Module 5: Advanced inference methods

5th June 2019

Variational inference

Overview

- So far we have discussed MAP and MCMC methods for approximating posteriors.
- These strategies represent extremes in terms of accuracy and computational complexity.
- There are many other approaches which lie between MAP and MCMC with different trade-offs between computational complexity and accuracy.
- Variational inference (VI) is one such approach.
  - In practice VI is typically not much more computationally expensive than MAP, but provides much better posterior approximations.
- There has been a renewed interest in VI over the last five years.
  - As a result VI is much easier to use on a wide variety of problems.

Basics

- The main idea is to frame posterior approximation as an optimization function.
- To objective of the optimisation function is some measure of similarity between the posterior, $p(\theta|X)$, and the variational approximation $q(\theta|\lambda)$.
  - The parameter λ is the variational parameters which are optimised to find the best solution.
- The typical measure of similarity is the KL divergence.

$$\begin{aligned} KL(q,p) &= \int \log \frac{p(\theta|X)}{q(\theta|\lambda)} q(\theta|\lambda) \mathrm{d}\theta \\ &= \mathbb{E}_q[\log p(\theta, X)] - \mathbb{E}_q[\log q(\theta|\lambda)] + \log p(X) \end{aligned}$$

- Since $p(X)$ is a constant with respect to λ optimising the KL is equivalent to optimising the evidence lower bound (ELBO)

$$\mathcal{L} := \mathbb{E}_q[\log p(\theta, X)] - \mathbb{E}_q[\log q(\theta|\lambda)]$$

Mean field variational inference

- Historically, mean field variational inference (MFVI) often called variational Bayes was the most common approach.
- MFVI assumes the approximating distribution has a factorised form

$$q(\theta|\lambda) = \prod_i q_i(\theta_i|\lambda_i)$$

- The factorisation is part of the approximation, and has nothing to do with the model structure.
- If the model is in the conjugate exponential family then it is possible to obtain an algorithm similar to EM with closed form updates.
  - In general this is not possible because $\mathbb{E}_q[\log p(\theta, X)]$ is hard to evaluate
- The main weakness of MFVI is the approximate posterior cannot capture correlation between parameters.

Monte Carlo and VI

- To optimise the ELBO we need to be able to compute

$$\nabla_\lambda \mathcal{L} \quad = \quad \nabla_\lambda [\mathbb{E}_q[\log p(\theta, X)] - \mathbb{E}_q[\log q(\theta|\lambda)]]$$

- One line of recent work has been figuring out how to make this computation feasible in general.
  - One strategy is to reparameterise the variational distribution so the objective is differentiable.
  - Another strategy is to move the gradient inside the expectation.
- The main insight of the new approaches is that we can use Monte Carlo methods to approximate expectations.
- Applying these methods, a much broader class of variational distributions can used.
  - This leads to better approximation of the posterior.

Stochastic VI

- The other major highlight of recent VI research has been the observation you can sub-sample data.
- The key idea is that we do not need to exactly evaluate the gradient of the ELBO.
- Instead we can use unbiased estimates to perform stochastic gradient descent.
- By using sub-sampling VI can efficiently scale to extremely large problems.

Advance Monte Carlo Methods

Motivation

- MH can struggle on high dimensional problems.
- The blocking procedure helps, but works poorly if the parameters are highly correlated between blocks.
- MH can also be slow to explore multi-modal posteriors or get trapped in local optima.

Simulated annealing

- Simulated annealing (SA) addresses the problem of local optima.
- The basic idea is to introduce a sequence of distributions such as

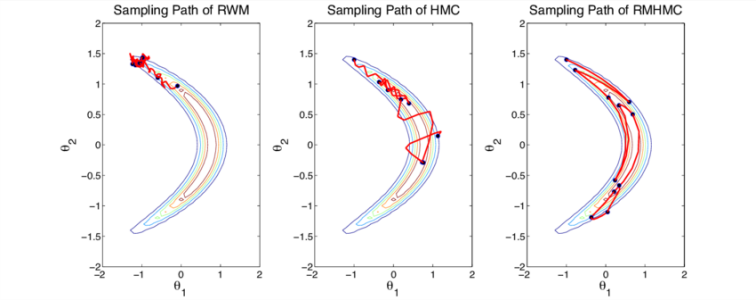
$$p_\beta(\theta|X) \quad = \quad p(X|\theta)^\beta p(\theta)$$

  with $0 = \beta_1 < \beta_2 < \ldots < \beta_T = 1$.
- The hope is we start with a distribution that is easy to sample from, and the slowly change the distribution until we arrive at the one of interest.
- We start with β_1 and use any MCMC update for the parameters. After several iterations we then update to β_2 and so on.
- We collect samples when we reach β_T.

Hamiltonian Monte Carlo

- HMC is an efficient way to update high dimensional continuous parameters.
- The basic idea is to propose a new value for the MCMC sampler by deterministically moving along a path defined by Hamiltonian dynamics.
- Traditionally tuning the algorithm parameters has been hard. Recent work has provided automated tuning solutions that seem to work well.
- The probabilistic programming language STAN uses HMC and automatic differentiation to efficiently sample from a large class of user defined models.

HMC in action

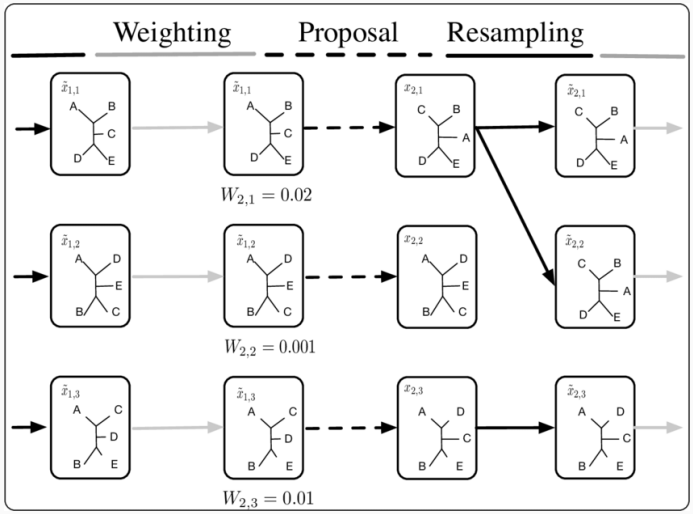


Lan et al. Journal Journal of Computational and Graphical Statistics 2015

Sequential Monte Carlo

- SMC is a very general algorithm for sampling from high dimensional distributions.
- The basic idea is to break the sampling problem down into smaller and easier to sample problems.
- SMC uses a collection of particles to approximate the posterior.
- The particles are iteratively update using local moves and resampling.
- Traditionally SMC has been used for models with a natural sequential structure.
- Recent work has shown that it can be applied more generally for problems such as phylogenetics.

SMC for phylogenetics

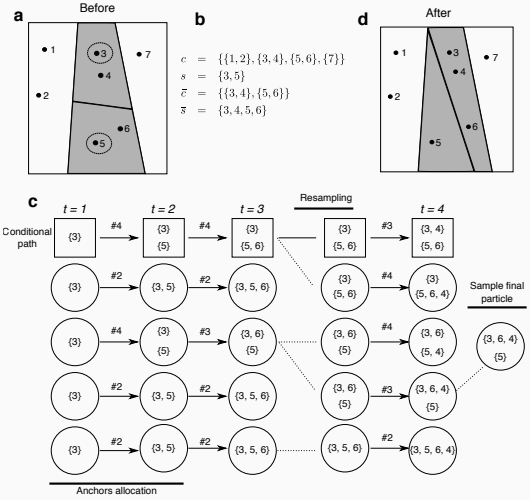


https://arxiv.org/abs/1806.08813

Particle Gibbs

- The basic idea of PG is to use SMC to perform block updates of a subset of parameters in a larger MCMC framework.
- One example would be to use SMC update the hidden states of an HMM. MH or other moves could then be used to update the other parameters.
- The obvious way to implement this is incorrect. So PG relies on conditional SMC to produce a valid algorithm.
- Like SMC, PG was originally applied to models with sequential structure.
- Recent work has show it is more general and can be applied to problems such as Bayesian mixture models.

Particle Gibbs for mixture models

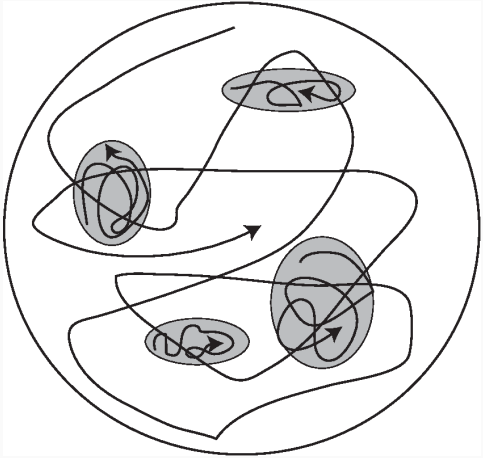


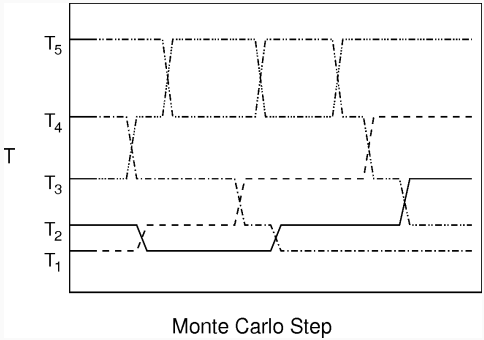
Bouchard-Côté et al. JMLR 2017

Parallel tempering

- PT is a general approach to sample from multi-modal or or hard to sample distributions.
- Like SA we define a sequence of distributions.
- The difference is we sample from all distributions simultaneously.
- The chains interact by swapping parameters occasionally.
  - This allows values from the easy to sample from chain move to the chain of interest.

PT mode hopping



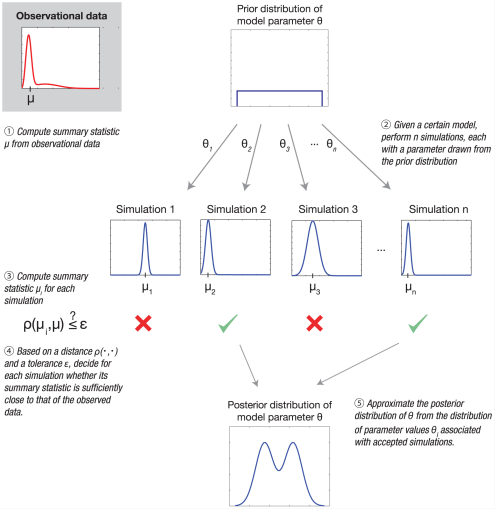


Earl et al. Physical chemistry chemical physics 2005

Approximate Bayesian computation

- ABC has emerged as an extremely general method for computing posterior approximations.
- Unlike the other approaches, ABC does require evaluation of the likelihood.
- This can be extremely useful in population genetics models where it is easy to simulate from the model, but hard to evaluate the data probability.
- The basic idea of ABC is to propose parameters from the prior, and then simulate data with these parameters. If the simulated data is close enough to the observed data then we add the parameters to the posterior approximation.

ABC



Sunnåker et al. PLOS Computational Biology 2013