

# The Data Science and Statistical Learning Journal Club @ CSU: Introduction

**DSSL @ CSU Team**

Sep, 9th, 2020

## DSSL @ CSU

About

Schedule

Archives

RSS



## About

The Data Science and Statistical Learning Journal Club @ CSU meets weekly to discuss papers and current work on topics relevant to data science and statistical learning. The journal club began meeting in Fall 2020 and is organized by [Wen Zhou](#), [Andee Kaplan](#), and [Haonan Wang](#), all in the [Department of Statistics @ CSU](#).

At the beginning of each semester, together with all participants, we will select a few interesting and latest manuscripts to study. Students are expected to actively participate in the discussion.

## How to Join

To accommodate the current pandemic situation, we will use Zoom to meet weekly. Each meeting will last for approximately one hour.

- Meeting times: TBD
- Zoom link: TBD

For a password to join the meeting, please send an e-mail to [dssl.csu@gmail.com](mailto:dssl.csu@gmail.com) with subject "Zoom Password for Weekly Meeting".

- Weekly meeting on Wednesday 4pm (MST)
- Contact email for DSSL: `dssl@stat.colostate.edu` or `dssl.csu@gmail.com`
- Papers or manuscripts from interesting topics will be picked up by the group, presented by students, and discussed
- Each paper may take 2-3 weeks for presentation and discussion
- Some meetings may have speakers from outside
- Research oriented

“One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.”

Statistical Science  
2001, Vol. 16, No. 3, 199-231

## Statistical Modeling: The Two Cultures

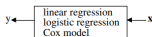
Leo Breiman

*Abstract.* There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

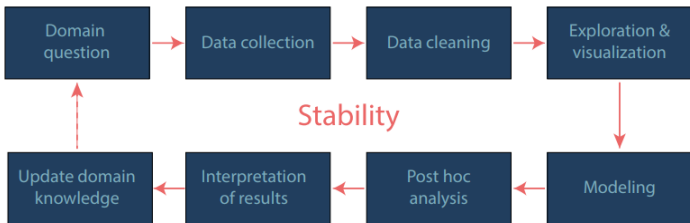
### 1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables  $\mathbf{x}$  (independent variables) go in one side, and on the other side the response variables  $\mathbf{y}$  come out. Inside the black box, nature functions to associate the predictor variables with the response variables so the picture is like this:

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



“Three core principles, **predictability, computability, and stability (PCS)**, provide the foundation for such a data driven language and a unified data analysis framework. They serve as minimum requirements for veridical data science.”

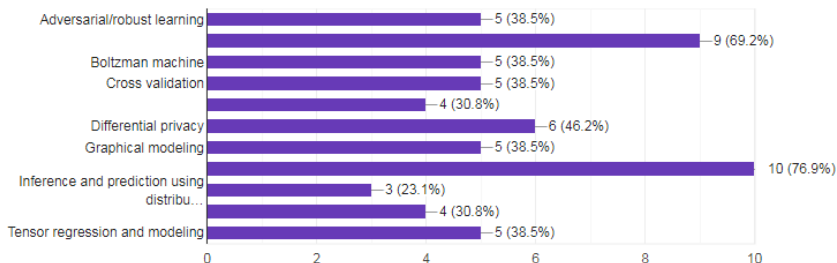


– Prof. B. Yu, PNAS, 2020

# Some Topics

- Adversarial/robust learning
- Bayesian network and causal inference
- Boltzman machine
- Cross validation
- Conformal prediction (and/or knock off)
- Differential privacy
- Graphical modeling
- Statistical understanding on neural networks: AMP, double descent, mean field, RF model
- Inference and prediction using distributed optimization
- Topic learning and mining
- Tensor regression and modeling

13 responses



# Adversarial/robust learning

## Theoretically Principled Trade-off between Robustness and Accuracy

Hongyang Zhang<sup>\*</sup>  
CMU & TTIC  
hongyanz@cs.cmu.edu

Yaodong Yu<sup>†</sup>  
University of Virginia  
yy8m@virginia.edu

Jiantao Jiao  
UC Berkeley  
jiantao@eecs.berkeley.edu

Eric P. Xing  
CMU & Petuum Inc.  
eping@cs.cmu.edu

Laurent El Ghaoui  
UC Berkeley  
elghaoui@berkeley.edu

Michael I. Jordan  
UC Berkeley  
jordan@cs.berkeley.edu

j] 24 Jun 2019

### Certifying Some Distributional Robustness with Principled Adversarial Training

Aman Sinha<sup>\*1</sup> Hongseok Namkoong<sup>\*2</sup> Riccardo Volpi<sup>3</sup> John Duchi<sup>1,4</sup>

<sup>1</sup>Departments of <sup>1</sup>Electrical Engineering, <sup>2</sup>Management Science & Engineering,

<sup>3</sup>Pattern Analysis & Computer Vision, and <sup>4</sup>Statistics

<sup>1,2,4</sup>Stanford University, <sup>3</sup>Istituto Italiano di Tecnologia

{[@stanford.edu](mailto:aman,hnamk,jduchi), [riccardo.volpi@iit.it](mailto:riccardo.volpi@iit.it)

#### Abstract

Neural networks are vulnerable to adversarial examples and researchers have proposed many heuristic attack and defense mechanisms. We address this problem through the principled lens of distributionally robust optimization, which guarantees performance under adversarial input perturbations. By considering a Lagrangian penalty formulation of perturbing the underlying data distribution in a Wasserstein ball, we provide a training procedure that augments model parameter updates with worst-case perturbations of training data. For smooth losses, our procedure provably achieves moderate levels of robustness with little computational or statistical cost relative to empirical risk minimization. Furthermore, our statistical guarantees allow us to efficiently certify robustness for the population loss. For imperceptible perturbations, our method matches or outperforms heuristic approaches.

## 1 Introduction

Consider the classical stochastic optimization problem, in which we minimize an expected loss  $E_{\theta_0}[\ell(\theta; Z)]$  over a parameter  $\theta \in \Theta$ , where  $Z \sim P_0$ ,  $P_0$  is a distribution on a space  $Z$ , and  $\ell$  is a loss function. In many systems, robustness to changes in the data-generating distribution  $P_0$  is desirable, whether they be from covariate shifts, changes in the underlying domain [3], or adversarial attacks [28, 38]. As deep networks become prevalent in modern performance-critical systems—prominent examples include perception systems for self-driving cars, and automated detection of

#### Abstract

We identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Although this problem has been widely studied empirically, much remains unknown concerning the theory underlying this trade-off. In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. Inspired by our theoretical analysis, we also design a new defense method, TRADES, made adversarial robustness off against accuracy. Our proposed algorithm performs well experimentally in datasets. The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision which we won the 1st place out of ~2,000 submissions, surpassing the runner-up approach by terms of mean  $\ell_2$  perturbation distance.

#### Introduction

sponse to the vulnerability of deep neural networks to small perturbations around input data [SZS<sup>+</sup>13], rial defenses have been an imperative object of study in machine learning [HPG<sup>+</sup>17], computer <sup>N<sup>+</sup>18, XWZ<sup>+</sup>17, MC17</sup>, natural language processing [JL17], and many other domains. In machine % of adversarial defenses has led to significant advances in understanding and defending against \*ar [HW<sup>+</sup>17]. In computer vision and natural language processing, adversarial defenses pensable building blocks for a range of security-critical systems and applications, such as ars and speech recognition authorization. The problem of adversarial defenses can be stated as g a classifier with high test accuracy on both natural and adversarial examples. The adversarial given labeled data  $(x, y)$  is a data point  $x'$  that causes a classifier  $c$  to output a different label



## Cross-validation Confidence Intervals for Test Error

Pierre Bayle\*  
Princeton University  
pbayle@princeton.edu

Alexandre Bayle\*  
Harvard University  
alexandre.bayle@g.harvard.edu

Lucas Janson  
Harvard University  
ljanson@fas.harvard.edu

Lester Mackey  
Microsoft Research New England  
lmackey@microsoft.com

### Abstract

This work develops central limit theorems for cross-validation and consistent estimators of its asymptotic variance under weak stability conditions on the learning algorithm. Together, these results provide practical, asymptotically-exact confidence intervals for  $k$ -fold test error and valid, powerful hypothesis tests of whether one learning algorithm has smaller  $k$ -fold test error than another. These results are also the first of their kind for the popular choice of leave-one-out cross-validation. In our real-data experiments with diverse learning algorithms, the resulting intervals and tests outperform the most popular alternative methods from the literature.

### 1 Introduction

Cross-validation (CV) [48, 25] is a de facto standard for estimating the test error of a prediction rule. By partitioning a dataset into  $k$  equal-sized validation sets, fitting a prediction rule with each validation set held out, evaluating each prediction rule on its corresponding held-out set, and averaging the  $k$  error estimates, CV produces an unbiased estimate of the test error with lower variance than a single train-validation split could provide. However, these properties alone are insufficient for high-stakes applications in which the uncertainty of an error estimate impacts decision-making. In predictive cancer prognosis and mortality prediction for instance, scientists and clinicians rely on *test error confidence intervals* based on CV and other repeated sample splitting estimators to avoid spurious findings and improve reproducibility [41, 44]. Unfortunately, the confidence intervals most often used have no correctness guarantees and can be severely misleading [29]. The difficulty comes from the dependence across the  $k$  averaged error estimates: if the estimates were independent, one could derive an asymptotically exact confidence interval for test error using a standard central limit theorem. However, the error estimates are seldom independent, due to the overlap amongst training sets and between different training and validation sets. Thus, new tools are needed to develop valid,

## A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation

Kamilar Rahnama Rad  Arian Maleki

First published: 20 June 2020 | <https://doi.org/10.1111/rssb.12374>

Read the full text >

 PDF  TOOLS  SHARE

### Summary

The paper considers the problem of out-of-sample risk estimation under the high dimensional settings where standard techniques such as  $k$ -fold cross-validation suffer from large biases. Motivated by the low bias of the leave-one-out cross-validation method, we propose a computationally efficient closed form approximate leave-one-out formula ALO for a large class of regularized estimators. Given the regularized estimate, calculating ALO requires a minor computational overhead. With minor assumptions about the data-generating process, we obtain a finite sample upper bound for the difference between leave-one-out cross-validation and approximate leave-one-out cross-validation,  $|\text{LO}-\text{ALO}|$ . Our theoretical analysis illustrates that  $|\text{LO}-\text{ALO}| \rightarrow 0$  with overwhelming probability, when  $n, p \rightarrow \infty$ , where the dimension  $p$  of the feature vectors may be comparable with or even greater than the number of observations,  $n$ . Despite the high dimensionality of the problem, our theoretical results do not require any sparsity assumption on the vector of regression coefficients. Our extensive numerical experiments show that  $|\text{LO}-\text{ALO}|$  decreases as  $n$  and  $p$  increase, revealing the excellent finite sample performance of approximate leave-one-out cross-validation. We further illustrate the usefulness of our proposed out-of-sample risk estimation method by an example of real recordings from spatially sensitive neurons (grid cells) in the medial entorhinal cortex of a rat.

Journal of the American Statistical Association  
Volume 113, 2018 - Issue 523

3,409 views  
25 CrossRef citations to date  
12 keywords

Theory and Methods  
**Distribution-Free Predictive Inference for Regression**  
Jing Lei, Maya Q'Sel, Alessandro Rinaldo, Ryan J. Tibshirani & Larry Wasserman  
Pages 1584-1611 | Received 21 Apr 2018, Accepted author version posted online 03 Apr 2019, Published online 08 Jun 2018

Download citation | <https://doi.org/10.1080/01621459.2017.1307116> | Check for updates

Full Article | Figures & data | References | Supplemental | Citations | Metrics | Reprints & Permissions

## ABSTRACT

We develop a general framework for distribution-free predictive inference in regression, using conformal inference. The proposed methodology allows for the construction of a prediction band for the response variable using any estimator of the regression function. The resulting prediction band preserves consistency properties of the original estimator under standard assumptions, while guaranteeing sample marginal coverage even when these assumptions do not hold. We analyze and compare empirically and theoretically, the two major variants of our conformal framework: full confidence and split conformal inference, along with a related jackknife method. These methods offer different tradeoffs between statistical accuracy (length of resulting prediction intervals) and computational efficiency. In particular, we develop a method for constructing valid in-sample prediction intervals. For split conformal inference, which has essentially the same computational efficiency as the full confidence method, we also describe an extension of our procedures for producing prediction intervals that adapt to heteroscedasticity in the data. Finally, we propose a model-free

## Conformal Prediction Under Covariate Shift

Ryan J. Tibshirani   Rina Foygel Barber   Emmanuel J. Candès   Aaditya Ramdas

### Abstract

We extend conformal prediction methodology beyond the case of exchangeable data. In particular, we show that a weighted version of conformal prediction can be used to compute distribution-free prediction intervals for problems in which the test and training covariate distributions differ, but the likelihood ratio between these two distributions is known—or, in practice, can be estimated accurately with access to a large set of unlabeled data (test covariate points). Our weighted extension of conformal prediction also applies more generally, to settings in which the data satisfies a certain weighted notion of exchangeability. We discuss other potential applications of our new conformal methodology, including latent variable and missing data problems.

## 1 Introduction

Let  $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ ,  $i = 1, \dots, n$  denote training data that is assumed to be i.i.d. from an arbitrary distribution  $P$ . Given a desired coverage rate  $1 - \alpha \in (0, 1)$ , consider the problem of constructing a band  $\hat{C}_n: \mathbb{R}^d \rightarrow \{\text{subsets of } \mathbb{R}\}$ , based on the training data such that, for a new i.i.d. point  $(X_{n+1}, Y_{n+1})$ ,

$$\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})\} \geq 1 - \alpha, \quad (1)$$

where this probability is taken over the  $n + 1$  points  $(X_i, Y_i)$ ,  $i = 1, \dots, n + 1$  (the  $n$  training points and the test point). Crucially, we will require (1) to hold with no assumptions whatsoever on the common distribution  $P$ .

Conformal prediction, a framework pioneered by Vladimir Vovk and colleagues in the 1990s, provides a means for achieving this goal, relying only on exchangeability of the training and test data. The definitive reference is the book

19v3 [stat.ME] 6 Jul 2020

## Gaussian Differential Privacy

Jinshuo Dong\* Aaron Roth† Weijie J. Su‡

May 2019; Revised April 2020

### Abstract

In the past decade, differential privacy has seen remarkable success as a rigorous and practical formalization of data privacy. This privacy definition and its divergence based relaxations, however, have several acknowledged weaknesses, either in handling composition of private algorithms or in analyzing important primitives like privacy amplification by subsampling. Inspired by the hypothesis testing formulation of privacy, this paper proposes a new relaxation of differential privacy, which we term “ $f$ -differential privacy” ( $f$ -DP). This notion of privacy has a number of appealing properties and, in particular, avoids difficulties associated with divergence based relaxations. First,  $f$ -DP faithfully preserves the hypothesis testing interpretation of differential privacy, thereby making the privacy guarantees easily interpretable. In addition,  $f$ -DP allows for lossless reasoning about composition in an algebraic fashion. Moreover, we provide a powerful technique to import existing results proven for the original differential privacy definition to  $f$ -DP and, as an application of this technique, obtain a simple and easy-to-interpret theorem of privacy amplification by subsampling for  $f$ -DP.

In addition to the above findings, we introduce a canonical single-parameter family of privacy notions within the  $f$ -DP class that is referred to as “Gaussian differential privacy” (GDP), defined based on hypothesis testing of two shifted Gaussian distributions. GDP is the focal privacy definition among the family of  $f$ -DP guarantees due to a central limit theorem for differential privacy that we prove. More precisely, the privacy guarantees of any hypothesis testing based definition of privacy (including the original differential privacy definition) converges to GDP in the limit under composition. We also prove a Berry–Esseen style version of the central limit theorem, which gives a computationally inexpensive tool for tractably analyzing the exact composition of private algorithms.

## Deep Learning with Gaussian Differential Privacy

Zhiqi Bu\* Jinshuo Dong† Qi Long‡ Weijie J. Su‡

University of Pennsylvania

November 25, 2019

### Abstract

Deep learning models are often trained on datasets that contain sensitive information such as individuals' shopping transactions, personal contacts, and medical records. An increasingly important line of work therefore has sought to train neural networks subject to privacy constraints that are specified by differential privacy or its divergence-based relaxations. These privacy definitions, however, have weaknesses in handling certain important primitives (composition and subsampling), thereby giving loose or complicated privacy analyses of training neural networks. In this paper, we consider a recently proposed privacy definition termed  $f$ -differential privacy [17] for a refined privacy analysis of training neural networks. Leveraging the appealing properties of  $f$ -differential privacy in handling composition and subsampling, this paper derives analytically tractable expressions for the privacy guarantees of both stochastic gradient descent and Adam used in training deep neural networks, without the need of developing sophisticated techniques as [3] did. Our results demonstrate that the  $f$ -differential privacy framework allows for a new privacy analysis that improves on the prior analysis [3], which in turn suggests tuning certain parameters of neural networks for a better prediction accuracy without violating the privacy budget. These theoretically derived improvements are confirmed by our experiments in a range of tasks in image classification, text classification, and recommender systems.

## 1 Introduction

In many applications of machine learning, the datasets contain sensitive information about individuals such as location, personal contacts, media consumption, and medical records. Exploiting the output of the machine learning algorithm, an adversary may be able to identify some individuals in the dataset, thus presenting serious privacy concerns. This reality gave rise to a broad and pressing call for developing privacy-preserving data analysis methodologies. Accordingly, there have been numerous investigations in the scholarly literature of many fields—statistics, cryptography, machine learning, and beyond—for the protection of privacy in data analysis.

# Statistical understanding on neural networks

## RESEARCH ARTICLE

### A mean field view of the landscape of two-layer neural networks

Song Mei<sup>1</sup>, Andrea Montanari<sup>1</sup>, and Phan-Minh Nguyen

PLoS August 14, 2018 | 11(8): e0260767 | doi:10.1371/journal.pone.0260767

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved June 21, 2018 (received for review April 18, 2018)

Article Figures & SI Info & Metrics PDF

#### Significance

Multi-layer neural networks have proven extremely successful in image classification to robotics. However, the reasons for their precise domain of applicability are unknown. Learning is solving a complex optimization problem with millions of stochastic gradient descent (SGD) algorithms. We study and derive a compact description of the differential equation. Among other things, we show that the landscape becomes more complex when the number of layers increases.

8.05355v3 [math.ST] 22 Oct 2019

#### The generalization error of random features regression: Precise asymptotics and double descent curve

Song Mei<sup>1</sup> and Andrea Montanari<sup>1</sup>

October 23, 2019

#### Abstract

Deep learning methods operate in regimes that defy the traditional statistical student. Neural network architectures often contain more parameters than training samples, and so we risk that they interpolate the observed labels, even if the labels are replaced by pure noise. Despite their large number of parameters, these models exhibit a surprising generalization error on real data.

This phenomenon has been understood in terms of a so-called “double descent” curve. As the model complexity increases, the test error follows the usual U-shaped curve of the hypothesis, but decreases and then peaks around the interpolation threshold (before the model achieves vanishing training error). However, it decreases again as model complexity exceeds this threshold. The global minimum of test error is found above the interpolation threshold, often in the extreme overparameterization regime where the number of parameters is much larger than the number of samples. The facts being a property of deep neural networks, elements of this behavior have been demonstrated in weak string settings, including linear regression with random covariates.

In this paper we consider the problem of learning an unknown function over the  $d$ -dimensional sphere  $S^{d-1}$ , from  $n$  i.i.d. samples  $(x_i, y_i) \in S^{d-1} \times \mathbb{R}$ ,  $i \leq n$ . We perform ridge regression on  $N$  random features of the form  $\sigma(x_i \cdot w_j)$ ,  $w_j \in S^{d-1}$ . This can be equivalently described as a two-layer neural network with random first-layer features. We compute the precise asymptotics of the test error, in the limit  $N, n, d \rightarrow \infty$  with  $N/n$  and  $n/d$  fixed. This provides the first analytically tractable model that captures all the features of the double-descent phenomenon without assuming any low-complexity structure. In particular, above a critical value of the signal-to-noise ratio, minimum test error is achieved by extremely overparameterized interpolation, i.e., networks that have a number of parameters much larger than the sample size, and vanishing training error.

.MLJ 19 Feb 2020

### Implicit Regularization of Random Feature Models

Arthur Jacot<sup>1\*</sup>, Berfin Şimşek<sup>1,2</sup>, Francesco Spadaro<sup>1</sup>, Clément Hongler<sup>1</sup>, Franck Gabriel<sup>1</sup>

#### Abstract

Random Feature (RF) models are used as efficient parametric approximations of kernel methods. We investigate, by means of random matrix theory, the connection between Gaussian RF models and Kernel Ridge Regression (KRR). For a Gaussian RF model with  $P$  features,  $N$  data points, and a ridge  $\lambda$ , we show that the average (i.e. expected) RF predictor is close to a KRR predictor with an effective ridge  $\tilde{\lambda}$ . We show that  $\tilde{\lambda} > \lambda$  and  $\tilde{\lambda} \searrow \lambda$  monotonically as  $P$  grows, thus revealing the implicit regularization effect of finite RF sampling. We then compare the risk (i.e. test error) of the  $\lambda$ -KRR predictor with the average risk of the  $\lambda$ -RF predictor and obtain a precise and explicit bound

due to the random initialization of the parameters and to the stochasticity of the training algorithm; for random forests, to the random branching; for random feature models to the sampling of random features. The somehow surprising generalization behavior of these models has been the subject of increasing attention. In general, the (i.e. test error) is a random variable with two sources of randomness: the usual one due to the sampling of the training set, and the second one due to the randomness of the model itself.

We consider the Random Feature (RF) model (Rahin Rezaei, 2008) with features sampled from a Gaussian Process (GP) and study the RF predictor  $f$  minimizing the regularized least squares error, isolating the randomness of model by considering fixed training data points. RF is

## A Learning Algorithm for Boltzmann Machines\*

DAVID H. ACKLEY  
GEOFFREY E. HINTON  
Computer Science Department  
Carnegie-Mellon University  
TERRENCE J. SEJNOWSKI  
Biophysics Department  
The Johns Hopkins University

The computational power of massively parallel networks of simple processing elements resides in the communication bandwidth provided by the hardware connections between elements. These connections can allow a significant fraction of the knowledge of the system to be applied to an instance of a problem in a very short time. One kind of computation for which massively parallel networks appear to be well suited is large constraint  $s$  but to use the connections efficiently two conditions  $r$  search technique that is suitable for parallel networks in there must be some way of choosing internal representa presenting hardware connections to be used efficiently straints in the domain being searched. We describe a ge method, based on statistical mechanics, and we show a eral learning rule for modifying the connection strength knowledge about a task domain in an efficient way. We examples in which the learning algorithm creates one that are demonstrably the most efficient way of using th. rity structure.

### 1. INTRODUCTION

Evidence about the architecture of the brain and its VLSI technology have led to a resurgence of interest

\* The research reported here was supported by grants from Foundation. We thank Peter Brown, Francis Crick, Mark Der Friedman, Stuart Geman, Gail Gong, John Hopfield, Jay McEl-Harry Prietz, Dave Rumelhart, Tim Shallice, Paul Smolensky, F-man Venkatasubramanian for helpful discussions.

Reprint requests should be addressed to David Ackley, Co-Carnegie-Mellon University, Pittsburgh, PA 15213.

Cope Copied (2016) 8:1384-1373  
DOI: 10.1016/j.cog.2016.06.029

### Weight Uncertainty in Boltzmann Machine

Jian Zhang<sup>1,2</sup>, Shidi Ding<sup>1,2</sup>, Nan Zhang<sup>1,2</sup>, Yu Xue<sup>1</sup>

Received: 14 November 2015 / Accepted: 18 August 2016 / Published online: 11 August 2016  
© Springer Science+Business Media New York 2016

**Abstract** Based on restricted Boltzmann machine (RBM), the deep learning models can be roughly divided into deep belief networks (DBNs) and deep Boltzmann machine (DBM). However, the overfitting problems commonly exist in neural networks and RBM models. In order to alleviate the overfitting problem, less of research has been done. This paper addressed the overfitting problem in RBM and proposed the weight uncertainty semi-restricted Boltzmann machine (WSRBM) to improve the ability of image recognition and image reconstruction. **Methods** First, this paper built weight uncertainty RBM model based on maximum likelihood estimation. And in the experimental section, this paper verified the effectiveness of the weight uncertainty deep belief network and the weight uncertainty deep Boltzmann machine. Second, in order to obtain better reconstructed images, this paper used the semi-restricted Boltzmann machine (SRBM) as the feature extractor and built the WSRBM. Lastly, this paper

introduced DBM were effective compared with the deep-out method. And the WSRBM model performed well in image recognition and image reconstruction as well. **Conclusions** This paper introduced the weight uncertainty method to RBM, and proposed a WSRBM model, which was effective in image recognition and image reconstruction.

**Keywords** RBM · DBM · DBN · Weight uncertainty

### Introduction

In the viewpoint of supervised learning, deep neural networks can be regarded as multilayer perceptrons. The problem that the network converged in the error curved surface depended on the initialized weights. However, the error curved surface of the multilayer perceptron is complex. And the network may converge to different local

## An Infinite Restricted Boltzmann Machine

Marc-Alexandre Côté  
marc-alexandre.cote@usherbrooke.ca  
Hugo Larochelle  
hugo.larochelle@usherbrooke.ca  
Department of Computer Science, Université de Sherbrooke,  
Sherbrooke, QC J1K 2R1, Canada

We present a mathematical construction for the restricted Boltzmann machine (RBM) that does not require specifying the number of hidden units. In fact, the hidden layer size is adaptive and can grow during training. This is obtained by first extending the RBM to be sensitive to the ordering of the chosen definition of the energy. Finally, we show that naturally chosen definition of the energy infinitely many hidden units is well maximum likelihood training can be naturally and adaptively added. We empirically study the behavior of performance is competitive to that tuning of a hidden layer size.





## Belief networks, hidden Markov models, and Markov random fields: A unifying view

Padhraic Smyth<sup>1,1</sup>

Show more ▾

[https://doi.org/10.1016/S0167-8655\(97\)91050-7](https://doi.org/10.1016/S0167-8655(97)91050-7)

[Get rights and content](#)

### Abstract

The use of graphs to represent independence structure in multivariate probability models has been pursued in a relatively independent fashion across a wide variety of research disciplines since the beginning of this century. This paper provides a brief overview of the current status of such research with particular attention to recent developments which have served to unify such seemingly disparate topics as probabilistic expert systems, statistical physics, image analysis, genetics, decoding of error-correcting codes, Kalman filters, and speech recognition with Markov models.

[◀ Previous article in issue](#)

[Next article in issue ▶](#)

## Fusion, Propagation, and Structuring in Belief Networks\*

Judea Pearl

Cognitive Systems Laboratory, Computer Science Department,  
University of California, Los Angeles, CA 90024, U.S.A.

Recommended by Patrick Hayes

### ABSTRACT

*Belief networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify direct dependencies between the linked propositions, and the strengths of these dependencies are quantified by conditional probabilities. A network of this sort can be used to represent the generic knowledge of a domain expert, and it turns into a computational architecture if the links are used not merely for storing factual knowledge but also for directing and activating the data flow in the computations which manipulate this knowledge.*

*The first part of the paper deals with the task of fusing and propagating the impacts of new information through the networks in such a way that, when equilibrium is reached, each proposition will be assigned a measure of belief consistent with the axioms of probability theory. It is shown that if the network is singly connected (e.g. tree-structured), then probabilities can be updated by local propagation in an isomorphic network of parallel and autonomous processors and that the impact of new information can be imparted to all propositions in time proportional to the longest path in the network.*

*The second part of the paper deals with the problem of finding a tree-structured representation for a collection of probabilistically coupled propositions using auxiliary (dummy) variables, colloquially called "hidden causes." It is shown that if such a tree-structured representation exists, then it is possible to uniquely uncover the topology of the tree by observing pairwise dependencies among the available propositions (i.e., the leaves of the tree). The entire tree structure, including the strengths of all internal relationships, can be reconstructed in time proportional to  $n \log n$ , where  $n$  is the number of leaves.*

# To-Do

- Pick up your topics
- Pick up/select your paper to be presented
- Make our presenting schedule
- “Scribed notes” or slides?
- Your Comments?