Bayesian Deep Learning

An Incomplete Tour

Yijun Xiao

UC SANTA BARBARA

Why Bayesian Deep Learning?

Carles Gelada @carlesgelada
Follow

Controversial opinion: Bayesian NNs make no sense. You only want to use Bayes rule if you have a reasonable prior of what the parameters should be. Nobody knows what is encoded by any prior over the weights of a NN. So why would we use such a prior? 1/4

9:20 PM - 21 Dec 2019

121 Retweets 676 Likes

31 121 676

Carles Gelada @carlesgelada · 21 Dec 2019
Sure many regularizations have a Bayesian interpretation… but everyone and their mama has an interpretation of what regularizations do. The question is: What did we gain from it being Bayesian? 2/4

1 1 65

Carles Gelada @carlesgelada · 21 Dec 2019
You could say that BNNs allow us to empirically find regularizations which can't be implemented as anything other than Bayesian priors. Yes, but what is the reason to believe that this space of regularizations is more interesting than any other? Others are easier to work with 3/4

2 1 53

Carles Gelada @carlesgelada · 21 Dec 2019
Another way BNNs could be used is as a Bayesian-meta-learning framework, where we meta learn a prior so that the learning of a new task's weights is fast. Potentially useful but again, no reason to believe this will be a better meta-learning framework than any other. 4/4

6 1 69

Federico Vaggi @F_Vaggi · 22 Dec 2019

Andrew Gordon Wilson @andrewgwils
Follow

Bayesian methods are *especially* compelling for deep neural networks. The key distinguishing property of a Bayesian approach is marginalization instead of optimization, not the prior, or Bayes rule. This difference will be greatest for underspecified models like DNNs. 1/18

Carles Gelada @carlesgelada
Controversial opinion: Bayesian NNs make no sense. You only want to use Bayes rule if you have a reasonable prior of what the parameters should be. Nobody knows what is encoded by any prior over the weights of a NN. So why would we use such a prior? 1/4
Show this thread

4:17 PM - 26 Dec 2019

432 Retweets 1,669 Likes

13 432 1.7K

Andrew Gordon Wilson @andrewgwils · 26 Dec 2019
In particular, the predictive distribution we often want to find is p(y|x,D) = \int p(y|x,w) p(w|D) dw. 'y' is an output, 'x' an input, 'w' the weights, and D the data. This is not a controversial equation, it is simply the sum and product rules of probability. 2/18

1 2 63

Andrew Gordon Wilson @andrewgwils · 26 Dec 2019
Rather than betting everything on a single hypothesis, we want to use every setting of parameters, weighted by posterior probabilities. This procedure is known as a Bayesian model average (BMA). 3/18

2 3 70

Andrew Gordon Wilson @andrewgwils · 26 Dec 2019
Classical training can be viewed as approximate Bayesian inference, where the

UCSB

Why Bayesian Deep Learning?

[Wilson 2019]

- Bayesian is marginalization instead of optimization
- The prior that matters is the prior in function space, not parameter space
- Priors without marginalization are simply regularization, but Bayesian methods are not about regularization

UCSB

Why Bayesian Deep Learning?

[Ghahramani 2016]

- Calibrated model and prediction uncertainty: getting systems that know when they don't know
- Automatic model complexity control and structure learning (Bayesian Occam's Razor)

UCSB

Why Bayesian Deep Learning?

[Teh 2017]

- A normative account of "best" learning given model and data
- Explicit expression of all prior knowledge/inductive biases in model
- Unified treatment of uncertainties
- Common language with statistics, applied sciences

UCSB

Outline

- Preliminaries
- Bayesian Neural Networks and Uncertainty Quantification
- Deep Latent Variable Models

"Deep Bayesian Learning"

UCSB

Outline

- Preliminaries
- Bayesian Neural Networks and Uncertainty Quantification
- Deep Latent Variable Models

UCSB

Bayesian Learning

Given data set

$$X = \{x_i\}_{i=1}^N, Y = \{y_i\}_{i=1}^N$$

In many situations, we want to model the distribution of y given an input x and all the observations

$$p(y|x, X, Y)$$

so that we could predict y for any new input x

UCSB

Bayesian Learning

Notice the distribution can be written as

$$p(y|x, X, Y) = \int_w p(y|x, w)p(w|X, Y)dw$$

where w is the set of weights for a function f(x;w)

UCSB

Bayesian Learning

Depending on the task, we have different definitions for p(y|x,w):

- Regression:

$$p(y|x, w) = \mathcal{N}(y; f(x; w), \sigma^2)$$

- Classification:

$$p(y = k|x, w) = \frac{\exp(f_k(x; w))}{\sum_i \exp(f_i(x; w))}$$

UCSB

Bayesian Learning

In this context, model training becomes finding the posterior distribution of w:

$$p(w|X, Y)$$

This term is closely related to maximum a posteriori (MAP) optimization. So is MAP Bayesian? No

UCSB

Bayesian Inference

The true posterior distribution of w usually can not be solved analytically.

- Markov Chain Monte Carlo [Neal 1995, Welling & Teh 2011]
- Variational Inference [Hinton & van Camp 1993]

UCSB

Variational Inference

Idea: propose a variational distribution of the variable and push it close to the true posterior

$$\theta^* = \arg\min_{\theta} \mathrm{KL}(q(w|\theta) \| p(w|X,Y))$$

Minimizing the KL divergence is equivalent to maximizing the evidence lower bound (ELBO)

$$\theta^* = \arg\min_{\theta} \mathrm{KL}(q(w|\theta) \| p(w)) - \mathbb{E}_{q(w|\theta)}[\log p(Y|X,w)]$$

Complexity cost

Likelihood cost

UCSB

Variational Inference

What are the challenges?

- How to evaluate the gradient of the expectation?
- How to choose prior and posterior distribution family?
- How to adapt to large scale training data?

UCSB

Outline

- Preliminaries
- Bayesian Neural Networks and Uncertainty Quantification
- Deep Latent Variable Models

UCSB

Bayesian Neural Networks

Earlier attempts [Hinton & van Camp 1993, Barber & Bishop 1998, Graves 2011] face challenges with scaling to more complex neural network structures. More recent works:

- Bayes by Backprop [Blundell et al. 2015]
- Monte Carlo Dropout [Gal & Ghahramani 2016]

UCSB

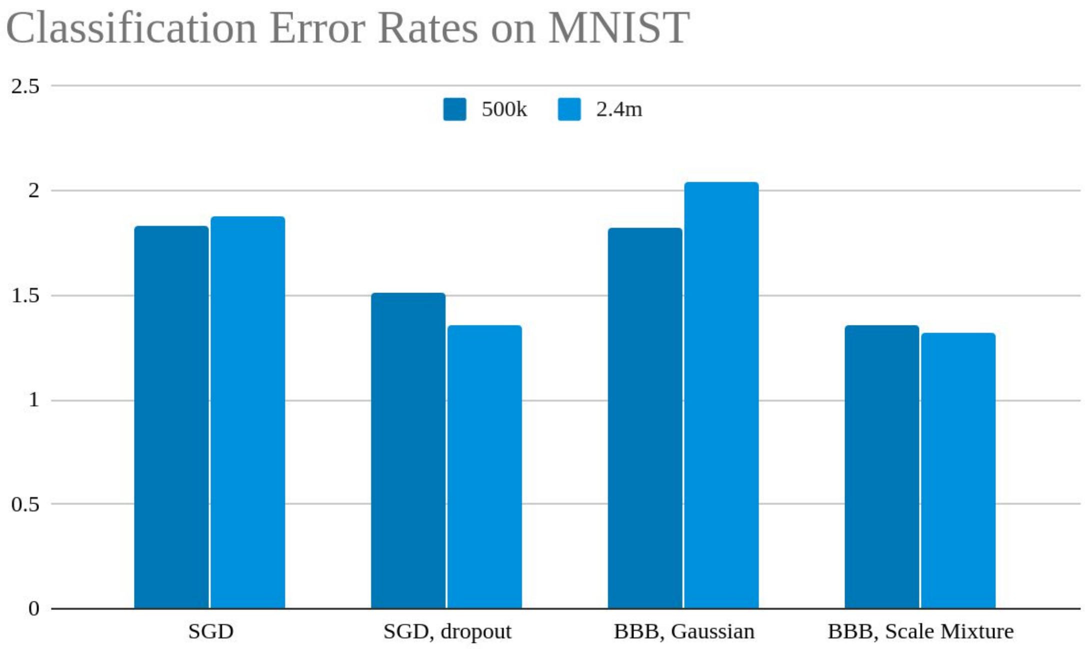
Bayes by Backprop

- Applies reparameterization trick to obtain unbiased low-variance estimate of the gradients
- Allows for broader prior and posterior families without closed-form complexity cost

$$\frac{\partial}{\partial \theta}\mathbb{E}_{q(\mathbf{w}|\theta)}[f(\mathbf{w},\theta)] = \mathbb{E}_{q(\epsilon)}\left[\frac{\partial f(\mathbf{w},\theta)}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial \theta} + \frac{\partial f(\mathbf{w},\theta)}{\partial \theta}\right]$$

UCSB

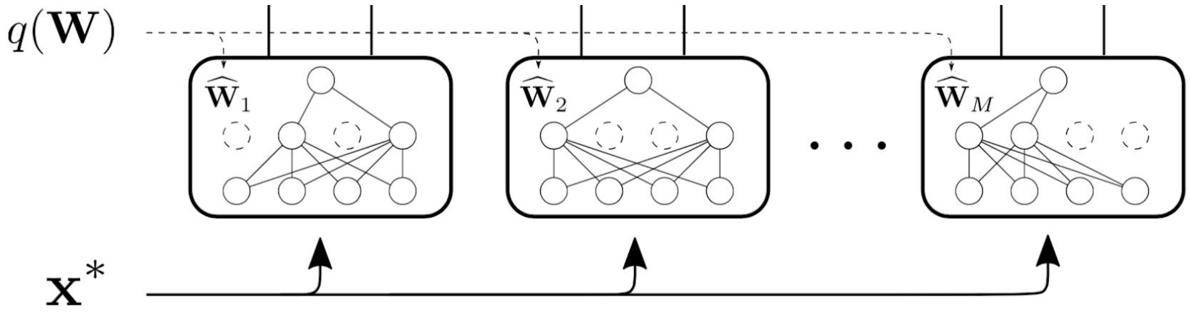
Bayes by Backprop



UCSB

Monte Carlo Dropout

DNNs with dropout layers trained with SGD perform variational inference.



UCSB

Monte Carlo Dropout

Pros:

- Exactly the same model implementation if dropout is present
- Number of parameters is the same instead of 2x

Cons:

- There might be underlying assumptions that are not obvious [Osband 2016]

UCSB

Outline

- Preliminaries
- Bayesian Neural Networks and Uncertainty Quantification
- Deep Latent Variable Models

UCSB

Non-Bayesian Approaches

- Model Calibration [Guo et al. 2017]
- Out-Of-Distribution Input Detection [Hendrycks & Gympel 2016, Liang et al. 2017]
- Deep Ensembles [Lakshminarayanan et al. 2017]

UCSB

Types of Uncertainties

Law of total variance:

$$\mathrm{Var}(y) = \mathrm{Var}\left(\mathbb{E}[y|x]\right) + \mathbb{E}\left[\mathrm{Var}(y|x)\right]$$

- Epistemic (model) uncertainties arise from the uncertainties about the model parameters
- Aleatoric (data) uncertainties are inherent uncertainties in the data or the measurement

UCSB

How to Quantify Uncertainties?

Epistemic Uncertainty

- Bayesian Neural Networks (BNN) model weights as Gaussian random variables
- Sample weights from the posterior distribution and measure output variance. e.g. MC Dropout [Gal & Ghahramani 2016]

UCSB

How to Quantify Uncertainties?

Aleatoric Uncertainty

- Model outputs a Gaussian distribution instead of a point estimate

$$y \sim \mathcal{N}\left(\mu(\mathbf{x}), \sigma(\mathbf{x})^2\right)$$

UCSB

How to Quantify Uncertainties?

Aleatoric Uncertainty

- Minimizing the negative log likelihood instead of the conventional MSE

$$\mathcal{L}_{\text{rgs}}(\mathbf{W}) = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i | \mu(\mathbf{x}_i), \sigma(\mathbf{x}_i))$$

UCSB

What About Classification?

- Measure variance in the logit space [Kendall & Gal 2017]
- Decompose entropy [Depeweg et al. 2018]

$$\mathrm{H}[y_\star|\mathbf{x}_\star] - \mathbf{E}_{q(\mathcal{W})}[\mathrm{H}(y_\star|\mathcal{W},\mathbf{x}_\star)] = I(y_\star,\mathcal{W})$$

- Dirichlet Prior Networks [Malinin & Gales 2018]

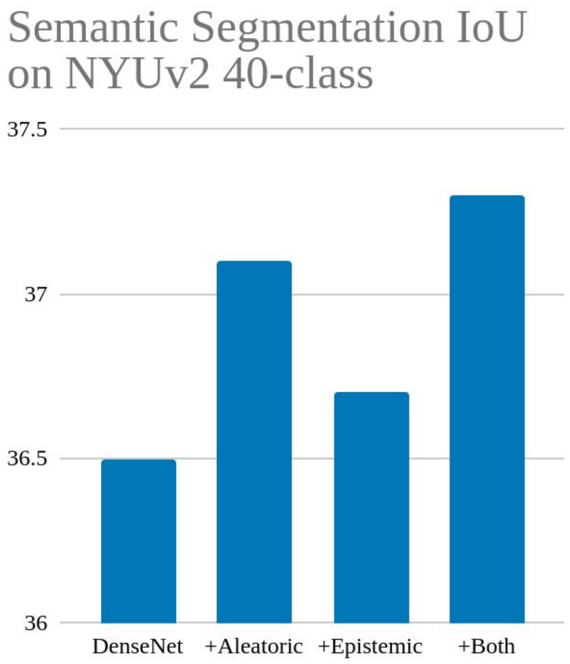
UCSB

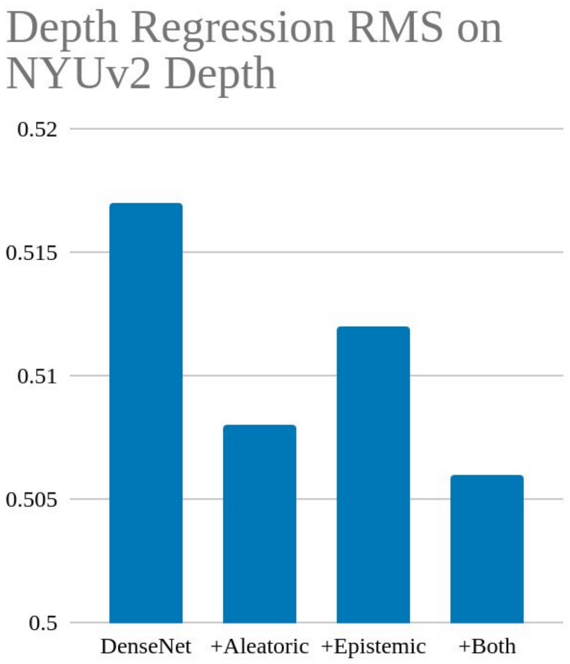
Uncertainty Quantification

- What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? [Kendall & Gal 2017]
- Deep and Confident Prediction for Time Series at Uber [Zhu & Laptev 2017]

UCSB

Kendall & Gal 2017





UCSB

Kendall & Gal 2017

Train dataset	Test dataset	RMS	Aleatoric variance	Epistemic variance
Make3D / 4	Make3D	5.76	0.506	7.73
Make3D / 2	Make3D	4.62	0.521	4.38
Make3D	Make3D	3.87	0.485	2.78
Make3D / 4	NYUv2	-	0.388	15.0
Make3D	NYUv2	-	0.461	4.87

(a) Regression

Train dataset	Test dataset	IoU	Aleatoric entropy	Epistemic logit variance ($\times 10^{-3}$)
CamVid / 4	CamVid	57.2	0.106	1.96
CamVid / 2	CamVid	62.9	0.156	1.66
CamVid	CamVid	67.5	0.111	1.36
CamVid / 4	NYUv2	-	0.247	10.9
CamVid	NYUv2	-	0.264	11.8

(b) Classification

Table 3: Accuracy and aleatoric and epistemic uncertainties for a range of different train and test dataset combinations. We show aleatoric and epistemic uncertainty as the mean value of all pixels in the test dataset. We compare reduced training set sizes (1, ½, ¼) and unrelated test datasets. This shows that aleatoric uncertainty remains approximately constant, while epistemic uncertainty decreases the closer the test data is to the training distribution, demonstrating that epistemic uncertainty can be explained away with sufficient training data (but not for out-of-distribution data).

Zhu & Laptev 2017

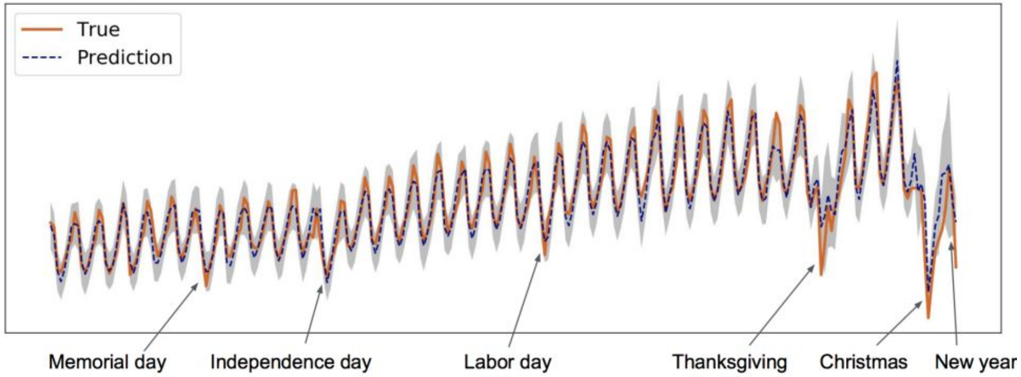


Figure 2. Daily completed trips in San Francisco during eight months of the testing set. True values are shown with the orange solid line, and predictions are shown with the blue dashed line, where the 95% prediction band is shown as the grey area. Exact values are anonymized.

UCSB

Outline

- Preliminaries
- Bayesian Neural Networks and Uncertainty Quantification
- Deep Latent Variable Models

UCSB

Why Use Latent Variables?

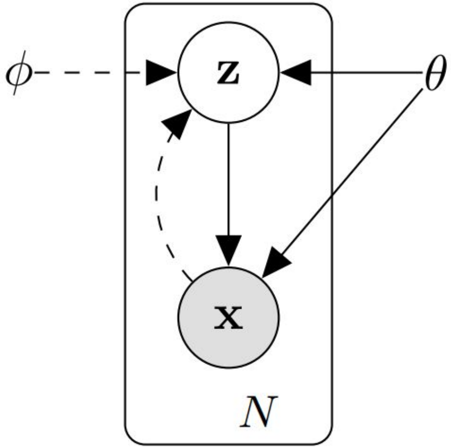
- Semi-supervised learning
- Incorporating prior knowledge
- Modeling multimodal distribution
- Interpretable representation
- ...

UCSB

Variational Autoencoders

What makes it deep?

p(x|z) and q(z|x) are neural networks



UCSB

Variational Autoencoders

$$\mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right]$$

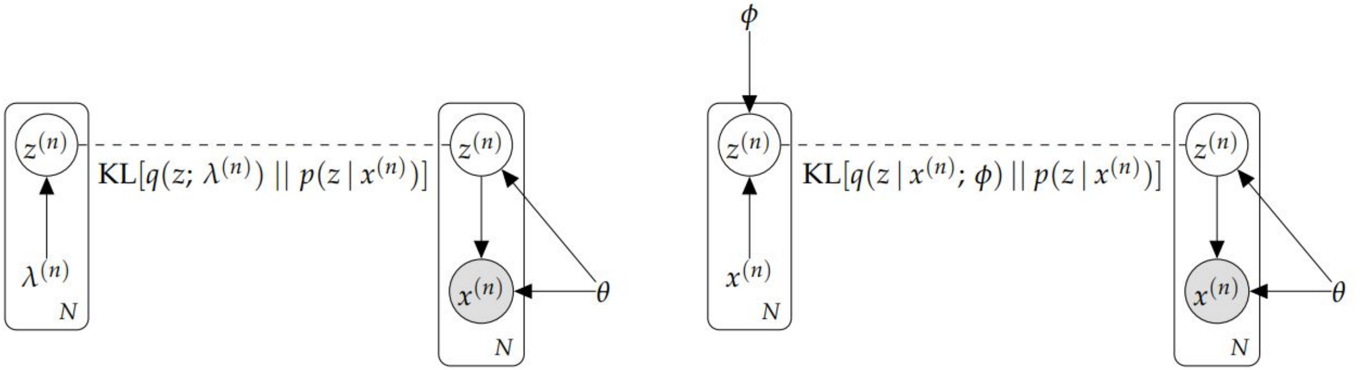


Figure 9: (Left) Traditional variational inference uses variational parameters $\lambda^{(n)}$ for each data point $x^{(n)}$. (Right) Amortized variational inference employs a global inference network ϕ that is run over the input $x^{(n)}$ to produce the local variational distributions.

[Kim et al. 2019]

UCSB

Variational Autoencoder

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right]$$

How to optimize this? SGVB

Why do we need reparameterization trick? Low variance gradient estimate

UCSB

Variational Autoencoder for Text

Posterior Collapse / KL Vanishing

- KL annealing [Bowman et al. 2016]
- Drop word [Bowman et al. 2016]
- Different decoders [Miao et al. 2016, Yang et al. 2017]
- Bag-of-word loss [Zhao et al. 2017]
- ...

UCSB

Non-Gaussian Latent Distributions

- Discrete [Maddison et al. 2017, Jang et al. 2017]
- Gaussian Mixture [Dilokthanakul et al. 2017] (Clustering)
- Logistic Normal [Srivastava & Sutton 2017] (Topic Modeling)
- von Mises-Fisher [Davidson et al. 2018, Xu & Durrett 2018]
- Gaussian Process [Tran et al. 2016]
- Stick Breaking Process [Nalisnick & Smyth 2017]
- …

UCSB

Tightening the Gap

ELBO is the lower bound of the evidence (hence the name).

- Normalizing Flow [Rezende & Mohamed 2015]
- Importance Weighted Autoencoders [Burda et al. 2016]

UCSB

Normalizing Flow

- Transform a simple distribution (e.g. a simple Gaussian) into a complex one through a chain of invertible transformations.
- Density of the complex variable can be derived using change of variable theorem (need Jacobian of the invertible transformations).

$$
\begin{aligned}
\mathbf{z}_0 &\sim q(\mathbf{z}_0 \mid x; \phi) = \mathcal{N}(\boldsymbol{\mu}(x), \boldsymbol{\sigma}^2(x)) \\
\mathbf{z}_K &= f_K \circ f_{K-1} \circ \cdots \circ f_1(\mathbf{z}_0).
\end{aligned}
$$

$$
\begin{aligned}
\log q_K(\mathbf{z}_K \mid x; \phi) &= \log q(\mathbf{z}_0 \mid x; \phi) + \sum_{k=1}^{K} \log \left| \frac{\partial f_k^{-1}}{\partial \mathbf{z}_k} \right| \\
&= \log q(\mathbf{z}_0 \mid x; \phi) - \sum_{k=1}^{K} \log \left| \frac{\partial f_k}{\partial \mathbf{z}_{k-1}} \right|
\end{aligned}
$$

UCSB

Importance Weighted Autoencoders

$$\mathcal{L}_k(\mathbf{x}) = \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k \sim q(\mathbf{h}|\mathbf{x})} \left[\log \frac{1}{k} \sum_{i=1}^{k} \frac{p(\mathbf{x}, \mathbf{h}_i)}{q(\mathbf{h}_i|\mathbf{x})} \right].$$

Theorem 1. *For all k, the lower bounds satisfy*

$$\log p(\mathbf{x}) \geq \mathcal{L}_{k+1} \geq \mathcal{L}_k.$$

Moreover, if $p(\mathbf{h}, \mathbf{x})/q(\mathbf{h}|\mathbf{x})$ is bounded, then $\mathcal{L}_k$ approaches $\log p(\mathbf{x})$ as k goes to infinity.

UCSB

Thank you!

UCSB

References

Bayesian Learning and Bayesian Neural Networks

1. Hinton, Geoffrey E., and Drew Van Camp. "Keeping the neural networks simple by minimizing the description length of the weights." Proceedings of the sixth annual conference on Computational learning theory. 1993.
2. Neal, Radford M. BAYESIAN LEARNING FOR NEURAL NETWORKS. Diss. University of Toronto, 1995.
3. Barber, David, and Christopher M. Bishop. "Ensemble learning in Bayesian neural networks." Nato ASI Series F Computer and Systems Sciences 168 (1998): 215-238.
4. Welling, Max, and Yee W. Teh. "Bayesian learning via stochastic gradient Langevin dynamics." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.
5. Graves, Alex. "Practical variational inference for neural networks." Advances in neural information processing systems. 2011.
6. Blundell, Charles, et al. "Weight Uncertainty in Neural Network." International Conference on Machine Learning. 2015.
7. Fortunato, Meire, Charles Blundell, and Oriol Vinyals. "Bayesian recurrent neural networks." arXiv preprint arXiv:1704.02798 (2017).
8. Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." international conference on machine learning. 2016.
9. Osband, Ian. "Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout." NIPS Workshop on Bayesian Deep Learning. Vol. 192. 2016.
10. Sun, Shengyang, et al. "Functional variational bayesian neural networks." arXiv preprint arXiv:1903.05779 (2019).
11. Wilson, Andrew Gordon. "The Case for Bayesian Deep Learning." NYU Courant Technical Report. 2019.

UCSB

References

Uncertainty Quantification

1. Guo, Chuan, et al. "On calibration of modern neural networks." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
2. Hendrycks, Dan, and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks." arXiv preprint arXiv:1610.02136 (2016).
3. Liang, Shiyu, Yixuan Li, and Rayadurgam Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks." arXiv preprint arXiv:1706.02690 (2017).
4. Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems. 2017.
5. Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." Advances in neural information processing systems. 2017.
6. Depeweg, Stefan, et al. "Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning." International Conference on Machine Learning. 2018.
7. Malinin, Andrey, and Mark Gales. "Predictive uncertainty estimation via prior networks." Advances in Neural Information Processing Systems. 2018.
8. Zhu, Lingxue, and Nikolay Laptev. "Deep and confident prediction for time series at uber." 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017.
9. Xiao, Yijun, and William Yang Wang. "Quantifying uncertainties in natural language processing tasks." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

UCSB

References

Deep Latent Variable Models and Variational Autoencoders

1. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
2. Bowman, Samuel, et al. "Generating Sentences from a Continuous Space." Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. 2016.
3. Miao, Yishu, Lei Yu, and Phil Blunsom. "Neural variational inference for text processing." International conference on machine learning. 2016.
4. Yang, Zichao, et al. "Improved variational autoencoders for text modeling using dilated convolutions." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
5. Zhao, Tiancheng, Ran Zhao, and Maxine Eskenazi. "Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.
6. Dilokthanakul, Nat, et al. "Deep unsupervised clustering with gaussian mixture variational autoencoders." arXiv preprint arXiv:1611.02648 (2016).
7. Srivastava, Akash, and Charles Sutton. "Autoencoding variational inference for topic models." arXiv preprint arXiv:1703.01488 (2017).
8. Xu, Jiacheng, and Greg Durrett. "Spherical latent spaces for stable variational autoencoders." arXiv preprint arXiv:1808.10805 (2018).
9. Davidson, Tim R., et al. "Hyperspherical variational auto-encoders." arXiv preprint arXiv:1804.00891 (2018).
10. Tran, Dustin, Rajesh Ranganath, and David M. Blei. "The variational Gaussian process." 4th International Conference on Learning Representations, ICLR 2016. 2016.

UCSB

References

Deep Latent Variable Models and Variational Autoencoders

11. Maddison, Chris J., Andriy Mnih, and Yee Whye Teh. "The concrete distribution: A continuous relaxation of discrete random variables." International Conference on Learning Representations (ICLR 2017). OpenReview. net, 2017.
12. Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical Reparametrization with Gumble-Softmax." International Conference on Learning Representations (ICLR 2017). OpenReview. net, 2017.
13. Nalisnick, Eric, and Padhraic Smyth. "Stick-breaking variational autoencoders." International Conference on Learning Representations (ICLR). 2017.
14. Rezende, Danilo, and Shakir Mohamed. "Variational Inference with Normalizing Flows." International Conference on Machine Learning. 2015.
15. Tomczak, Jakub M., and Max Welling. "Improving variational auto-encoders using householder flow." arXiv preprint arXiv:1611.09630 (2016).
16. Kingma, Durk P., et al. "Improved variational inference with inverse autoregressive flow." Advances in neural information processing systems. 2016.
17. Louizos, Christos, and Max Welling. "Multiplicative normalizing flows for variational bayesian neural networks." arXiv preprint arXiv:1703.01961 (2017).
18. Burda, Yuri, Roger Grosse, and Ruslan Salakhutdinov. "Importance weighted autoencoders." arXiv preprint arXiv:1509.00519 (2015).
19. Kim, Yoon, Sam Wiseman, and Alexander M. Rush. "A tutorial on deep latent variable models of natural language." arXiv preprint arXiv:1812.06834 (2018).

UCSB