Optimization Operators as Evidence Estimators

Dmitry Bylinkin **Andrei Semenov** **Alexander Shestakov**
Vladimir Solodkin

1 Introduction

In this blog, we will talk about how optimization operators can be useful for evidence estimation. Why does this matter? Well, estimating evidence allows models to more accurately capture data distributions, especially when generating new data, where understanding the true likelihood is often a complex challenge. In this blog, we are focused on the so-called early-stoping technique, when the data are small enough and we do not want to form a validation set from them.

Traditionally, likelihoods are estimated directly through variational inference techniques. However, research suggests that optimization operators can act as effective proxies for evidence, ultimately enhancing our model's ability to capture data patterns. In this post, we'll explore some mathematics behind this approach and figure out how it works in practice.

We present a library that implements methods from basic to state-of-the-art as pytorch classes.

2 Overview of Evidence Estimate Approaches

2.1 Basic SGD Approach

Let's do some not-so-formal math around our topic. The problem of finding the optimal distribution $q(\theta)$ that approximates posterior $p(\theta|x)$ is usually posed as the problem of minimizing the forward Kulback-Leibler divergence:

$$\min_{p(x)} \left[KL(q(\theta)||p(\theta|x)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta|x)} dx \right].$$

It can be shown - we leave it without proof - that minimizing KL is equivalent to maximizing evidence. Evidence consists of two terms: energy and entropy. The energy term measures how well q fits the data and the entropy term encourages the probability mass of q to spread out, preventing overfitting. This means that, for our purposes, we need to somehow learn how to compute the entropy change in order to stop learning in time. Let q_0 be some initial distribution. Then the entropic term with accuracy to constants can be written as

$$S[q_T] = \sum_{t=0}^{T-1} \log |J(\theta_t)|.$$

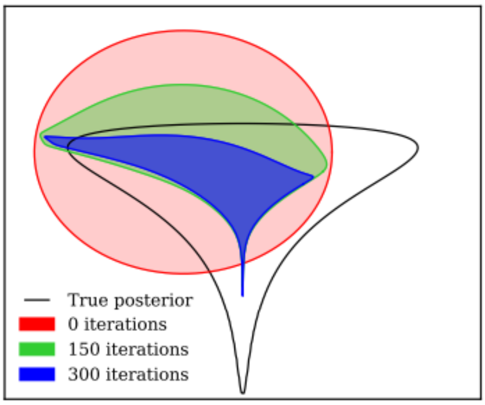

Figure 1: The variational distribution implied by the modified, "entropy-friendly", SGD algorithm

Using the SGD step, we obtain

$$S[q_{t+1}] \approx S[q_t] + \log \left| I - \alpha \nabla^2 (-\log p_t(\theta_t, x)) \right| .$$

Thus, we can update the entropy value at every iteration of SGD (Maclaurin et al., 2015). In low dimensional spaces, we can compute the log determinant exactly. If the dimensionality is large and it is computationally challenging, there are a number of approximate approaches that are implemented in our library along with exact computation.

2.2 Applying Langevin Dynamics to Evidence Estimate

TODO (Welling and Teh, 2011)

2.3 Stochastic Gradient Fisher Scoring

In practical problems both sampling accuracy and mixing rate are important. If the posterior is close to Gaussian, we would like to take advantage of high mixing rate. However, if we need to capture a highly non-Gaussian posterior, we should be able to trade-off mixing rate for sampling accuracy. SGFS can be seen as a scalable Bayesian MCMC algorithm, where variational perspective is used to rederive the Fisher scoring update. The basic idea is preconditioned gradient and added Gaussian noise. More precisely,

$$\theta_{t+1} = \theta_t - \varepsilon H \widehat{\nabla} \theta_t + \sqrt{\varepsilon} H \mathcal{N}(0, Q_t)$$

(Stephan et al., 2017)

2.4 Constant SGD as Variational EM

TODO

3 Implementation

The most famous and developed Python library with a built-in differentiation engine is PyTorch (Paszke et al., 2019). For this reason, we chose PyTorch as the foundation for our library. All methods we have implemented are modifications of SGD in one way or another, so we inherit from the `torch.optim.SGD` class to develop our own ones.

4 Demo

TODO

5 Conclusion

In summary, our library could be useful for researchers and practitioners interested in model selection tools. By offering a comprehensive set of approaches, our library aims to make model selection process more efficient and accessible. We invite you to explore our library, test out the demo, and join us in advancing its development.

References

Dougal Maclaurin, David Duvenaud, and Ryan P Adams. 2015. Early stopping is nonparametric variational inference. *arXiv preprint arXiv:1504.01344* (2015).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

Mandt Stephan, Matthew D Hoffman, David M Blei, et al. 2017. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research* 18, 134 (2017), 1–35.

Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 681–688.