applications: An Overview
Astha Saini, Petre Stoica, Prabhu Babu and Aakash Arora
ISBN: 978-1-63828-466-6

Causal Deep Learning: Encouraging Impact on Real-world Problems Through Causality
Jeroen Berrevoets, Krzysztof Kacprzyk, Zhaozhi Qian and Mihaela van der Schaar
ISBN: 978-1-63828-400-0

Energy-Based Models with Applications to Speech and Language Processing
Zhijian Ou
ISBN: 978-1-63828-306-5

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

Discrete Latent Structure in Neural Networks

Vlad Niculae

Caio Corro

Nikita Nangia

Tsvetomila Mihaylova

André F. T. Martins

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

V. Niculae *et al.*. *Discrete Latent Structure in Neural Networks*. Foundations and Trends® in Signal Processing, vol. 19, no. 2, pp. 99–211, 2025.

ISBN: 978-1-63828-571-7

Foundations and Trends® in Signal Processing
Volume 19, Issue 2, 2025

Editorial Board

Editorial Scope

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, 2025, Volume 19, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Contents

Discrete Latent Structure in Neural Networks

Vlad Niculae[1], Caio Corro[2], Nikita Nangia[3], Tsvetomila Mihaylova[4] and André F. T. Martins[5,6,7]

[1] *Language Technology Lab, Informatics Institute, Faculty of Science, University of Amsterdam, Netherlands*
[2] *INSA Rennes, IRISA, Inria, CNRS, Université de Rennes, France*
[3] *Amazon, USA*
[4] *Department of Electrical Engineering and Automation, Aalto University, Finland*
[5] *Instituto Superior Técnico, Portugal*
[6] *Instituto de Telecomunicações, Portugal*
[7] *Unbabel, Portugal*

ABSTRACT

Many types of data from fields including natural language processing, computer vision, and bioinformatics are well represented by discrete, compositional structures such as trees, sequences, or matchings. Latent structure models are a powerful tool for learning to extract such representations, offering a way to incorporate structural bias, discover insight about the data, and interpret decisions. However, effective training is challenging as neural networks are typically designed for continuous computation.

This text explores three broad strategies for learning with discrete latent structure: continuous relaxation, surrogate gradients, and probabilistic estimation. Our presentation relies on consistent notations for a wide range of models.

Vlad Niculae, Caio Corro, Nikita Nangia, Tsvetomila Mihaylova and André F. T. Martins (2025), "Discrete Latent Structure in Neural Networks", Foundations and Trends® in Signal Processing: Vol. 19, No. 2, pp 99–211. DOI: 10.1561/2000000134.

As such, we reveal many new connections between latent structure learning strategies, showing how most consist of the same small set of fundamental building blocks, but use them differently, leading to substantially different applicability and properties.

Notation

Vectors, matrices, and indexing.

$u, \boldsymbol{v}, \boldsymbol{W}, \mathcal{X}$	a scalar, a vector, a matrix, and a set.
v_i	the ith element of vector $\boldsymbol{v}$.
$\boldsymbol{w}_j$	the jth column of matrix $\boldsymbol{W}$.
$\|\boldsymbol{v}\|_p$	$\coloneqq \left(\sum_{i=1}^{d} \|v_i\|^p\right)^{1/p}$, the p-norm of $\boldsymbol{v} \in \mathbb{R}^d$.

Probabilities and distributions.

Y	a random variable with values $y \in \mathcal{Y}$.
$p(\mathsf{Y} = y)$	probability that Y take the specific value y.
$p(y \mid x)$	short for $p(\mathsf{Y} = y \mid \mathsf{X} = x)$ when unambiguous.
$\mathbb{E}_{p(\mathsf{Y})}[\mathsf{Y}]$	$\coloneqq \sum_{y \in \mathcal{Y}} y p(y)$, the expected value of Y.

Differentiation.

$\partial_i f$	the partial derivative of $f : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n} \to \mathbb{R}^d$ *w.r.t.* the ith argument. $(\partial_i f)(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a linear $\mathbb{R}^d \to \mathbb{R}^{d_i}$ map (the pullback of f), identified with a $d_i \times d$ matrix: the Jacobian transpose. For single-argument $f : \mathbb{R}^{d_1} \to \mathbb{R}^d$ we omit the subscript, and if $\boldsymbol{J}_{\boldsymbol{x}}$ is the Jacobian of f at $\boldsymbol{x}$ then $\partial f(\boldsymbol{x})(\boldsymbol{v}) = \boldsymbol{J}_{\boldsymbol{x}}^\top \boldsymbol{v}$. This transposed convention is more convenient for backpropagation.
$\partial_\theta(expr.)$	interprets the (possibly complicated) expression as a single-argument function of θ and applies ∂.

Convex sets.

$\mathbb{R}^d_+$	$\coloneqq \{\boldsymbol{\alpha} \in \mathbb{R}^d; \alpha_i \geq 0 \text{ for all } 1 \leq i \leq d\}$, the non-negative orthant;
$\triangle_d$	$\coloneqq \{\boldsymbol{\alpha} \in \mathbb{R}^d_+; \sum_i \alpha_i = 1\}$, the simplex with d bins, containing all probability distributions over d choices;
$\text{conv}(\mathcal{Z})$	the convex hull of $\mathcal{Z}$, *i.e.*, the smallest convex set containing $\mathcal{Z}$.

1

Introduction

1.1 Motivation

Machine learning (ML) is often employed to build predictive models for analyzing rich data, such as images, text, or sound. Most such data is governed by underlying *structured representations*, such as segmentations, hierarchy, or graph structure. For example, natural language sentences can be analyzed in terms of their *dependency structure*, yielding an arborescence of directed grammatical relationships between words (Figure 1.1).

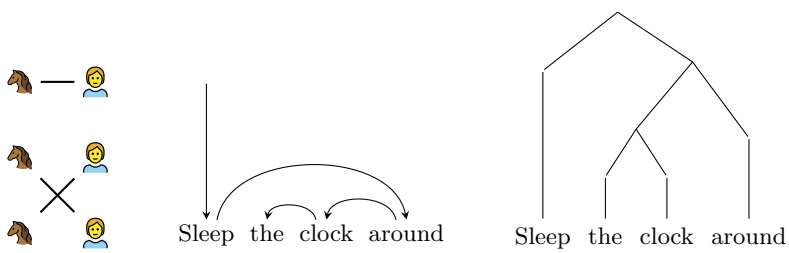

Figure 1.1: Some example structures. Left: linear assignment (matching); center: dependency parse tree (directed arborescence); right: binary constituency parse tree (binary tree).

It is common for practical ML systems to be structured as **pipelines**, including off-the-shelf components (analyzers) that produce structured representations of the input, used as features in subsequent steps of the pipeline. On the one hand, such architectures require availability of these analyzers (or of the data to train them). Since the analyzer may not be built with the downstream goal in mind, a disadvantage of pipelines is that they are prone to error propagation. On the other hand, they are transparent: the predicted structures can be directly inspected and used to interpret downstream predictions. In contrast, *deep neural networks* rival and even outperform pipelines by learning dense, continuous representations of the data, solely driven by the downstream objective.

However, the popular success of end-to-end deep learning hides some fundamental challenges. For example, large language models are still based on a pipeline system in which tokenization is an independent pre-processing step. Another known limitation is the structural generalization problem [222]: sequential architectures (both recurrent neural networks and self-attentive networks) have difficulties to generalize to unseen (recursive) combinations of known parts. It is possible to tackle this problem by inducing latent structured representations [22, 124]. Similar limitations are known for length generalization [4, 229]. Another important research direction in the natural language processing community is intermediate plan-based representations for text generation [125, 151], where latent structures may play an important role, for example, when learning with limited information [221]. Beside natural language processing, latent structure inference is also a key topic in computer vision for unsupervised segmentation and learning object-centric representations [55, 69, 129, 196].

This text is about neural network models that induce **discrete latent structure**, combining the strengths of both end-to-end and pipeline systems. In the following, we do not assume a specific downstream application in natural language processing nor computer vision. Our presentation follows an abstract framework that allows to focus on technical aspects related to end-to-end learning with deep neural networks.

1.2 Supervised Learning

We begin by establishing the common setup of predictive machine learning. A prediction function is a map associating to input $x \in \mathcal{X}$ an output $y \in \mathcal{Y}$. Prediction functions usually rely on a *scoring* function

$$M(x, y;\ \boldsymbol{\theta}), \tag{1.1}$$

which returns the score, or preference, for some candidate $y \in \mathcal{Y}$, given an input x. In our setting, M is a parametric function with learnable parameters $\boldsymbol{\theta}$. For simple classification problems, M could be a feed-forward network with x as input, and a $|\mathcal{Y}|$-dimensional output, such that the yth position of the output is $M(x, y;\ \boldsymbol{\theta})$. Our notation allows for more involved setting like predicting structured objects (for examples graphs, see Section 2). To make predictions, we search for the output of maximum weight

$$\hat{y}(x;\ \boldsymbol{\theta}) \coloneqq \arg\max_{y' \in \mathcal{Y}} M(x, y';\ \boldsymbol{\theta})\,. \tag{1.2}$$

In many cases, we are also interested in a distribution over outputs. Assuming $\mathcal{Y}$ is a finite set, a common choice is to rely on a Boltzmann-Gibbs distribution, also called *softmax* [25], defined as follows:

$$\begin{aligned} p(y \mid x) &= \frac{\exp M(x, y;\ \boldsymbol{\theta})}{\sum_{y' \in \mathcal{Y}} M(x, y';\ \boldsymbol{\theta})} \quad \text{for } y \in \mathcal{Y}\,, \\ &\propto \exp M(x, y;\ \boldsymbol{\theta})\,. \end{aligned}$$

Note that the most probable output under distribution $p(\cdot|x)$ is equal to $\hat{y}(x;\ \boldsymbol{\theta})$.

In the supervised learning scenario, we assume access to a dataset $\mathcal{D}$ containing samples of input/output pairs $(x, y) \in \mathcal{D}$. Parameters $\boldsymbol{\theta}$ are fixed to minimize the empirical risk

$$L_{\text{avg}}(\boldsymbol{\theta}) \coloneqq \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} L(y, x;\ \boldsymbol{\theta})\,, \tag{1.3}$$

where L is a loss function [208]. For practical reasons, the loss function used for classification problems is usually not the targeted evaluation function (for example the 0-1 loss which is equal to 1 if and only if

the model predicts the expected output) but a surrogate loss that is amenable for gradient-based optimization. Statistical consistency of such surrogates has been widely studied [168, 179, 217]. A common choice is the cross-entropy loss,

$$L(x, y;\ \boldsymbol{\theta}) = -M(x, y;\ \boldsymbol{\theta}) + \log \sum_{y' \in \mathcal{Y}} \exp M(x, y';\ \boldsymbol{\theta})\,, \tag{1.4}$$

which is simply the model negative log-probability of gold output under a Boltzmann-Gibbs distribution. Then, Equation 1.3 can be interpreted as *maximum likelihood* estimation of $\boldsymbol{\theta}$. Non-probabilistic losses like the hinge loss or the perceptron loss fit the framework as well.

From a computational point of view, both training and prediction under such a model eventually requires evaluating or optimizing a function of the form

$$g(x, y;\ \boldsymbol{\theta}),$$

which may refer to either the scoring model M or the loss L. Therefore, we shall use the generic functional notation $g(x, y;\ \boldsymbol{\theta})$ in the following. In this text, we are interested in computing (or approximating) partial derivatives with respect to all values in θ via the backpropagation algorithm for automatic differentiation [123].

Gradient-based learning. The gradient method for minimizing a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ iterates

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} + \eta^{(t)} (\partial F)(\boldsymbol{\theta}^{(t)})\,, \tag{1.5}$$

where $\eta^{(t)}$ is a step size schedule, and $\partial F(\cdot)$ is identified with its column-vector Jacobian. This method converges to a stationary point of F under some assumptions on the step size [16, §1.2.2]. Often in machine learning evaluating F is slow and memory-intensive, as it depends on the entire training data; this is the case in Equation 1.3. In such cases, the stochastic gradient [SG, 182] method may be preferred. The SG method replaces the gradient with a stochastic direction G such that

$$\mathbb{E}[\mathsf{G}] = \partial F(\boldsymbol{\theta}^{(t)})\,, \tag{1.6}$$

followed by updating

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} + \eta^{(t)} \mathsf{G}\,. \tag{1.7}$$

This method also converges to a stationary point under mild assumptions [17] : mainly, requiring smooth F, square-summable decreasing step sizes, and a linear bound on the variance of G *w.r.t.* the norm of the gradient of F. If F takes the form of an average, *i.e.*, $F(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} F_i(\boldsymbol{\theta})$ (for instance Equation 1.3), then G may be chosen as a single sample $F_i(\boldsymbol{\theta})$ where i is drawn uniformly from $\{1, \dots, N\}$, or a mini-batch estimator. The gradient and stochastic gradient methods can be extended to a broader family using acceleration, momentum, and adaptivity [43, 99, 130, 154]. Algorithms in this family are the *de facto* choice in deep learning at the time of writing. For this reason, our work focuses on compatibility with gradient-based learning.

Backpropagation and the Chain Rule. Given a composition of functions $u : \mathbb{R}^m \to \mathbb{R}^n$, $v : \mathbb{R}^n \to \mathbb{R}^p$, $w : \mathbb{R}^p \to \mathbb{R}^q$, and their composition $(w \circ v \circ u)(\boldsymbol{\theta}) \coloneqq w(v(u(\boldsymbol{\theta})))$ we have:

$$\partial(w \circ v \circ u)(\boldsymbol{\theta}) = (\partial u)(\boldsymbol{\theta}) \circ (\partial v)(u(\boldsymbol{\theta})) \circ (\partial w)(v(u(\boldsymbol{\theta}))) \,. \tag{1.8}$$

The derivatives are applied in the opposite order compared to the computation. This is known as *backpropagation* or *reverse-mode automatic differentiation* [70] and is popular in deep learning, where models are built using such compositions, with the final layer w having a scalar output (loss). The *forward pass* computes and stores the intermediate values that appear in $w \circ v \circ u$, and the backward pass invokes the ∂ operator to propagate gradients from the output to the input. In the most popular software frameworks today [*e.g.*, 175], elementary building blocks are provided as composable modules, with implementations providing *forward* calls $f(\boldsymbol{\theta})$ and *backward* calls (vector-Jacobian products) $\partial f(\boldsymbol{\theta})(\boldsymbol{z})$, and the automatic differentiation engine handles the composition.

1.3 Latent Representations

Our main motivation is to go beyond direct mappings $x \to y$, toward machine learning models with latent representations. In this text, we take a rather inclusive view of what constitutes a latent representation [12]. We call a latent representation $z \in \mathcal{Z}$ an object designed to capture

some relevant property of a data point $x \in \mathcal{X}$, which can be inferred based on x, but is typically unobserved. In particular, we cover but do not require probabilistic modeling of z [19]. On the other hand, we are explicitly interested in discrete and structured latent representations.

Latent representations are often designed with downstream tasks in mind: we may look for a model of $y \in \mathcal{Y}$ that has access not only to x but also to the representation z:

$$g(x, y, z;\ \boldsymbol{\theta}_g)\,. \qquad \textit{(downstream model)} \tag{1.9}$$

Remark. During prediction from a pretrained model, we may think of g as a classifier returning the score of class y. For training the model, however, we may want to think of g as some loss function on top of the same classifier. Mathematically, this distinction is irrelevant for the purpose of our text, which is the modelling of z, so we henceforth use $g(x, y, z;\ \boldsymbol{\theta}_g)$ to denote either. Practitioners should exercise caution.

A downstream model as in Equation 1.9 is not directly usable, since z is unknown both at training and at test time. Therefore, the problem we are concerned with in this text is jointly learning to predict z from x using an encoder $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$:

$$f(x, z;\ \boldsymbol{\theta}_f), \qquad \textit{(encoder model)} \tag{1.10}$$

assigning higher values to better-fitting choices of z to the given x.

The key challenge of learning latent variable models is that we cannot learn $\boldsymbol{\theta}_f$ using standard supervised approaches, since z is not observed. This text is about how to learn a good encoder model f jointly with the downstream model only from pairs (x, y). During training, the downstream model gets direct supervision, but the encoder model only gets a form of *distant supervision*, its only learning signal is coming in the form of gradients propagated through the downstream model. Joint learning with latent structure in this scenario is the main topic of our text. The next three paragraphs outline the main ways to train end-to-end models in such encoders; the main part of our text (Sections 3 to 5) later goes into detail.

Pretraining and pipelines. A first strategy is to sidestep the issue altogether and obtain supervision. This poses no challenge mathematically,

and is not studied further in this text, but serves as a motivating base case: If in fact some training pairs (x, z) are available, it is promising to first train a model $f(x, z;\ \boldsymbol{\theta}_f)$ and then deploy a two-step pipeline:

1. predict $\hat{z} = \arg\max_{z' \in \mathcal{Z}} f(x, z';\ \boldsymbol{\theta}_f)$,
2. use downstream model $g(x, y, \hat{z};\ \boldsymbol{\theta}_g)$.

The parameters of the downstream model $\boldsymbol{\theta}_g$ can now be trained in a fully-supervised fashion, since $\hat{z}$ is a known fixed input. This corresponds to the time-tested approach of using off-the-shelf analysis models (parsers, object detection, entity recognizers, etc.) as a pre-processing step. This approach is vulnerable to two main sources of error: *domain shift*, due to the fact that $\boldsymbol{\theta}_f$ is likely trained on samples coming from a different distribution than the one $\mathcal{D}$ is drawn from, and *error propagation*, due to the lack of mechanism for improving $\boldsymbol{\theta}_f$ if the model makes errors. The latent representation treatment we propose mitigates both these concerns by allowing the fine-tuning of $\boldsymbol{\theta}_f$ with signal from downstream, see [167] for examples.

Deterministic latent representations. A straightforward idea for end-to-end learning would be to characterize the mapping from x to a promising candidate $\hat{z}$ as a function,

$$\hat{z}(x; \boldsymbol{\theta}_f),$$

which implicitly defined by the encoder f. (For example, $\hat{z}(x) = \arg\max_{z \in \mathcal{Z}} f(x, z)$.) Then, an end-to-end model emerges as a composition of functions:

$$g\left(x, y, \hat{z}(x; \boldsymbol{\theta}_f);\ \boldsymbol{\theta}_g\right). \tag{1.11}$$

This resembles the pipeline approach, but now we aim to train $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$ jointly using gradient methods. Depending on how $\hat{z}$ is constructed, we may have an end-to-end differentiable relaxed model (Section 3) or a discrete model optimized with surrogate gradient heuristics (Section 4). Both cases will require further assumptions compared to the pipeline approach with frozen $\boldsymbol{\theta}_f$, but require no supervision on z.

Probabilistic latent variables. Alternatively, we can gain expressiveness by modelling latent representations as **random variables** whose distribution is induced by the encoder f. Notationally, we define a random variable Z taking values $z \in \mathcal{Z}$, with distribution $p(\mathsf{Z} = z \mid x;\ \boldsymbol{\theta}_f)$ parametrized in some way using f (*e.g.*, $p(z \mid x;\ \boldsymbol{\theta}_f) \propto \exp f(x, z;\ \boldsymbol{\theta}_f)$.) Then, the end-to-end model will consider not a single value of z but the expectation over all possible values $z \in \mathcal{Z}$:

$$\begin{aligned} \bar{g}(x, y;\ \boldsymbol{\theta}_f, \boldsymbol{\theta}_g) &:= \mathbb{E}_{\mathsf{Z}}\left[g(x, y, \mathsf{Z}; \boldsymbol{\theta}_g)\right] \\ &= \sum_{z \in \mathcal{Z}} g(x, y, z; \boldsymbol{\theta}_g) p(z \mid x; \boldsymbol{\theta}_f). \end{aligned} \tag{1.12}$$

The expected loss depends on both $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$, and so provides a learning signal to both the encoder and the downstream model. In particular, some choices of g can correspond to a probabilistic treatment of Y as well, making this strategy interesting for generative modelling. We study methods for probabilistic latent variables in Section 5. Broadly speaking, these methods tend to require fewer assumptions compared to deterministic ones, but come at a higher computational cost.

Remark. What sets apart a latent representation from an arbitrary "hidden layer" is that the former is designed to capture a specific aspect of x, relevant to the modeler. In this text, we focus on discrete z with structural constraints that can guide it to take a certain form of interest (*e.g.*, alignments, syntax.) This is often (but not necessarily) reflected in the more transparent, informed way in which the way the downstream model g accesses z.

1.4 Further History and Scope

Latent variable models have a long history in ML, especially for unsupervised learning. In this section, we briefly survey this history and clarify the scope of this work.

Shallow models. Many popular models fall under this umbrella, typically with linear f and g. Factor analysis (FA) is an unsupervised representation learning model ($\mathcal{Y} = \mathbb{R}^d$, $\mathcal{X} = \varnothing$) with continuous latent variables ($\mathcal{Z} = \mathbb{R}^k$) defined by [6, §21.1]

$$f(\boldsymbol{y}, \boldsymbol{z};\ \boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{z} - \boldsymbol{\mu})\,, \tag{1.13}$$

where the covariance $\boldsymbol{\Sigma}$ is a diagonal matrix. If $\boldsymbol{\Sigma}$ is further constrained to be isotropic, FA reduces to probabilistic PCA. The discrete counterpart is the Gaussian mixture model (GMM) where $\mathcal{Z} = \{1, 2, \ldots, k\}$ is a discrete variable, and we have

$$f(\boldsymbol{y}, z;\ \boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_z)^\top \boldsymbol{\Sigma}_z^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_z)\,. \tag{1.14}$$

For supervised regression of continuous $\boldsymbol{y}$ given $\boldsymbol{x}$, the counterpart of FA is the linear mixed effect model

$$f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x};\ \boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{z} - \boldsymbol{W}\boldsymbol{x})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{z} - \boldsymbol{W}\boldsymbol{x})\,. \tag{1.15}$$

and the counterpart of the GMM is the mixture of linear regressions

$$f(\boldsymbol{y}, z, \boldsymbol{x};\ \boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{W}_z\boldsymbol{x})^\top \boldsymbol{\Sigma}_z^{-1}(\boldsymbol{y} - \boldsymbol{W}_z\boldsymbol{x})\,, \tag{1.16}$$

corresponding to learning a separate linear regression model for each cluster component. All of the above can be fit by expectation-maximization algorithms, with the notable exception of probabilistic PCA, for which the exact solution can be found from a single SVD of the design matrix. Extensions to categorical (*i.e.*, classification) models of $\mathcal{Y}$ are mostly studied in the context of mixed effects models within the framework of hierarchical generalized linear models.

Unsupervised linguistic structure discovery. An important line of work in natural language processing is the use of latent structures for language modeling (*i.e.*, learning a distribution over sentences) in a Bayesian setting, that is by defining a Bayesian network whose observations are sentences and latent variables include structure modeling. Then, parameter inference from raw texts can provide structured representation of texts. Although useful for unsupervised and semi-supervised structured prediction, it is important to bear in mind that part of this line of work is also motivated by the goal of automatically discovering structures that may be useful for linguistic research.

Segmentation models are often used for discovering word boundaries [24, 66, 210], especially in languages that do not have explicit boundary

markers and speech processing for (non-written) low resource languages [226]. Unsupervised tagging models learn to group similar words in the same class [67]. They are mainly based on hidden Markov models, possibly with an infinite number of classes [10]. Syntactic models aim to represent more complex relations between words in a sentence than as a sequence of words. We often differentiate two types of syntactic structures:

- Such models are mainly based on latent probabilistic context-free grammars [32, 88, 117]. Phrase structures or constituency trees that model syntax by grouping words in hierarchical spans.
- Dependency trees that model syntax using bilexical dependencies between words. The main approach is called *dependency model with valence* [104].

Beyond the sentence level, previous work considered latent modeling of discourse structures [29] and topic segmentation, which aims to model topical changes in a document [50, 53].

Note that these works are not covered in this manuscript. Cohen [34] covers all basic techniques in the purely probabilistic setting (*e.g.*, parameter inference techniques like Markov chain Monte Carlo and variational inference) including the use of priors to bias models toward linguistically plausible structures. These approaches exploit probability distribution structures and their (simple) parametrization, which is not possible with the neural network setting that we cover in this manuscript. We instead focus on techniques for learning neural models in end-to-end approaches with limited assumptions, including but not limited to techniques described by Kim *et al.* [98].

Deep models. Sigmoid belief networks [SBN, 152] and Boltzmann machines [BM, 1] are popular generative neural networks with discrete latent variables that have a long history in machine learning. They are graphical models (Bayesian network in the case of SBN, factor graph in the case of BM) that use implicit parametrization using a small neural network instead of explicit contingency tables. SBNs can naturally describe deep architectures with several layers of latent variables whereas

RBMs can be stacked to achieve a similar goal [80, 191]. Straightforward approaches to fit these models are based on Markov chain Monte Carlo estimation of the gradient [79, 80, 152], which can be slow in practice. Generalization of the expectation-maximization [EM, 46] algorithm using mean field theory approximation [164] allows fast training of these models [170, 192]. A downside of EM is that it relies on strong assumption on factors' parametrization (*i.e.*, simple linear projection), and therefore does not extend to complex neural parametrization. This contrasts with methods studied in this manuscript that focus on techniques for learning discrete latent variables that (1) can learn more complex latent structures than binary variables and (2) are compatible with the modern end-to-end learning framework. Moreover, some of the techniques we described do not have a probabilistic interpretation of latent variables.

Nonlinear models parametrized by neural networks have proven themselves effective for generative modeling. Prominent among them is the *variational auto-encoder* [VAE, 100, 181], which is a Bayesian network where conditional distributions are parametrized by deep neural networks. This means that variational methods used for SBN are not applicable anymore. Key to the success of the VAE is the "evidence lower bound" (ELBO) objective

$$\begin{aligned} L(\boldsymbol{x};\boldsymbol{\theta}_g) &= -\log \mathbb{E}_{p(\mathsf{Z})}\left[p(x \mid \mathsf{Z};\boldsymbol{\theta}_g)\right] \\ &\leq \mathrm{KL}[p(\mathsf{Z} \mid x;\boldsymbol{\theta}_f), p(\mathsf{Z})] - \underbrace{\mathbb{E}_{p(\mathsf{Z}\mid x;\boldsymbol{\theta}_f)}[\log p(x \mid \mathsf{Z};\boldsymbol{\theta}_g)]}_{\text{reconstruction term}} \\ &:= \bar{L}(\boldsymbol{x},\boldsymbol{\theta}_g,\boldsymbol{\theta}_f)\,, \end{aligned} \tag{1.17}$$

where KL denotes the Kullback–Leibler divergence and $p(\mathsf{Z} \mid x;\boldsymbol{\theta}_f)$ corresponds to the approximate posterior. The conditional and approximate posterior distributions are fully specified by the Gibbs distributions

$$p(\boldsymbol{x} \mid \boldsymbol{z};\boldsymbol{\theta}_g) \propto \exp f(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta}_g)\,, \qquad p(\boldsymbol{z} \mid \boldsymbol{x};\boldsymbol{\theta}_f) \propto \exp g(\boldsymbol{z};\boldsymbol{x},\boldsymbol{\theta}_f)\,.$$

As such, the reconstruction term of the ELBO is similar to Equation 1.12.

In our framework, we may take $\boldsymbol{y} = \boldsymbol{x}$ to represent an autoencoding task, and set, for a Gaussian latent and Gaussian output VAE,

$$\begin{aligned} f(\boldsymbol{z};\boldsymbol{x},\boldsymbol{\theta}_f) &= (\boldsymbol{z} - \boldsymbol{\mu}_{\boldsymbol{z}}(\boldsymbol{x};\boldsymbol{\theta}_f))^\top \boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}(\boldsymbol{x};\boldsymbol{\theta}_f)(\boldsymbol{z} - \boldsymbol{\mu}_{\boldsymbol{z}}(\boldsymbol{x};\boldsymbol{\theta}_f))\,, \\ g(\boldsymbol{x};\boldsymbol{z};\boldsymbol{\theta}_g) &= (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}(\boldsymbol{z};\boldsymbol{\theta}_g))^\top \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\boldsymbol{z};\boldsymbol{\theta}_g)(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}(\boldsymbol{z};\boldsymbol{\theta}_g))\,, \end{aligned} \tag{1.18}$$

i.e., a neural network is used to generate the parameters of an observation distribution and of an approximate posterior; this strategy is known as amortization.

In this text, we focus on deep models with discrete, structured latent variables. This differs from works that extend the original VAE with richer priors or structured inference networks [89, 122, 166, 228, amongst others]. For a tutorial on latent variable learning with a focus on probabilistic models for language, we refer the reader to the thorough tutorial by Kim *et al.* [98].

1.5 Roadmap

Before getting into the matter of discrete latent structure, in Section 2 we revisit the tools of the trade of (supervised) structure prediction; they will prove essential for the latent case as well. Sections 3 to 5 form the main part of our text, covering three different directions to take for learning deep networks with discrete latent structure. In Section 3 we explore a deterministic approach to learning latent structure, using a fundamental *relaxation* strategy, at the cost of partially abandoning discreteness. Then, in Section 4 we discuss a range of methods that regain discreteness by introducing a gap between the learning objective and the desired model. Finally, in Section 5 we study strategies for approximately minimizing the true stochastic objective, allowing for the most flexible latent structure models, at a controllable computational cost. Section 6 summarizes the field and provides a table of various trade-offs and applicability of the discussed methods, along with pointers to prominent libraries.

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive science*, vol. 9, no. 1, 1985, pp. 147–169.

[2] R. P. Adams and R. S. Zemel, "Ranking via sinkhorn propagation," *preprint arXiv:1106.1925*, 2011.

[3] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and Z. Kolter, "Differentiable convex optimization layers," in *Advances in Neural Information Processing Systems*, 2019.

[4] C. Anil, Y. Wu, A. Andreassen, A. Lewkowycz, V. Misra, V. Ramasesh, A. Slone, G. Gur-Ari, E. Dyer, and B. Neyshabur, "Exploring length generalization in large language models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, pp. 38 546–38 556, Curran Associates, Inc., 2022. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/fb7451e43f9c1c35b774bcfad7a5714b-Paper-Conference.pdf.

[5] J. K. Baker, "Trainable grammars for speech recognition," *The Journal of the Acoustical Society of America*, vol. 65, no. S1, 1979, S132–S132. DOI: 10.1121/1.2017061.

[6] D. Barber, *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[7] J. Bastings, W. Aziz, and I. Titov, "Interpretable neural predictions with differentiable binary variables," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963–2977, Florence, Italy: Association for Computational Linguistics, 2019. DOI: 10.18653/v1/P19-1284.

[8] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process," 1972. URL: https://api.semanticscholar.org/CorpusID:60804212.

[9] C. Baziotis, I. Androutsopoulos, I. Konstas, and A. Potamianos, "Seq^3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., pp. 673–681, Minneapolis, Minnesota: Association for Computational Linguistics, 2019. DOI: 10.18653/v1/N19-1071.

[10] M. Beal, Z. Ghahramani, and C. Rasmussen, "The infinite hidden markov model," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14, MIT Press, 2001. URL: https://proceedings.neurips.cc/paper_files/paper/2001/file/e3408432c1a48a52fb6c74d926b38886-Paper.pdf.

[11] R. Bellman, "The theory of dynamic programming," *Bulletin of the American Mathematical Society*, vol. 60, no. 6, 1954, pp. 503–515.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 2013, pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

[13] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," in *Proc. of NIPS*, 2013. URL: https://arxiv.org/abs/1305.2982.

[14] Q. Berthet, M. Blondel, O. Teboul, M. Cuturi, J.-P. Vert, and F. Bach, "Learning with differentiable pertubed optimizers," *Advances in neural information processing systems*, vol. 33, 2020, pp. 9508–9519.

[15] D. P. Bertsekas, "The auction algorithm: A distributed relaxation method for the assignment problem," *Annals of Operations Research*, vol. 14, no. 1, 1988, pp. 105–123.

[16] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific Belmont, 1999. URL: http://www.athenasc.com/nonlinbook.html.

[17] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, 2000, pp. 627–642.

[18] G. Birkhoff, "Tres observaciones sobre el algebra lineal," *Univ. Nac. Tucumán Rev. Ser. A*, vol. 5, 1946, pp. 147–151.

[19] C. M. Bishop, "Latent variable models," in *Learning in Graphical Models*, M. I. Jordan, Ed., pp. 371–403, Dordrecht: Springer Netherlands, 1998. DOI: 10.1007/978-94-011-5014-9_13.

[20] M. Blondel, Q. Berthet, M. Cuturi, R. Frostig, S. Hoyer, F. Llinares-López, F. Pedregosa, and J.-P. Vert, "Efficient and modular implicit differentiation," *arXiv preprint arXiv:2105.15183*, 2021.

[21] M. Blondel, A. F. Martins, and V. Niculae, "Learning with Fenchel-Young losses," *Journal of Machine Learning Research*, vol. 21, no. 35, 2020, pp. 1–69.

[22] B. Bogin, S. Subramanian, M. Gardner, and J. Berant, "Latent compositional representations improve systematic generalization in grounded question answering," *Transactions of the Association for Computational Linguistics*, vol. 9, B. Roark and A. Nenkova, Eds., 2021, pp. 195–210. DOI: 10.1162/tacl_a_00361.

[23] J. Borwein and A. S. Lewis, *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[24] M. R. Brent, "An efficient, probabilistically sound algorithm for segmentation and word discovery," *Machine Learning*, vol. 34, 1999, pp. 71–105.

[25] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 2, Morgan-Kaufmann, 1989. URL: https://proceedings.neurips.cc/paper_files/paper/1989/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf.

[26] P. Brucker, "An $O(n)$ algorithm for quadratic knapsack problems," *Operations Research Letters*, vol. 3, no. 3, 1984, pp. 163–166. URL: https://www.sciencedirect.com/science/article/pii/0167637784900105.

[27] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, 1997, pp. 41–75.

[28] G. Casella and C. P. Robert, "Rao-blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, 1996, pp. 81–94. URL: http://www.jstor.org/stable/2337434 (accessed on 11/06/2024).

[29] H. Chen, S. Branavan, R. Barzilay, and D. R. Karger, "Content modeling using latent permutations," *Journal of Artificial Intelligence Research*, vol. 36, 2009, pp. 129–163.

[30] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar: Association for Computational Linguistics, 2014. DOI: 10.3115/v1/D14-1179.

[31] Y.-J. Chu and T.-H. Liu, "On the shortest arborescence of a directed graph," *Science Sinica*, vol. 14, 1965, pp. 1396–1400.

[32] A. Clark, "Unsupervised induction of stochastic context-free grammars using distributional clustering," in *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*, 2001. URL: https://aclanthology.org/W01-0713.

[33] J. Cocke and J. T. Schwartz, "Programming languages and their compilers: Preliminary notes," Tech. Rep., 1970.

[34] S. Cohen, *Bayesian Analysis in Natural Language Processing*, ser. Synth. Lect. Human Lang. Technol. Morgan & Claypool, 2019. DOI: 10.1007/978-3-031-02170-1.

[35] S. B. Cohen, C. Gómez-Rodrıguez, and G. Satta, "Elimination of spurious ambiguity in transition-based dependency parsing," *preprint arXiv:1206.6735*, 2012.

[36] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Mathematical Programming*, vol. 158, no. 1-2, 2016, pp. 575–585. URL: https://hal.archives-ouvertes.fr/hal-01056171.

[37] G. M. Correia, V. Niculae, and A. F. Martins, "Adaptively sparse transformers," in *Proceedings of EMNLP-IJCNLP*, 2019.

[38] G. M. Correia, V. Niculae, W. Aziz, and A. F. T. Martins, "Efficient Marginalization of Discrete and Structured Latent Variables via Sparsity," in *Proceedings of NeurIPS*, 2020. URL: http://arxiv.org/abs/2007.01919.

[39] C. Corro and I. Titov, "Differentiable perturb-and-parse: Semi-supervised parsing with a structured variational autoencoder," in *Proc. of ICLR*, 2019. URL: https://arxiv.org/abs/1807.09875.

[40] C. Corro and I. Titov, "Learning latent trees with stochastic perturbations and differentiable dynamic programming," in *Proc. of ACL*, 2019. URL: https://www.aclweb.org/anthology/P19-1551/.

[41] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proceedings of NeurIPS*, 2013.

[42] M. Cuturi and M. Blondel, "Soft-DTW: A differentiable loss function for time-series," in *Proceedings of ICML*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017. URL: https://proceedings.mlr.press/v70/cuturi17a.html.

[43] A. d'Aspremont, D. Scieur, and A. Taylor, "Acceleration methods," *Foundations and Trends® in Optimization*, vol. 5, no. 1-2, 2021, pp. 1–245. DOI: 10.1561/2400000036.

[44] G. B. Dantzig, A. Orden, and P. Wolfe, "The generalized simplex method for minimizing a linear form under linear inequality restraints," *Pacific Journal of Mathematics*, vol. 5, no. 2, 1955, pp. 183–195. URL: https://msp.org/pjm/1955/5-2/pjm-v5-n2-s.pdf.

[45] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020. URL: https://mml-book.github.io/.

[46] 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: Series B (Methodological)*, vol. 39, no. 1, 1977, pp. 1–22. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

[47] Y. Deng, Y. Kim, J. Chiu, D. Guo, and A. Rush, "Latent alignment and variational attention," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[48] J. Domke, "Learning graphical model parameters with approximate marginal inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, 2013, pp. 2454–2467.

[49] A. Drozdov and S. Bowman, "The coadaptation problem when learning how and what to compose," in *Proc of ReplNLP*, 2017.

[50] L. Du, W. Buntine, and M. Johnson, "Topic segmentation with a structured topic model," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, L. Vanderwende, H. Daumé III, and K. Kirchhoff, Eds., pp. 190–200, Atlanta, Georgia: Association for Computational Linguistics, 2013. URL: https://aclanthology.org/N13-1019.

[51] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proc. of ICML*, 2008. URL: https://stanford.edu/~jduchi/projects/DuchiShSiCh08.pdf.

[52] J. Edmonds, "Optimum branchings," *J. Res. Nat. Bur. Stand.*, vol. 71B, 1967, pp. 233–240. DOI: 10.6028/jres.071b.032.

[53] J. Eisenstein and R. Barzilay, "Bayesian unsupervised topic segmentation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, M. Lapata and H. T. Ng, Eds., pp. 334–343, Honolulu, Hawaii: Association for Computational Linguistics, 2008. URL: https://aclanthology.org/D08-1035.

[54] J. Eisner, "Inside-outside and forward-backward algorithms are just backprop (tutorial paper)," in *Proceedings of the Workshop on Structured Prediction for NLP*, K.-W. Chang, M.-W. Chang, A. Rush, and V. Srikumar, Eds., pp. 1–17, Austin, TX: Association for Computational Linguistics, 2016. DOI: 10.18653/v1/W16-5901.

[55] G. Elsayed, A. Mahendran, S. van Steenkiste, K. Greff, M. C. Mozer, and T. Kipf, "Savi++: Towards end-to-end object-centric learning from real-world videos," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, pp. 28 940–28 954, Curran Associates, Inc., 2022. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/ba1a6ba05319e410f0673f8477a871e3-Paper-Conference.pdf.

[56] A. Farinhas, W. Aziz, V. Niculae, and A. F. Martins, "Sparse communication via mixed distributions," in *International Conference on Learning Representations*, 2022.

[57] G. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, 1973, pp. 268–278. DOI: 10.1109/PROC.1973.9030.

[58] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Nav. Res. Log.*, vol. 3, no. 1-2, 1956, pp. 95–110. DOI: 10.1002/nav.3800030109.

[59] S. Frühwirth-Schnatter, "Data augmentation and dynamic linear models," *Journal of Time Series Analysis*, vol. 15, no. 2, 1994, pp. 183–202. DOI: https://doi.org/10.1111/j.1467-9892.1994.tb00184.x.

[60] Y. Fu, C. Tan, B. Bi, M. Chen, Y. Feng, and A. Rush, "Latent template induction with gumbel-crfs," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, pp. 20 259–20 271, Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc / paper / 2020 / file / ea119a40c1592979f51819b0bd38d39d - Paper.pdf.

[61] M. R. Garey and D. S. Johnson, *Computers and intractability*, vol. 174. Freeman San Francisco, 1979.

[62] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*, G. K. Thiruvathukal, Y.-H. Lu, J. Kim, Y. Chen, and B. Chen, Eds., Chapman and Hall/CRC, 2022.

[63] P. Glasserman, *Gradient estimation via perturbation analysis*, vol. 116. Springer Science & Business Media, 1990.

[64] P. W. Glynn, "Likelihood ratio gradient estimation for stochastic systems," *Communications of the ACM*, vol. 33, no. 10, 1990, pp. 75–84.

[65] A. A. Goldstein, "Convex programming in hilbert space," *Bulletin of the American Mathematical Society*, vol. 70, no. 5, 1964, pp. 709–710.

[66] S. Goldwater, T. L. Griffiths, and M. Johnson, "Contextual dependencies in unsupervised word segmentation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, N. Calzolari, C. Cardie, and P. Isabelle, Eds., pp. 673–680, Sydney, Australia: Association for Computational Linguistics, 2006. DOI: 10.3115/1220175.1220260.

[67] S. Goldwater and T. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, A. Zaenen and A. van den Bosch, Eds., pp. 744–751, Prague, Czech Republic: Association for Computational Linguistics, 2007. URL: https://aclanthology.org/P07-1094.

[68] W. Grathwohl, D. Choi, Y. Wu, G. Roeder, and D. Duvenaud, "Backpropagation through the void: Optimizing control variates for black-box gradient estimation," in *Proc. ICLR*, 2018. URL: https://arxiv.org/pdf/1711.00123.pdf.

[69] K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner, "Multi-object representation learning with iterative variational inference," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, pp. 2424–2433, PMLR, 2019. URL: https://proceedings.mlr.press/v97/greff19a.html.

[70] A. Griewank and A. Walther, *Evaluating Derivatives*, Second. Society for Industrial and Applied Mathematics, 2008. DOI: 10.1137/1.9780898717761.

[71] S. Gu, S. Levine, I. Sutskever, and A. Mnih, "Muprop: Unbiased backpropagation for stochastic neural networks," in *Proc. ICLR*, 2016. URL: https://arxiv.org/pdf/1511.05176.pdf.

[72] E. J. Gumbel, *Statistical theory of extreme values and some practical applications: a series of lectures*, vol. 33. US Government Printing Office, 1954.

[73] S. Havrylov and I. Titov, "Emergence of language with multi-agent games: Learning to communicate with sequences of symbols," in *Proc. NeurIPS*, 2017. URL: http://arxiv.org/abs/1705.11192.

[74] T. Hazan and T. Jaakkola, "On the partition function and random maximum a-posteriori perturbations," in *Proceedings of ICML*, 2012.

[75] T. Hazan, S. Maji, and T. Jaakkola, "On sampling from the gibbs distribution with random maximum a-posteriori perturbations," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26, Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/443cb001c138b2561a0d90720d6ce111-Paper.pdf.

[76] M. Held, P. Wolfe, and H. P. Crowder, "Validation of subgradient optimization," *Mathematical Programming*, vol. 6, no. 1, 1974, pp. 62–88. URL: https://link.springer.com/article/10.1007/BF01580223.

[77] G. E. Hinton, *Neural networks for machine learning*, Coursera, video lectures, 2012.

[78] G. E. Hinton and Z. Ghahramani, "Generative models for discovering sparse distributed representations," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 352, no. 1358, 1997, pp. 1177–1190.

[79] G. E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," *Neural Computation*, vol. 14, no. 8, 2002, pp. 1771–1800. DOI: 10.1162/089976602760128018.

[80] 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. DOI: 10.1162/neco.2006.18.7.1527.

[81] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of mathematics and physics*, vol. 20, no. 1–4, 1941, pp. 224–230.

[82] Y. Ho and X. Cao, "Optimization and perturbation analysis of queueing networks," *Journal of Optimization Theory and Applications*, vol. 40, no. 4, 1983, pp. 559–582.

[83] E. Hoogeboom, J. Peters, R. van den Berg, and M. Welling, "Integer discrete flows and lossless compression," in *Advances in Neural Information Processing Systems*, 2019. URL: https://papers.neurips.cc/paper/2019/hash/9e9a30b74c49d07d8150c8c83b1ccf07-Abstract.html.

[84] L. Huang, "Advanced dynamic programming in semiring and hypergraph frameworks," in *Coling 2008: Advanced Dynamic Programming in Computational Linguistics: Theory, Algorithms and Applications - Tutorial notes*, L. Huang, Ed., pp. 1–18, Manchester, UK: Coling 2008 Organizing Committee, 2008. URL: https://aclanthology.org/C08-5001.

[85] I. A. Huijben, W. Kool, M. B. Paulus, and R. J. Van Sloun, "A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[86] H. Ishwaran, J. S. Rao, *et al.*, "Spike and slab variable selection: Frequentist and bayesian strategies," *Annals of Statistics*, vol. 33, no. 2, 2005, pp. 730–773.

[87] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. of ICLR*, 2017. URL: https://arxiv.org/abs/1611.01144.

[88] M. Johnson, T. Griffiths, and S. Goldwater, "Bayesian inference for PCFGs via Markov chain Monte Carlo," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, C. Sidner, T. Schultz, M. Stone, and C. Zhai, Eds., pp. 139–146, Rochester, New York: Association for Computational Linguistics, 2007. URL: https://aclanthology.org/N07-1018.

[89] M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta, "Composing graphical models with neural networks for structured representations and fast inference," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/7d6044e95a16761171b130dcb476a43e-Paper.pdf.

[90] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, 1987, pp. 325–340. URL: https://link.springer.com/article/10.1007/BF02278710.

[91] D. Jurafsky and J. H. Martin, *Speech and Language Processing (3rd ed.)* draft, 2018. URL: https://web.stanford.edu/~jurafsky/slp3/.

[92] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, "Regularization techniques for learning with matrices," *Journal of Machine Learning Research*, vol. 13, 2012, pp. 1865–1890. URL: https://arxiv.org/abs/0910.0610.

[93] L. Kantorovich, "On the translocation of masses," *Dokl Asad Nauk SSSR*, vol. 37, no. 7–8, 1942, pp. 227–229.

[94] J. Kasai, K. Sakaguchi, R. Le Bras, D. Radev, Y. Choi, and N. A. Smith, "A call for clarity in beam search: How it works and when it stops," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., pp. 77–90, Torino, Italia: ELRA and ICCL, 2024. URL: https://aclanthology.org/2024.lrec-main.7.

[95] T. Kasami, "An efficient recognition and syntax-analysis algorithm for context-free languages," Tech. Rep., 1966.

[96] K. Keith, S. L. Blodgett, and B. O'Connor, "Monte Carlo syntax marginals for exploring and using dependency parses," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 2018. DOI: 10.18653/v1/N18-1084.

[97] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," in *Proc. of ICLR*, 2017. URL: https://arxiv.org/abs/1702.00887.

[98] Y. Kim, S. Wiseman, and A. M. Rush, "A tutorial on deep latent variable models of natural language," *arXiv preprint arXiv:1812.06834*, 2018.

[99] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR*, 2015. URL: https://arxiv.org/abs/1412.6980.

[100] D. P. Kingma and M. Welling, "Auto-encoding Variational Bayes," in *Proceedings of ICLR*, 2014.

[101] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/file/d523773c6b194f37b938d340d5d02232-Paper.pdf.

[102] E. Kiperwasser and Y. Goldberg, "Simple and accurate dependency parsing using bidirectional lstm feature representations," *TACL*, vol. 4, 2016, pp. 313–327. URL: https://aclweb.org/anthology/Q16-1023.

[103] G. Kirchhoff, "Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird," *Annalen der Physik*, vol. 148, no. 12, 1847, pp. 497–508.

[104] D. Klein and C. Manning, "Corpus-based induction of syntactic structure: Models of dependency and constituency," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 478–485, Barcelona, Spain, 2004. DOI: 10.3115/1218955.1219016.

[105] N. Komodakis, N. Paragios, and G. Tziritas, "Mrf optimization via dual decomposition: Message-passing revisited.," in *ICCV*, vol. 1, p. 5, 2007.

[106] W. Kool, H. van Hoof, and M. Welling, "Ancestral gumbel-top-k sampling for sampling without replacement," *Journal of Machine Learning Research*, vol. 21, no. 47, 2020, pp. 1–36. URL: http://jmlr.org/papers/v21/19-985.html.

[107] W. Kool, H. van Hoof, and M. Welling, "Estimating gradients for discrete random variables by sampling without replacement," in *Proc. ICLR*, 2020. URL: https://openreview.net/forum?id=rklEj2EFvB.

[108] W. Kool, H. Van Hoof, and M. Welling, "Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, pp. 3499–3508, PMLR, 2019. URL: https://proceedings.mlr.press/v97/kool19a.html.

[109] E. van Krieken, J. M. Tomczak, and A. t. Teije, "Storchastic: A framework for general stochastic automatic differentiation," vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 7574–7587. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/3dfe2f633108d604df160cd1b01710db-Paper.pdf.

[110] H. W. Kuhn, "The hungarian method for the assignment problem," *Nav. Res. Log.*, vol. 2, no. 1-2, 1955, pp. 83–97. URL: http://onlinelibrary.wiley.com/doi/10.1002/nav.3800020109/abstract.

[111] P. Kumaraswamy, "A generalized probability density function for double-bounded random processes," *Journal of hydrology*, vol. 46, no. 1-2, 1980, pp. 79–88.

[112] A. Kyrillidis, S. Becker, V. Cevher, and C. Koch, "Sparse projections onto the simplex," in *Proc. ICML*, 2013. URL: http://proceedings.mlr.press/v28/kyrillidis13.pdf.

[113] P. L'Ecuyer, "Note: On the interchange of derivative and expectation for likelihood ratio derivative estimators," *Management Science*, vol. 41, no. 4, 1995, pp. 738–747.

[114] S. Lacoste-Julien and M. Jaggi, "On the global linear convergence of frank-wolfe optimization variants," in *Proc. of NIPS*, 2015. URL: https://arxiv.org/abs/1511.05932.

[115] J. Lafferty, A. McCallum, F. Pereira, *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Icml*, Williamstown, MA, vol. 1, p. 3, 2001.

[116] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proceedings of ICLR*, 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[117] K. Lari and S. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm," *Computer Speech & Language*, vol. 4, no. 1, 1990, pp. 35–56. DOI: https://doi.org/10.1016/0885-2308(90)90022-X.

[118] A. Lazaridou, A. Peysakhovich, and M. Baroni, "Multi-agent cooperation and the emergence of (natural) language," in *Proc. ICLR*, 2017. URL: http://arxiv.org/abs/1612.07182.

[119] E. S. Levitin and B. T. Polyak, "Constrained minimization methods," *USSR Computational mathematics and mathematical physics*, vol. 6, no. 5, 1966, pp. 1–50.

[120] Z. Li and J. Eisner, "First- and second-order expectation semirings with applications to minimum-risk training on translation forests," in *Proc. of EMNLP*, 2009. URL: http://www.mt-archive.info/EMNLP-2009-Li.pdf.

[121] P. Liang, M. I. Jordan, and D. Klein, "Learning programs: A hierarchical Bayesian approach," in *Proceedings of ICML*, 2010.

[122] W. Lin, M. E. Khan, and N. Hubacher, "Variational message passing with structured inference networks," in *International Conference on Learning Representations*, 2018. URL: https://openreview.net/forum?id=HyH9lbZAW.

[123] S. Linnainmaa, "The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors," Ph.D. dissertation, Master's Thesis (in Finnish), Univ. Helsinki, 1970.

[124] C. Liu, S. An, Z. Lin, Q. Liu, B. Chen, J.-G. Lou, L. Wen, N. Zheng, and D. Zhang, "Learning algebraic recombination for compositional generalization," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., pp. 1129–1144, Online: Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.findings-acl.97.

[125] D. Liu, M. Lapata, and F. Keller, "Visual storytelling with question-answer plans," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., pp. 5800–5813, Singapore: Association for Computational Linguistics, 2023. DOI: 10.18653/v1/2023.findings-emnlp.386.

[126] R. Liu, J. Regier, N. Tripuraneni, M. Jordan, and J. Mcauliffe, "Rao-blackwellized stochastic gradients for discrete distributions," in *Proc. ICML*, 2019. URL: http://proceedings.mlr.press/v97/liu19c/liu19c.pdf.

[127] Y. Liu, M. Gardner, and M. Lapata, "Structured alignment networks for matching sentences," in *Proceedings of EMNLP*, Association for Computational Linguistics, 2018. DOI: 10.18653/v1/D18-1184.

[128] Y. Liu and M. Lapata, "Learning structured text representations," *TACL*, vol. 6, 2018, pp. 63–75. URL: https://arxiv.org/abs/1705.09207.

[129] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, pp. 11 525–11 538, Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/8511df98c02ab60aea1b2356c013bc0f-Paper.pdf.

[130] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[131] C. Louizos, M. Welling, and D. P. Kingma, "Learning sparse neural networks through L_0 regularization," in *International Conference on Learning Representations*, 2018. URL: https://openreview.net/forum?id=H1Y8hhg0b.

[132] B. T. Lowerre, "The HARPY speech recognition system," PhD, Carnegie-Mellon University, 1976.

[133] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *Proc. of ICLR*, 2017. URL: https://arxiv.org/abs/1611.00712.

[134] A. Martins, N. Smith, and E. Xing, "Concise integer linear programming formulations for dependency parsing," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, K.-Y. Su, J. Su, J. Wiebe, and H. Li, Eds., pp. 342–350, Suntec, Singapore: Association for Computational Linguistics, 2009. URL: https://aclanthology.org/P09-1039.

[135] A. F. Martins and R. F. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proc. of ICML*, 2016. URL: https://arxiv.org/abs/1602.02068.

[136] A. F. Martins, M. A. Figueiredo, P. M. Aguiar, N. A. Smith, and E. P. Xing, "Ad3: Alternating directions dual decomposition for map inference in graphical models," *JMLR*, vol. 16, no. 1, 2015, pp. 495–545. URL: http://jmlr.org/papers/v16/martins15a.html.

[137] A. F. Martins, N. A. Smith, and E. P. Xing, "Concise integer linear programming formulations for dependency parsing," in *Proc. of ACL-IJCNLP*, 2009. URL: http://www.aclweb.org/anthology/P09-1039.

[138] C. Meister, A. Amini, T. Vieira, and R. Cotterell, "Conditional Poisson stochastic beams," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., pp. 664–681, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.emnlp-main.52.

[139] C. Meister, R. Cotterell, and T. Vieira, "If beam search is the answer, what was the question?" In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., pp. 2173–2185, Online: Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.emnlp-main.170.

[140] C. Meister, T. Vieira, and R. Cotterell, "Best-first beam search," *Transactions of the Association for Computational Linguistics*, vol. 8, M. Johnson, B. Roark, and A. Nenkova, Eds., 2020, pp. 795–809. DOI: 10.1162/tacl_a_00346.

[141] G. Mena, D. Belanger, S. Linderman, and J. Snoek, "Learning latent permutations with gumbel-sinkhorn networks," in *Proc. of ICLR*, 2018. URL: https://arxiv.org/abs/1802.08665.

[142] A. Mensch and M. Blondel, "Differentiable dynamic programming for structured prediction and attention," in *Proc. of ICML*, 2018. URL: https://arxiv.org/abs/1802.03676.

[143] O. Meshi, M. Mahdavi, and A. G. Schwing, "Smooth and strong: MAP inference with linear convergence," in *Proc. of NIPS*, 2015. URL: https://papers.nips.cc/paper/5710-smooth-and-strong-map-inference-with-linear-convergence.

[144] N. Metropolis and S. Ulam, "The monte carlo method," *Journal of the American Statistical Association*, vol. 44, no. 247, 1949, pp. 335–341. DOI: 10.1080/01621459.1949.10483310.

[145] T. Mihaylova, V. Niculae, and A. F. T. Martins, "Understanding the mechanics of SPIGOT: Surrogate gradients for latent structure learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.emnlp-main.171.

[146] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, 1988, pp. 1023–1032.

[147] A. Mnih and K. Gregor, "Neural variational inference and learning in belief networks," in *Proceedings of ICML*, 2014. URL: http://arxiv.org/abs/1402.0030.

[148] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, "Monte carlo gradient estimation in machine learning.," *J. Mach. Learn. Res.*, vol. 21, no. 132, 2020, pp. 1–62.

[149] M. Mohri, "Semiring frameworks and algorithms for shortest-distance problems," *J. Autom. Lang. Comb.*, vol. 7, no. 3, 2002, pp. 321–350.

[150] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL: probml.ai.

[151] S. Narayan, J. Maynez, R. K. Amplayo, K. Ganchev, A. Louis, F. Huot, A. Sandholm, D. Das, and M. Lapata, "Conditional generation with a question-answering blueprint," *Transactions of the Association for Computational Linguistics*, vol. 11, 2023, pp. 974–996. DOI: 10.1162/tacl_a_00583.

[152] R. M. Neal, "Connectionist learning of belief networks," *Artificial Intelligence*, vol. 56, no. 1, 1992, pp. 71–113. DOI: https://doi.org/10.1016/0004-3702(92)90065-6.

[153] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, 1970, pp. 443–453. DOI: https://doi.org/10.1016/0022-2836(70)90057-4.

[154] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," in *Dokl akad nauk Sssr*, vol. 269, p. 543, 1983.

[155] V. Niculae and M. Blondel, "A regularized framework for sparse and structured neural attention," in *Proc. of NIPS*, 2017. URL: https://arxiv.org/abs/1705.07704.

[156] V. Niculae and A. Martins, "LP-SparseMAP: Differentiable relaxed optimization for sparse structured prediction," in *Proceedings of ICML*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020. URL: https://proceedings.mlr.press/v119/niculae20a.html.

[157] V. Niculae, A. F. Martins, M. Blondel, and C. Cardie, "SparseMAP: Differentiable sparse structured inference," in *Proc. of ICML*, 2018.

[158] M. Niepert, P. Minervini, and L. Franceschi, "Implicit mle: Backpropagating through discrete exponential family distributions," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 14 567–14 579.

[159] J. Nocedal and S. Wright, *Numerical Optimization*. Springer New York, 1999. URL: https://doi.org/10.1007/b98874.

[160] A. van den Oord, O. Vinyals, and k. kavukcuoglu koray, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

[161] J. Paisley, D. M. Blei, and M. I. Jordan, "Variational bayesian inference with stochastic search," in *Proc. ICML*, 2012. URL: https://arxiv.org/abs/1206.6430.

[162] A. W. Palmer, A. J. Hill, and S. J. Scheding, "Methods for stochastic collection and replenishment (scar) optimisation for persistent autonomy," *Robotics and Autonomous Systems*, vol. 87, 2017, pp. 51–65.

[163] G. Papandreou and A. L. Yuille, "Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models," in *2011 International Conference on Computer Vision*, IEEE, pp. 193–200, 2011.

[164] G. Parisi and R. Shankar, "Statistical field theory," 1988.

[165] M. Paulus, D. Choi, D. Tarlow, A. Krause, and C. J. Maddison, "Gradient estimation with stochastic softmax tricks," *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5691–5704.

[166] M. Pearce, "The gaussian process prior vae for interpretable latent dynamics from pixels," in *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, C. Zhang, F. Ruiz, T. Bui, A. B. Dieng, and D. Liang, Eds., ser. Proceedings of Machine Learning Research, vol. 118, pp. 1–12, PMLR, 2020. URL: https://proceedings.mlr.press/v118/pearce20a.html.

[167] H. Peng, S. Thomson, and N. A. Smith, "Backpropagating through structured argmax using a SPIGOT," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1863–1873, Melbourne, Australia: Association for Computational Linguistics, 2018. DOI: 10.18653/v1/P18-1173.

[168] M. I. J. Peter L Bartlett and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, 2006, pp. 138–156. DOI: 10.1198/016214505000000907.

[169] B. Peters, V. Niculae, and A. F. Martins, "Sparse sequence-to-sequence models," in *Proc. ACL*, 2019. URL: https://arxiv.org/abs/1905.05702.

[170] C. Peterson, "A mean field theory learning algorithm for neural network," *Complex systems*, vol. 1, 1987, pp. 995–1019.

[171] G. Peyré and M. Cuturi, "Computational optimal transport," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, 2019, pp. 355–607.

[172] G. C. Pflug, *Optimization of stochastic models: the interface between simulation and optimization*, vol. 373. Springer Science & Business Media, 2012.

[173] M. V. Pogančić, A. Paulus, V. Musil, G. Martius, and M. Rolinek, "Differentiation of blackbox combinatorial solvers," in *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=BkevoJSYPB.

[174] V. Punyakanok, D. Roth, W.-t. Yih, and D. Zimak, "Semantic role labeling via integer linear programming inference," in *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1346–1352, 2004.

[175] PyTorch Foundation, *Pytorch*, 2017. URL: http://pytorch.org.

[176] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *P. IEEE*, vol. 77, no. 2, 1989, pp. 257–286. URL: https://doi.org/10.1109/5.18626.

[177] R. Ranganath, S. Gerrish, and D. Blei, "Black Box Variational Inference," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, S. Kaski and J. Corander, Eds., ser. Proceedings of Machine Learning Research, vol. 33, pp. 814–822, Reykjavik, Iceland: PMLR, 2014. URL: https://proceedings.mlr.press/v33/ranganath14.html.

[178] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, 2011, pp. 333–359.

[179] M. D. Reid and R. C. Williamson, "Composite binary losses," *Journal of Machine Learning Research*, vol. 11, no. 83, 2010, pp. 2387–2422. URL: http://jmlr.org/papers/v11/reid10a.html.

[180] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008–7024, 2017.

[181] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, pp. 1278–1286, Bejing, China: PMLR, 2014. URL: https://proceedings.mlr.press/v32/rezende14.html.

[182] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, 1951, pp. 400–407. DOI: 10.1214/aoms/1177729586.

[183] J. T. Rolfe, "Discrete variational autoencoders," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017. URL: https://openreview.net/forum?id=ryMxXPFex.

[184] D. Roth and W.-t. Yih, "Integer linear programming inference for conditional random fields," in *Proceedings of the 22nd international conference on Machine learning*, pp. 736–743, 2005.

[185] R. Y. Rubinstein, "Sensitivity analysis of discrete event systems by the "push out" method," *Annals of Operations Research*, vol. 39, no. 1, 1992, pp. 229–250.

[186] R. Rubinstein, "A Monte Carlo method for estimating the gradient in a stochastic network," *Unpublished manuscript, Technion, Haifa, Israel*, 1976.

[187] A. M. Rush, D. Sontag, M. Collins, and T. Jaakkola, "On dual decomposition and linear programming relaxations for natural language processing," Association for Computational Linguistics, 2010.

[188] A. M. Rush, *Torch-struct: Deep structured prediction library*, 2020. arXiv: 2002.00876 [cs.CL].

[189] K. Sagae and A. Lavie, "A classifier-based parser with linear runtime complexity," in *Proceedings of the Ninth International Workshop on Parsing Technology*, pp. 125–132, Vancouver, British Columbia: Association for Computational Linguistics, 2005. URL: https://aclanthology.org/W05-1513.

[190] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, 1978, pp. 43–49. DOI: 10.1109/TASSP.1978.1163055.

[191] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, D. van Dyk and M. Welling, Eds., ser. Proceedings of Machine Learning Research, vol. 5, pp. 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, 2009. URL: https://proceedings.mlr.press/v5/salakhutdinov09a.html.

[192] L. K. Saul, T. Jaakkola, and M. I. Jordan, "Mean field theory for sigmoid belief networks," *Journal of artificial intelligence research*, vol. 4, 1996, pp. 61–76.

[193] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *The annals of mathematical statistics*, vol. 35, no. 2, 1964, pp. 876–879.

[194] M. Stanojević, "Unbiased and efficient sampling of dependency trees," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., pp. 1691–1706, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. DOI: 10.18653/v1/2022.emnlp-main.110.

[195] M. Stanojević and L. Sartran, "SynJax: Structured Probability Distributions for JAX," *arXiv preprint arXiv:2308.03291*, 2023. URL: https://arxiv.org/abs/2308.03291.

[196] S. van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber, "Relational neural expectation maximization: Unsupervised discovery of objects and their interactions," in *International Conference on Learning Representations*, 2018. URL: https://openreview.net/forum?id=ryH20GbRW.

[197] V. Stoyanov, A. Ropson, and J. Eisner, "Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure," in *Proc. of AISTATS*, 2011. URL: http://proceedings.mlr.press/v15/stoyanov11a.html.

[198] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

[199] B. Taskar, "Learning structured prediction models: A large margin approach," Ph.D. dissertation, Stanford University, 2004. URL: https://homes.cs.washington.edu/~taskar/pubs/thesis.pdf.

[200] Y. Tay, D. Bahri, L. Yang, D. Metzler, and D.-C. Juan, "Sparse Sinkhorn attention," in *Proceedings of ICML*, 2020.

[201] M. Titsias and M. Lázaro-Gredilla, "Doubly stochastic variational bayes for non-conjugate inference," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, pp. 1971–1979, Bejing, China: PMLR, 2014. URL: https://proceedings.mlr.press/v32/titsias14.html.

[202] M. K. Titsias and M. Lázaro-Gredilla, "Local expectation gradients for black box variational inference," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/1373b284bc381890049e92d324f56de0-Paper.pdf.

[203] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, 2007, pp. 1–13. URL: https://EconPapers.repec.org/RePEc:igg:jdwm00:v:3:y:2007:i:3:p:1-13.

[204] G. Tucker, A. Mnih, C. J. Maddison, D. Lawson, and J. Sohl-Dickstein, "Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models," in *Proceedings of NeurIPS*, 2017. URL: http://arxiv.org/abs/1703.07370.

[205] W. T. Tutte, *Graph theory*, vol. 21. Cambridge university press, 2001.

[206] A. Vahdat, W. Macready, Z. Bian, A. Khoshaman, and E. Andriyash, "DVAE++: Discrete variational autoencoders with overlapping transformations," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, pp. 5035–5044, PMLR, 2018. URL: https://proceedings.mlr.press/v80/vahdat18a.html.

[207] L. Valiant, "The complexity of computing the permanent," *Theoretical Computer Science*, vol. 8, no. 2, 1979, pp. 189–201. DOI: https://doi.org/10.1016/0304-3975(79)90044-6.

[208] V. Vapnik, "Principles of risk minimization for learning theory," in *Advances in Neural Information Processing Systems*, J. Moody, S. Hanson, and R. Lippmann, Eds., vol. 4, Morgan-Kaufmann, 1991. URL: https://proceedings.neurips.cc/paper_files/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf.

[209] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[210] A. Venkataraman, "A statistical model for word discovery in transcribed speech," *Computational Linguistics*, vol. 27, no. 3, J. Hirschberg, Ed., 2001, pp. 351–372. DOI: 10.1162/089120101317066113.

[211] T. K. Vintsyuk, "Speech discrimination by dynamic programming," *Cybernetics*, vol. 4, no. 1, 1968, pp. 52–57.

[212] M. Vinyes and G. Obozinski, "Fast column generation for atomic norm regularization," in *Proc. of AISTATS*, 2017. URL: http://proceedings.mlr.press/v54/vinyes17a.html.

[213] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, 1967, pp. 260–269. DOI: 10.1109/TIT.1967.1054010.

[214] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. ACM*, vol. 21, no. 1, 1974, pp. 168–173. DOI: 10.1145/321796.321811.

[215] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, 2008, pp. 1–305. URL: https://people.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.pdf.

[216] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3-4, 1992, pp. 229–256. URL: http://link.springer.com/10.1007/BF00992696.

[217] R. C. Williamson, E. Vernet, and M. D. Reid, "Composite multiclass losses," *Journal of Machine Learning Research*, 2016. URL: http://jmlr.org/papers/v17/14-294.html.

[218] P. Wolfe, "Finding the nearest point in a polytope," *Mathematical Programming*, vol. 11, no. 1, 1976, pp. 128–149. URL: https://link.springer.com/article/10.1007/BF01580381.

[219] C. Wong, K. M. Ellis, J. Tenenbaum, and J. Andreas, "Leveraging language to learn program abstractions and search heuristics," in *Proceedings of ICML*, PMLR, 2021.

[220] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, pp. 2048–2057, Lille, France: PMLR, 2015. URL: https://proceedings.mlr.press/v37/xuc15.html.

[221] Y. Xu and M. Lapata, "Document summarization with latent queries," *Transactions of the Association for Computational Linguistics*, vol. 10, B. Roark and A. Nenkova, Eds., 2022, pp. 623–638. DOI: 10.1162/tacl_a_00480.

[222] Y. Yao and A. Koller, "Structural generalization is hard for sequence-to-sequence models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., pp. 5048–5062, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. DOI: 10.18653/v1/2022.emnlp-main.337.

[223] J. I. Yellott Jr., "The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution," *Journal of Mathematical Psychology*, vol. 15, no. 2, 1977, pp. 109–144.

[224] D. H. Younger, "Recognition and parsing of context-free languages in time n3," *Information and Control*, vol. 10, no. 2, 1967, pp. 189–208. DOI: https://doi.org/10.1016/S0019-9958(67)80007-X.

[225] C. Zălinescu, *Convex Analysis in General Vector Spaces*. World Scientific, 2002. URL: http : / / www . worldscientific . com / worldscibooks/10.1142/5021.

[226] M. Zanon Boito, B. Yusuf, L. Ondel, A. Villavicencio, and L. Besacier, "Unsupervised word segmentation from discrete speech units in low-resource settings," in *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, M. Melero, S. Sakti, and C. Soria, Eds., pp. 1–9, Marseille, France: European Language Resources Association, 2022. URL: https://aclanthology.org/2022.sigul-1.1.

[227] V. Zantedeschi, J. Kaddour, L. Franceschi, M. Kusner, and V. Niculae, "Dag learning on the permutahedron," in *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

[228] Y. Zhao and S. Linderman, "Revisiting structured variational autoencoders," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, pp. 42 046–42 057, PMLR, 2023. URL: https://proceedings.mlr.press/v202/zhao23c.html.

[229] H. Zhou, A. Bradley, E. Littwin, N. Razin, O. Saremi, J. M. Susskind, S. Bengio, and P. Nakkiran, "What algorithms can transformers learn? a study in length generalization," in *The Twelfth International Conference on Learning Representations*, 2024. URL: https://openreview.net/forum?id=AssIuHnmHX.

[230] R. Zmigrod, T. Vieira, and R. Cotterell, "Efficient sampling of dependency structure," in *Proceedings of EMNLP*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.emnlp-main.824.