Statistical Modeling, Causal Inference, and Social Science

Bayesian Learning via Stochastic Gradient Langevin Dynamics

Posted by Andrew on 4 August 2012, 9:58 am

Burak Bayramli writes:

> In this paper by Sunjin Ahn, Anoop Korattikara, and Max Welling and this paper by Welling and Yee Whye The, there are some arguments on big data and the use of MCMC. Both papers have suggested improvements to speed up MCMC computations. I was wondering what your thoughts were, especially on this paragraph:
>
> > When a dataset has a billion data-cases (as is not uncommon these days) MCMC algorithms will not even have generated a single (burn-in) sample when a clever learning algorithm based on stochastic gradients may already be making fairly good predictions. In fact, the intriguing results of Bottou and Bousquet (2008) seem to indicate that in terms of "number of bits learned per unit of computation", an algorithm as simple as stochastic gradient descent is almost optimally efficient. We therefore argue that for Bayesian methods to remain useful in an age when the datasets grow at an exponential rate, they need to embrace the ideas of the stochastic optimization literature."
>
> My [Bayramli's] argument against this is that Bayesian models are more expressive, and coupled with MCMC they allowed researchers to solve previously impossible problems. Performance issue can be solved in time, IMHO.

My reply:

I glanced at the papers only quickly but the general idea makes sense. I've thought for awhile that the Bayesian central limit theorem should allow efficient inference via data partitioning, but my only attempt was not particularly successful (which is why this 2005 paper with Zaiying Huang is unpublished; in fact I don't even recall if we submitted it anywhere). So, just in general terms, I like what Ahn et al. are doing.

I also feel warmly about ideas of combining stochastic optimization with Hamiltonian dynamics and MCMC sampling, as this is what we are doing with Nuts.

Finally, it often seems that methodological advances come from solving applied problems that are in our way. I like the papers you link to because they appear to be motivated by new applications rather than being new methods applied to benchmark problems.

Filed under Bayesian Statistics, Statistical computing | Permalink

4 Comments

1. *John Myles White* says:
   August 4, 2012 at 10:42 am

   

   For people interested in another potential competitor to SGD-style computations while allowing large-scale Bayesian analysis, I would suggest searching for the ArXiv paper on stochastic variational inference by Hoffman et al.

2. *Bob Carpenter* says:
   August 4, 2012 at 11:57 am

   

   The main problem from a Bayesian perspective of pure MAP estimates as produced by optimization algorithms such as stochastic gradient descent (or other optimization algorithms) is that the MAP estimate can be degenerate. For instance, even a normal mixture model that estimates variance (aka soft K-means in the machine learning literature) will devolve to zero-variance clusters. This paper is using SGD to get close to the posterior modes (which will be the MAP estimate in a unimodal problem such as non-hierarchical logistic regression), then transition to Langevin (aka one-step Hamiltonian Monte Carlo) to get both an estimate of the posterior mean and variance.

   A popular alternative is to use variational methods (algorithmically similar to EM after some analytical work up front), which provide approximations to posterior means and to posterior variance. I don't know if anyone has worked on transitioning variational methods to full sampling (like Langevin or HMC).

   Andrew often suggests doing something like this using point estimates (such as provided by Doug Bates's LMER package) and then Metropolis steps after that. The problem with this approach is that LMER isn't nearly as scalable as something like SGD.

   A big practical issue with SGD is setting step size, to which it's very sensitive (for the same reasons that HMC is so sensitive to step size).

   - *Jared* says:
     August 4, 2012 at 5:31 pm

     

     The *real* problem with MAP from a Bayesian perspective is that there's nothing especially Bayesian about a MAP estimate when it isn't motivated by a 0-1 loss function. :)

3. *Bob Carpenter* says:
   August 6, 2012 at 10:24 am

   

   Absolutely — point estimates of any kind are just approximations to full Bayes.