Probabilistic Machine Learning: Models, Algorithms and a Programming Library

Jun Zhu
Department of Computer Science and Technology, Tsinghua Lab of Brain and Intelligence
State Key Lab for Intell. Tech & Sys., BNRist Lab, Tsinghua University, 100084, China
dcszj@tsinghua.edu.cn

Abstract

Probabilistic machine learning provides a suite of powerful tools for modeling uncertainty, performing probabilistic inference, and making predictions or decisions in uncertain environments. In this paper, we present an overview of our recent work on probabilistic machine learning, including the theory of regularized Bayesian inference, Bayesian deep learning, scalable inference algorithms, a probabilistic programming library named ZhuSuan, and applications in representation learning as well as learning from crowds.

1 Introduction

The world is an uncertain place because of physical randomness, incomplete knowledge, ambiguities, and contradictions. Drawing inference from noisy or ambiguous data is an important part of intelligent systems, where probability theory (in particular Bayes' theorem) serves as a principled framework of combining prior knowledge and empirical evidence. The past 30 years have seen tremendous progress in developing both Bayesian and nonparametric Bayesian methods for resolving model complexity and adapting to stochastic and changing environments with data-driven learning algorithms.

However, conventional probabilistic inference is facing great challenges in dealing with large-scale complex data, arising from unstructured, noisy, and dynamic environments such as the Web, which records massive digital traces of human activities. To address these challenges, our research consists of developing flexible Bayesian inference methods and scalable algorithms to solve important problems in scientific and engineering domains. This paper presents an overview of our recent progress on probabilistic machine learning.

2 Bayesian Inference with Posterior Regularization

At the core of Bayesian inference is the Bayes' rule (a.k.a Bayes' theorem). Though intuitively simple, the standard Bayesian inference with Bayes' rule is lacking of a mechanism to directly control the target posterior distribution, as the inference process is a "one-way" procedure that projects the prior distribution to the posterior by observing empirical data. In many settings, such as supervised learning and reinforcement learning, our ultimate goal is to apply the posterior to learning tasks with some measurement on the performance (e.g., prediction error or expected reward). A good posterior distribution should have a low prediction error or a high expected reward. Furthermore, as the large-scale knowledge bases are built [Suchanek *et al.*, 2007; Carlson *et al.*, 2010] and crowdsourcing platforms [Raykar *et al.*, 2010] are widely adopted to collect human data, it is desirable to incorporate the external side information into statistical modeling and inference when building an intelligent system.

We generalized the flexibility of Bayesian methods by presenting a new framework for Bayesian inference with posterior regularization, named regularized Bayesian inference (RegBayes) [Zhu *et al.*, 2014b], which can capture the desired properties or side information into the inference process by solving an optimization problem. RegBayes introduces a new dimension (i.e., posterior regularization) to the standard Bayesian inference, and makes it significantly more flexible in optimizing the objectives of learning/decision-making tasks or incorporating domain knowledge. When the posterior regularization is derived from the discriminative max-margin principle, RegBayes sets up a bridge between (nonparametric) Bayesian methods and max-margin learning, two important subfields in machine learning that have taken largely disjoint paths for decades. As illustrating examples, we have applied RegBayes to learn discriminative latent representations from text [Zhu *et al.*, 2014a], image [Li *et al.*, 2015], and network [Zhu, 2012] data, where the posterior regularization term measures the prediction error. We have also applied RegBayes to perform robust Bayesian inference with domain knowledge that is represented in FOL (first-order logic) form and collected from crowds [Mei *et al.*, 2014], and demonstrated that the idea can be naturally generalized to unsupervised learning and semi-supervised learning settings [Chen *et al.*, 2014]. Finally, we extended the RegBayes theory to the reproducing kernel Hilbert space [Song *et al.*, 2016], which avoids parametric assumptions of prior distribution and likelihood function.

3 Scalable Inference Algorithms

It is generally intractable to infer the posterior distribution of a non-trivial Bayesian model using either the vanilla Bayes' rule or the more flexible RegBayes principle, especially when we consider the applications with streaming and/or massive data as well as the deep Bayesian models that have a hierarchical structure of latent variables. The intractability also poses challenges for learning the parameters of a probabilistic model that has latent variables. To address the computational challenges, we have developed scalable algorithms in both online and distributed settings [Zhu *et al.*, 2017], and further developed a probabilistic programming library to enable the fast development and application of deep Bayesian models.

For the online setting, the conventional Bayes' rule is essentially sequential, that is, the posterior at time t is actually playing the role of a prior for the data at time $t + 1$ for the Bayesian updating. We generalized this sequential property to the more flexible RegBayes for sequential learning. Specifically, we have presented a form of streaming RegBayes inference — online Bayesian passive-aggressive learning (BayesPA) [Shi and Zhu, 2017], where a Bayesian classifier is sequentially updated by considering two cases. If the incoming sample is correctly classified by the current posterior, a passive update strategy is adopted (i.e., no update); otherwise, an aggressive update strategy is adopted which projects the current posterior to the one that makes an accurate prediction on the given sample. This passive-aggressive strategy was applied to learn supervised topic models, with about two-orders of magnitude speedup against the batch counterpart algorithms, where the intractable posterior distribution of latent topics was inferred by variational approximation methods.

We have also developed stochastic gradient algorithms for posterior inference with large-scale datasets. As the intriguing results of [Bottou and Bousquet, 2008] suggest, an algorithm as simple as stochastic gradient descent (SGD) can be optimally efficient in terms of "number of bits learned per unit of computation". We considered both stochastic variational and stochastic Monte Carlo methods for probabilistic inference, especially by exploring the structures of the models (e.g., manifolds or graphical structures). For instance, we developed the stochastic gradient geodesic Monte Carlo (SGGMC) [Liu *et al.*, 2016] on manifolds with known geodesic flow, and we developed Riemannian Stein Variational Gradient Descent (RSVGD) [Liu and Zhu, 2018], a Bayesian inference method that generalizes Stein Variational Gradient Descent (SVGD) to Riemann manifold. RSVGD has the advantage over SVGD of utilizing information geometry and brings the unique advantages of SVGD to the Riemannian world. Finally, we have generalized Stein variational inference to graphical models [Zhuo *et al.*, 2018], where a message passing protocol was developed to explore the graphical structure for efficient inference without degenerating the particle efficiency.

For the distributed setting, we consider both general Bayesian inference algorithms (e.g., MCMC) and the highly optimized algorithms for particular models (e.g., topic models). For generic algorithms, we presented a distributed posterior sampling algorithm with a moment sharing scheme under EP (i.e., expectation propagation) to achieve high accuracy [Xu *et al.*, 2014]. For the concrete models, we have studied extensively the scalable algorithms to learn topic models, which provide a suite of statistical tools to discover latent semantic structures from complex corpora, with latent Dirichlet allocation (LDA) [Blei *et al.*, 2003] as the most popular one. We have finished a series of work on scaling up the vanilla LDA [Chen *et al.*, 2016] as well as various representative extensions, including correlated topic models [Chen *et al.*, 2013], dynamic topic models [Bhadury *et al.*, 2016], and max-margin topic models [Zhu *et al.*, 2013]. In particular, we developed WarpLDA [Chen *et al.*, 2016], an LDA sampler which achieves both the best $O(1)$ time complexity per token and the best $O(K)$ scope of random access, where K is the number of topics. Our empirical results in a wide range of testing conditions demonstrate that WarpLDA is consistently 5-15x faster than the state-of-the-art competitors. With WarpLDA, users can learn up to one million topics from hundreds of millions of documents in a few hours, at an unprecedentedly throughput of 11G tokens per second. We further developed SaberLDA, a GPU-based LDA system that implements a sparsity-aware algorithm to achieve sublinear time complexity and scales well to learn a large number of topics [Li *et al.*, 2017c]. SaberLDA can learn from billions-token-scale data with up to 10,000 topics, which is almost two orders of magnitude larger than that of the previous GPU-based systems.

4 Bayesian Deep Learning and a Probabilistic Programming Library

Recently, the popularity of deep generative models have demonstrated the promise of combining deep neural networks with probabilistic modeling, which has shown superior results in image generation [Kingma and Welling, 2013; Goodfellow *et al.*, 2014], semi-supervised classification [Salimans *et al.*, 2016] and one-shot learning [Rezende *et al.*, 2016]. We call such an arising direction that conjoins the advantages of Bayesian methods and deep learning as *Bayesian Deep Learning* (BDL). The scope of BDL covers the traditional Bayesian methods, the deep learning methods where probabilistic inference plays a key role, and their intersection. One unique feature of BDL is that the deterministic transformation between random variables can be automatically learned from data under an expressive parametric formulation typically using deep neural networks, while in traditional Bayesian models, the transformation tends to have a simple analytical form (e.g., the exponential function or inner product). One key challenge for Bayesian deep learning is on posterior inference, which is typically intractable for such models and needs sophisticated approximation techniques.

We aim to address the newly arising challenges in Bayesian deep learning by developing efficient algorithms and a programming library. Namely, we have developed kernel implicit variational inference [Shi *et al.*, 2018a], which is a variational inference algorithm by using an implicit distribution (i.e., without a tractable density) as the variational pos-

terior. Our method addresses the issues of noisy estimation and computational infeasibility when applied to models with high-dimensional latent variables. We further developed a gradient estimator for implicit distributions based on Stein's identity and a spectral decomposition of kernel operators, where the eigenfunctions are approximated by the Nyström method. Unlike the previous works that only provide estimates at the sample points, our approach directly estimates the gradient function, thus allows for a simple and principled out-of-sample extension. We provide theoretical results on the error bound of the estimator and discuss the bias-variance tradeoff in practice [Shi *et al.*, 2018b].

Finally, we designed ZhuSuan[1] [Shi *et al.*, 2017], a python probabilistic programming library for Bayesian deep learning. We built ZhuSuan upon the popular deep learning library Tensorflow [Abadi *et al.*, 2016]. Unlike existing deep learning libraries, which are mainly designed for deterministic neural networks and supervised learning tasks, ZhuSuan is featured for its deep root into Bayesian inference, thus supporting various kinds of probabilistic models, including both the traditional hierarchical Bayesian models and recent deep generative models. ZhuSuan incorporates the recent advances on scalable inference/learning algorithms, including both variational and MCMC methods. We have developed many examples to illustrate the probabilistic programming on ZhuSuan, including Bayesian logistic regression, variational auto-encoders, deep sigmoid belief networks and Bayesian recurrent neural networks.

5 Applications

5.1 Learning Interpretable and Predictive Representations

Due to the complex nature of the environment, the collected data often presents various unfavorable properties, such as being noisy, incomplete, and dynamics. It is typically unwise to directly throw the raw data into a learning machine for further knowledge discovery or data management tasks. Extracting a proper representation from the raw data is playing a vital role to make a learning algorithm work well. Researchers have taken decades to manually design such good features for various types of data, including text, image, video, voice and networks. Recently, significant progress has been made on learning representations, especially when a hierarchical model structure is adopted which can lead to a compact and rich form to represent complex data [LeCun *et al.*, 2015]. In our research, we have studied extensively on various important aspects in representation learning, including discriminativeness, sparsity, model complexity as well as scalability.

For discriminativeness, we focus on developing flexible statistical models that can consider supervising information to guide the learning of feature representations. This is motivated by the increasing availability of free on-line information such as image tags, user ratings, etc. Various forms of side-information that can potentially offer "free" supervision have led to a need for new models and training schemes that can make effective use of such information to achieve better results, such as more discriminative latent representations of image contents, and more accurate image classifiers. Under the similar principle of RegBayes, we have developed various statistical models that learn discriminative representations for text, image, and network data. For example, we presented a max-margin supervised topic model (MedLDA) [Zhu *et al.*, 2012], which conjoins the principe of max-margin learning with Bayesian latent variable models. MedLDA can significantly improve the discriminativeness and interpretability of the learnt topics, as well as making accurate predictions in classification and regression tasks. The similar idea has been applied to learn latent representations shared by multimodal data [Chen *et al.*, 2012], demonstrating significant improvements on cue integration and prediction, as well as learning the social space representations for network data [Chen *et al.*, 2015]. When a hierarchy is adopted to represent the latent features, we derived a discriminative training algorithm for deep generative models (DGMs), which for the first time manages to lift DGMs to the same level of prediction accuracy with deep neural networks [Li *et al.*, 2015], while retaining (or even improving) the ability of generating out-set samples and completing missing values. We further demonstrated that a discriminative DGM [Li *et al.*, 2017b] can achieve state-of-the-art accuracy for semi-supervised learning, where only a small part of the labels are provided. For the same task of semi-supervised classification, we also carefully investigated the generative adversarial networks (GAN) and presented effective training schemes, including an adversarial game with three players (i.e., Triple-GAN) [Li *et al.*, 2017a] and structured GAN [Deng *et al.*, 2017]. Finally, we have also proposed novel spectral decomposition algorithms on learning supervised latent topic models [Wang and Zhu, 2014b; Ren *et al.*, 2017], which are provably correct and computationally efficient.

Sparsity is another desired property in latent representation learning. For example, very often it makes intuitive sense to assume that each document or word has a few salient topical meanings or senses, rather than letting every topic make a non-zero contribution; this is important in practice for large scale text mining endeavors such as those undertaken in Google or Yahoo, where it is not uncommon to learn hundreds or thousands of topics for hundreds of millions of documents; without an explicit sparcification procedure, it would be extremely challenging, if not impossible, to nail down the semantic meanings of a document or word. However, to achieve sparsity in a probabilistic topic model is non-trivial. Existing attempts by using a sparse prior could indirectly introduce a sparsity bias over the posterior representations. We addressed this problem by presenting sparse topical coding (STC) [Zhu and Xing, 2011], a novel non-probabilistic formulation of topic models for learning hierarchical latent representations of input samples (e.g., text documents). In STC, each individual input feature (e.g., a word count) is reconstructed from a linear combination of a set of bases, where the coefficient vectors (or codes) are unnormalized, and the representation of an entire document is derived via an aggregation strategy (e.g., truncated averaging) from the codes of all its individual features. When applied to text, we use the log-Poisson loss to model discrete word counts and learn the

[1] GitHub repository: https://github.com/thu-ml/zhusuan

topical bases that are unigram distributions over the terms in a vocabulary. The nonprobabilistic STC enjoys nice properties which make it an appealing alternative formulation of topic models. We also presented an online learning method to deal with large-scale datasets [Zhang *et al.*, 2013].

Finally, we addressed the model complexity issue of latent representation learning, that is, how to decide the dimensionality of the latent space, which is unknown a *priori*. Traditional methods usually resort to post-processing procedures such as cross-validation or likelihood ratio test. The recent success of nonparametric Bayesian techniques, such as the Dirichlet process (DP) mixtures in dealing with similar challenges in clustering (i.e., unknown number of clusters), offers a promising direction to bypass the model selection problem and automatically resolve the unknown number of experts. We followed this line of research by presenting novel nonparametric Bayesian models. Unlike the conventional Bayesian inference, we built our models under the RegBayes framework, which enables the adoption of the max-margin principle for learning discriminative representations. In particular, we have presented the infinite SVMs (iSVMs) [Zhu *et al.*, 2011a], a Dirichlet process mixture of large-margin kernel machines, and infinite latent SVMs (iLSVMs) [Zhu *et al.*, 2014b; 2011b], an SVM classifier with an unbounded number of latent features. Such techniques can be naturally generalized to learn latent representations with an unbounded dimensionality for relational data, such as social networks [Zhu, 2012] and the "user-movie" dyad data in recommendation systems [Xu *et al.*, 2013]. We further investigated the scalability issue by presenting more efficient algorithms under a small-variance asymptotic analysis. For instance, we performed the small-variance version of iSVMs and got a simple and efficient algorithm which monotonically optimizes a max-margin DP-means [Wang and Zhu, 2014a] problem, an extension of DP-means for both predictive learning and descriptive clustering. We also used small-variance analysis to derive DP-space [Wang and Zhu, 2015], a simple and efficient subspace learning method that can automatically decide the number of subspace components and the dimensionality of each component.

5.2 Learning from Crowds

Crowdsourcing provides an effective way to collect large-scale experimental data from distributed workers with a low cost. However, the labeling accuracy of web workers is often lower than expected due to their various backgrounds or lack of knowledge. To improve the accuracy, a typical strategy is to label every data (or task) multiple times by different workers, then the redundant labels can provide hints on resolving the true labels. To extract useful information from the cheap but potentially unreliable answers to tasks, we need sophisticated methods to identify reliable workers as well as unambiguous tasks. We consider both *supervised learning* (SL) and *unsupervised learning* (USL) settings. The SL setting aims to resolve the assumed but unknown ground truth label for each data sample, while the USL setting aims to understand the behaviors of workers and tasks, without a true label assumption. One common thread of our approaches in both settings is that we build sophisticated statistical models.

For the SL setting, existing work includes both generative approaches and discriminative approaches. A generative method builds a flexible probabilistic model for generating the noisy observations conditioned on the unknown true labels and some behavior assumptions, with examples of the Dawid-Skene estimator [Dawid and Skene, 1979] and the minimax entropy estimator [Zhou *et al.*, 2012]. In contrast, a discriminative approach does not model the observations; it directly identifies the true labels via some aggregation rules. Examples include majority voting and the weighted majority voting that takes worker reliability into consideration [Karger *et al.*, 2011]. We approach this problem from a very different perspective. Instead of building on the popular methods, we re-took the most classical rule of majority voting and present a significant extension by adopting the max-margin principle [Tian and Zhu, 2015]. Under the RegBayes theory, we were further able to present a Bayesian generalization that conjoins the advantages of both generative and discriminative approaches. The max-margin majority voting (M^3V) directly maximizes the margin between the aggregated score of a potential true label and that of any alternative label, and the Bayesian model consists of a flexible probabilistic model to generate the noisy observations by conditioning on the unknown true labels. We adopted the same approach as the classical Dawid-Skene estimator to build the probabilistic model by considering worker confusion matrices, though many other generative models are also possible. Then, we strongly coupled the generative model and M^3V by formulating a joint learning problem under the RegBayes framework, where the posterior regularization enforces a large margin between the potential true label and any alternative label. Naturally, our Bayesian model covers both the David-Skene estimator and M^3V as special cases by setting the regularization parameter to its extreme values (i.e., 0 or ∞). Experiments on real datasets suggest that max-margin learning can significantly improve the accuracy of majority voting, and the Bayesian estimators achieve better results than state-of-the-art estimators. With the same goal of learning accurate classifiers, we further considered the extremely challenging setting of zero-shot learning by leveraging crowdsourcing to learn human comprehensible and cross-category transferrable attributes. This is achieved by carefully designing the mechanism and crowdsourcing jobs as well as a hierarchical Bayesian model to aggregate the noisy labels [Tian *et al.*, 2017].

For the USL setting, we analyze the phenomenon of *schools of thought*, where each task may have multiple valid answers. This phenomenon is commonly observed in applying crowdsourcing for qualitative user studies, demographic survey or solving a hard problem, because of the subjectiveness of the tasks or the variety of workers' cultural and educational background. We built a nonparametric Bayesian model to identify worker reliability and task clarity without the assumption of ground truth labels [Tian and Zhu, 2012]. Our model was built on two mild assumptions on the grouping behavior that happens in schools of thought: 1) reliable workers tend to agree with other workers in many tasks; and 2) the answers to a clear task tend to form tight clusters. Finally, we have developed a scalable system using the above tech-

niques to detect crowd frauds in internet advertising [Tian *et al.*, 2015], which has been deployed in a giant internet company in China.

6 Conclusions

The above topics give some of the overall flavor of our research, including the fundamental theories of machine learning, the design of scalable algorithms and a probabilistic programming library, and the applications in various domains. We keep improving probabilistic machine learning. We are also working on some exciting but largely unaddressed problems, such as adversarial attack and defense of deep learning models [Dong *et al.*, 2018; Pang *et al.*, 2018] or machine learning models in general.

Acknowledgements

The work is supported by the National NSF of China (Nos. 61620106010 and 61621136008) and Beijing Natural Science Foundation (No. L172037).

References

[Abadi *et al.*, 2016] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[Bhadury *et al.*, 2016] Arnab Bhadury, Jianfei Chen, Jun Zhu, and Shixia Liu. Scaling up dynamic topic models. In *WWW*, 2016.

[Blei *et al.*, 2003] David Blei, Andrew Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *JMLR*, (3):993–1022, 2003.

[Bottou and Bousquet, 2008] Leon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *NIPS*, 2008.

[Carlson *et al.*, 2010] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[Chen *et al.*, 2012] Ning Chen, Jun Zhu, Fuchun Sun, and Eric P. Xing. Large-margin predictive latent subspace learning for multi-view data analysis. *IEEE Trans. on PAMI*, 34(12):2365–2378, 2012.

[Chen *et al.*, 2013] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. Scalable inference for logistic-normal topic models. In *NIPS*, 2013.

[Chen *et al.*, 2014] Chaoyou Chen, Jun Zhu, and Xinhua Zhang. Robust bayesian max-margin clustering. In *NIPS*, 2014.

[Chen *et al.*, 2015] Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. Discriminative relational topic models. *IEEE Trans. on PAMI*, 37(5):973–986, 2015.

[Chen *et al.*, 2016] Jianfei Chen, Kaiwei Li, Jun Zhu, and Wenguang Chen. WarpLDA: a Cache Efficient O(1) Algorithm for Latent Dirichlet Allocation. In *VLDB*, 2016.

[Dawid and Skene, 1979] Alexander P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *J. Roy. Statist. Soc. C (Applied Statistics)*, 2:20–28, 1979.

[Deng *et al.*, 2017] Zhijie Deng, Hao Zhang, Xiaodan Liang, Luona Yang, Shizhen Xu, Jun Zhu, and Eric Xing. Structured generative adversarial networks. In *NIPS*, 2017.

[Dong *et al.*, 2018] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[Karger *et al.*, 2011] David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *NIPS*, 2011.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, (521):436–444, 2015.

[Li *et al.*, 2015] Chongxuan Li, Jun Zhu, Tianlin Shi, and Bo Zhang. Max-margin deep generative models. In *NIPS*, 2015.

[Li *et al.*, 2017a] Chongxuan Li, Kun Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *NIPS*, 2017.

[Li *et al.*, 2017b] Chongxuan Li, Jun Zhu, and Bo Zhang. Max-margin deep generative models for (semi-)supervised learning. *IEEE TPAMI (in press)*, 2017.

[Li *et al.*, 2017c] Kaiwei Li, Jianfei Chen, Wenguang Chen, and Jun Zhu. SaberLDA: Sparsity-aware learning of topic models on gpus. In *ASPLOS*, 2017.

[Liu and Zhu, 2018] Chang Liu and Jun Zhu. Riemannian stein variational gradient descent for bayesian inference. In *AAAI*, 2018.

[Liu *et al.*, 2016] Chang Liu, Jun Zhu, and Yang Song. Stochastic gradient geodesic mcmc methods. In *NIPS*, 2016.

[Mei *et al.*, 2014] Shike Mei, Jun Zhu, and Xiaojin Zhu. Robust RegBayes: Selectively incorporating first-order logic domain knowledge into Bayesian models. In *ICML*, 2014.

[Pang *et al.*, 2018] Tianyu Pang, Chao Du, and Jun Zhu. Max-mahalanobis linear discriminant analysis networks. In *ICML*, 2018.

[Raykar *et al.*, 2010] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo H. Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *JMLR*, 11:1297–1322, 2010.

[Ren *et al.*, 2017] Yong Ren, Yining Wang, and Jun Zhu. Spectral learning for supervised topic models. *IEEE TPAMI (in press)*, 2017.

[Rezende *et al.*, 2016] Danilo Jimenez Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-shot generalization in deep generative models. *arXiv preprint arXiv:1603.05106*, 2016.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2226–2234, 2016.

[Shi and Zhu, 2017] Tianlin Shi and Jun Zhu. Online Bayesian passive-aggressive learning. *JMLR*, 18(33):1–39, 2017.

[Shi *et al.*, 2017] Jiaxin Shi, Jianfei Chen, Jun Zhu, Shengyang Sun, Yucen Luo, Yihong Gu, and Yuhao Zhou. ZhuSuan: A library for bayesian deep learning. *preprint arXiv:1709.05870*, 2017.

[Shi *et al.*, 2018a] Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. In *ICLR*, 2018.

[Shi *et al.*, 2018b] Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *ICML*, 2018.

[Song *et al.*, 2016] Yang Song, Jun Zhu, and Yong Ren. Kernel Bayesian inference with posterior regularization. *NIPS*, 2016.

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO – a core of semantic knowledge. In *WWW*, 2007.

[Tian and Zhu, 2012] Yuandong Tian and Jun Zhu. Learning from crowds in the presence of schools of thought. In *SIGKDD*, 2012.

[Tian and Zhu, 2015] Tian Tian and Jun Zhu. Max-margin majority voting for learning from crowds. In *NIPS*, 2015.

[Tian *et al.*, 2015] Tian Tian, Jun Zhu, Fen Xia, Xin Zhuang, and Tong Zhang. Crowd fraud detection in internet advertising. In *WWW*, 2015.

[Tian *et al.*, 2017] Tian Tian, Ning Chen, and Jun Zhu. Learning attributes from the crowdsourced relative labels. In *AAAI*, 2017.

[Wang and Zhu, 2014a] Yining Wang and Jun Zhu. Small-variance asymptotics for dirichlet process mixtures of svms. In *AAAI*, 2014.

[Wang and Zhu, 2014b] Yining Wang and Jun Zhu. Spectral methods for supervised topic models. In *NIPS*, 2014.

[Wang and Zhu, 2015] Yining Wang and Jun Zhu. Dp-space: Bayesian nonparametric subspace clustering with small-variance asymptotics. In *ICML*, 2015.

[Xu *et al.*, 2013] Minjie Xu, Jun Zhu, and Bo Zhang. Fast max-margin matrix factorization with data augmentation. In *International Conference on Machine Learning (ICML)*, pages 978–986, 2013.

[Xu *et al.*, 2014] Minjie Xu, Balaji Lakshminarayanan, Yee Whye Teh, Jun Zhu, and Bo Zhang. Distributed Bayesian posterior sampling via moment sharing. In *NIPS*, 2014.

[Zhang *et al.*, 2013] Aonan Zhang, Jun Zhu, and Bo Zhang. Sparse online topic models. In *WWW*, 2013.

[Zhou *et al.*, 2012] Dengyong Zhou, Sumit Basu, Yi. Mao, and John C. Platt. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, 2012.

[Zhu and Xing, 2011] Jun Zhu and Eric Xing. Sparse topic coding. In *UAI*, 2011.

[Zhu *et al.*, 2011a] Jun Zhu, Ning Chen, and Eric Xing. Infinite SVM: a Dirichlet process mixture of large-margin kernel machines. In *ICML*, pages 617–624, 2011.

[Zhu *et al.*, 2011b] Jun Zhu, Ning Chen, and Eric P. Xing. Infinite latent SVM for classification and multi-task learning. In *NIPS*, 2011.

[Zhu *et al.*, 2012] Jun Zhu, A. Ahmed, and E.P Xing. MedLDA: maximum margin supervised topic models. *JMLR*, 13:2237–2278, 2012.

[Zhu *et al.*, 2013] Jun Zhu, Xun Zheng, Li Zhou, and Bo Zhang. Scalable inference in max-margin topic models. In *SIGKDD*, 2013.

[Zhu *et al.*, 2014a] Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with data augmentation. *JMLR*, 15:949–986, 2014.

[Zhu *et al.*, 2014b] Jun Zhu, Ning Chen, and Eric P. Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *JMLR*, 15:1799–1847, 2014.

[Zhu *et al.*, 2017] Jun Zhu, Jianfei Chen, Wenbo Hu, and Bo Zhang. Big learning with bayesian methods. *National Science Review*, page nwx044, 2017.

[Zhu, 2012] Jun Zhu. Max-margin nonparametric latent feature models for link prediction. In *ICML*, 2012.

[Zhuo *et al.*, 2018] Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message passing stein variational gradient descent. In *ICML*, 2018.