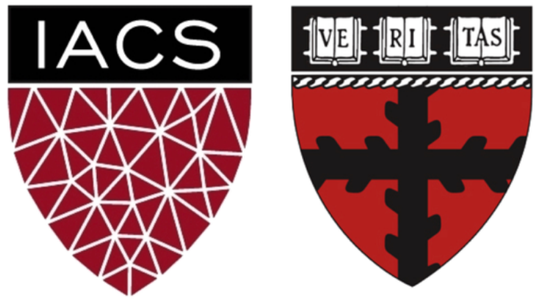
Lecture #19: Variational Inference in Context

AM 207: Advanced Scientific Computing

Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020


IACS
VE RI TAS

Outline

1. How to Evaluate Approximate Inference
2. How to Improve Approximate Inference
3. Why do I care?

How to Evaluate Approximate Inference

How Good is Your Variational Approximation of the True Posterior?

Question: Why is it hard to check that the variational posterior is a good approximation of the true posterior?

You can't simply visualize the true posterior, since it's 1) high dimensional, 2) intractable to sample from.

Question: What if we computed the KL-divergence between the true posterior and the varaitional approximation?

"Practical Posterior Error Bounds from Variational Objectives": Unfortunately, even when the KL-divergence between q and p is effectively zero, the difference between the means and variances of q and p can be **arbitrarily large**. That is, a small KL-divergence doesn't capture our intuition about what it means for two distributions to be similar.

Alternative Posterior Evaluation Metrics

In "Yes, But Did It Work: Evaluating Variational Inference", the authors proposes two alternative posterior evaluation metrics:

1. The Pareto Smoothed Importance Sampling Diagnostic: If the variational approximation q is very different than p, then the importance sampling MC estimate of $\mathbb{E}_{p(\theta|\text{Data})}\left[f(\theta)\right]$,

$$\mathbb{E}_{p(\theta|\text{Data})}\left[f(\theta)\right] \approx \frac{1}{S}\sum_{s=1}^{S}\frac{p(\theta|\text{Data})}{q(\theta)}f(\theta_s),\ \ \theta_s \sim q(\theta)$$

will have very high variance, due to the fact that the importance weights $\frac{p(\theta|\text{Data})}{q(\theta)}$ will be very heterogeneous. Thus, methods of smoothing the weights, like ***Pareto Smoothed Importance Sampling (PSIS)*** will have poor performance. The efficacy of PSIS can be used as a diagnostic tool for testing if q is similar to p.

2. The Variational Simulation-Based Calibration (VSBC) Diagnostic: Given a Bayesian model $p(y, \theta) = p(y|\theta)p(\theta)$, if your variational inference proceedure is good at approximating the posterior $p(\theta|y)$ then it should do well for most sets of data that is generated from the model $p(y, \theta)$.

We generate M sets of synthetic data from $p(y, \theta)$. For each set of the data we approximate the posterior $q_m(\theta)$. We then perform a calibration test on each variational approximation. This will reveal if your variational inference procedure is biased - produces approximation that are consistently flawed in some specific way.

How to Improve Approximate Inference

Your Variational Approximation Sucks, Can You Fix It?

The current research on improving variational inference can be categorize by which **design choice** each work tries to improve:

1. Choice of Divergence: In Lab #5, you explored the draw-backs of fitting an approximate posterior q to the true posterior p using KL-divergence. There are a huge number of works that explore performing variational inference using other types of divergences, for example:

- **a.** "Black-box α-divergence Minimization"
- **b.** "Stein Variational Gradient Descent"

2. Choice of Variational Family: In Homework #0, you've seen that even for a simple Bayesian linear regression model, the model parameters were corrlated in the posterior. In in-class 10.5, you've seen that minimizing the KL-divergence to fit an isotropic Gaussian means that you can only capture one mode in the posterior. There are a huge number of works that explore using different types of variational families:

- **a.** "Variational Inference with Normalizing Flows"
- **b.** "The Variational Gaussian Process"

3. Choice of Optimization Procedure: Even if your divergence measure and variational family are well-chosen, the optimization objective can still be non-convex! This means that your optimization procedure might return a q that is a only local-optimum. There are a large number of works that address how to jump out of local optima using SGD and some works that specifically address the optimization challenges of variational inference:

- **a.** "Proximity Variational Inference"

Are There Any Good Properties of Variational Approximations that We Are Sure About?

In "Frequentist Consistency of Variational Bayes", the authors prove that under **certain assumptions** on $p(\theta)$ and $p(y|\theta)$:

1. as the number of observation increases, the variational approximation converges (in Total Variation distance) to the q that minimizes the KL-divergence to a normal distribution centered at the ground truth parameters θ_{true} that generated the data.

2. the mean of the variational approximation is consistent and asymptoptically normal.

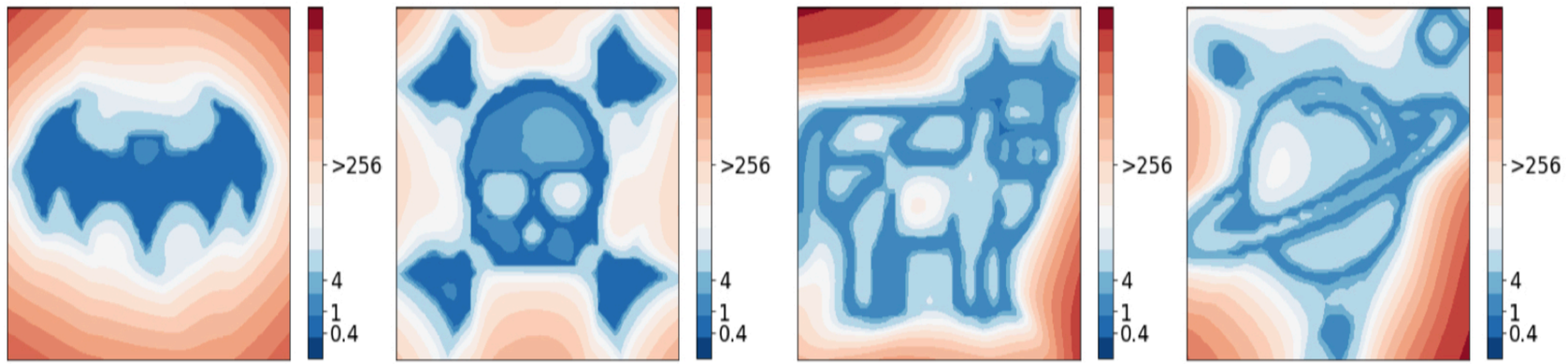
Unfortunately, the assumptions required by the theorems **do not hold for neural network likelihood models**.

Why is Inference for Neural Networks Hard?

Research on the likelihoods (and hence posteriors) of neural networks are beginning to give us ways of visualizing these high-dimensional functions using low-dimensional (3D) projections. Loss Landsacapes represents some of the latest efforts at visualization:

It's Weirder Than You Can Imagine

In *"Loss Landscape Sightseeing with Multi-Point Optimization"*, the authors show that neural network likelihoods (and thus posteriors) are so complicated that you can find a 2-D projection that can create any pattern you want:



(a) Loss surface on FashionMNIST dataset (b) Loss surface on CIFAR10 dataset

The lesson: take every low-dimensional visualization with a grain of salt (i.e. it far from accurately represents the entire landscape).

But why do I care?

Evaluation of Variational Approximation for Real Down-stream Tasks

In practice, you may only care about the quality of your posterior approximation in so far as it affects model performance on your down-stream task. So what do we want from our machine learning or statistical models, especially when they are used in safety, or fairness critical applications (e.g. personalized medicine, health-care resource allocation, criminal justic systems)?

Most of us in the community agree that we want models that *1)* makes accurate predictions *2)* gives realistic estimates of its prective uncertainty. So that the model can be held-accountable by humans in the system. There are a number of subcomunities in ML that focus on studying the social impact of machine learning models as well as how to design models whose negative impact can be mitigated.

a. "Concrete Problems in AI Safety"
b. "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer"

Predictiveness of the Variational Approximate Posterior

We can check the predictiveness of our varational approximation of the posterior by sampling models from the posterior and then making predictions using these models by sampling from the likelihood. We then check that these predictions align well with observed data by:

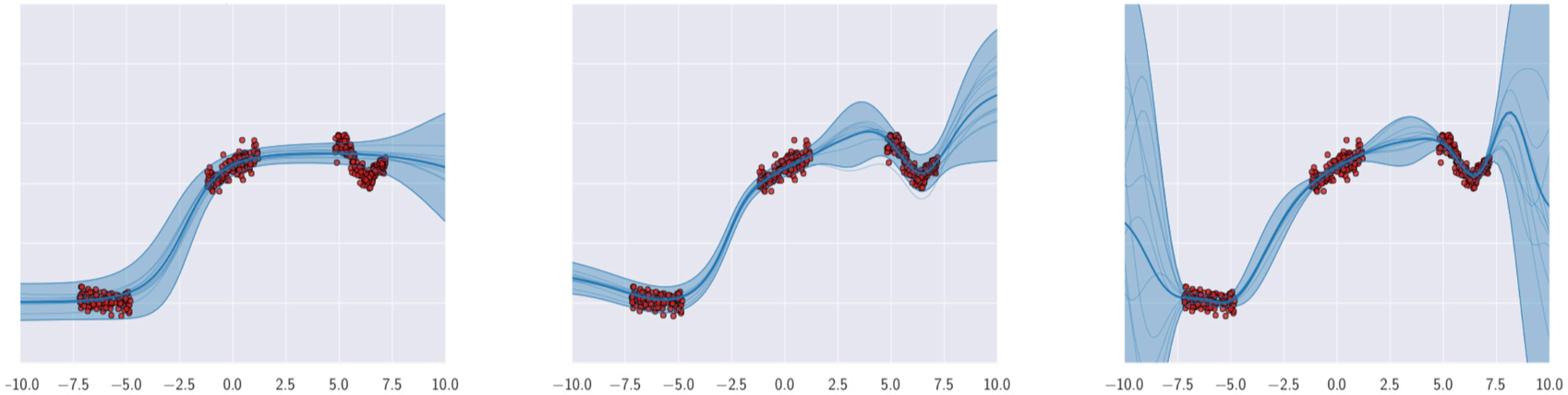
1. **Visualization:** visualizing the posterior predictive against the observed data. *This is generally impossible for high dimensional data.*

2. **Log Marginal Data Likelihood:** we compute the expected value of the log likelihood of the test data under our approximate posterior. *Log likelihood is only useful for comparing two different models, given a single model it is hard to say if a log likelihood value is "good enough".*

3. **Expected Mean Square Error and Accuracy:** we compute the expected MSE (for regression) or accuracy (for classification) on test data under our approximate posterior. *While these metric frames model quality in concrete task related terms, they are each misleading when the data contains outliers or imbalanced classes.*

Quality of Variational Approximate Posterior Predictive Uncertainty

We've argued in this course that good predictive uncertainty must involve an accurate assessment of both ***epistemic*** and ***aleatoric*** uncertainty. Since each type of uncertainty requires a different risk-management action in the down-stream task.



Unfortunately, there isn't a single good statistical metric for assessing the quality of the uncertainty of a model. "Good" epstemic uncertainty is especially hard to quantify. Rather than looking for statistical tests of uncertainty, some in the ML community advocate for assessing model uncertainty with respect to a set of benchmark down-stream tasks: Bayesian Deep Learning Benchmarks

Does a Poor Posterior Approximation Imply a Poor Posterior Predictive?

You've seen in HW#8 that even with an HMC sampler that is far from converged, you were able to produce posterior predictives that aligned well with the data and had good epistemic and aleatoric uncertainty. In fact, a number of works are showing that by capturing a little piece (if it is the right piece) of the true posterior of a BNN, you are able to capture most of the variation you want in the posterior predictive.

In "Subspace Inference for Bayesian Deep Learning", the authors provide toy examples where one obtains a posterior predictive distribution with good epistemic and aleatoric uncertainty by reducing a high dimensional parameter space to 2-dimensions and performing inference in the 2-dimensional subspace: