Markov Chain Monte Carlo and Variational Inference: Bridging the Gap, by Salimans, Kingma, Welling (2014)

Andi Wang

NN Reading group

28 April 2020

Introduction

This paper seeks to use Markov chain Monte Carlo (MCMC) to construct *auxiliary random variables*, which are used within a (stochastic) *variational inference* (VI) framework to perform *approximate posterior inference*.

Introduction

This paper seeks to use Markov chain Monte Carlo (MCMC) to construct *auxiliary random variables*, which are used within a (stochastic) *variational inference* (VI) framework to perform *approximate posterior inference*.

The hope is that this fusion can combine the fast, and explicit optimization of VI with the (asymptotically) exact and flexible nature of MCMC.

Notation

The paper uses notation which is typical for ML and SMC, which is ambiguous from a probabalistic point of view: the letter p denotes the true distribution of 'everything' and it is the argument which dictates which distribution we look at.

Notation

The paper uses notation which is typical for ML and SMC, which is ambiguous from a probabalistic point of view: the letter p denotes the true distribution of 'everything' and it is the argument which dictates which distribution we look at.

- Latent variables of interest are z, prior is $p(z)$.
- Observed data x, likelihood is $p(x|z)$, marginal likelihood is $p(x)$, true posterior is $p(z|x)$.
- Letter $q_\theta(z|x)$ denotes family of approximate variational posteriors, which are parameterized by θ.

Variational Inference

We wish to approximate the true posterior $p(z|x)$ using some 'nice' family of distributions

$$\{q_\theta(z|x) : \theta \in \Theta\}.$$

Variational Inference

We wish to approximate the true posterior $p(z|x)$ using some 'nice' family of distributions

$$\{q_\theta(z|x) : \theta \in \Theta\}.$$

E.g. Gaussian approximation: $q_\theta(z|x)$ could all be Gaussians with mean/covariance based on data x, parameterized by θ.

Variational Inference

We wish to approximate the true posterior $p(z|x)$ using some 'nice' family of distributions

$$\{q_\theta(z|x) : \theta \in \Theta\}.$$

E.g. Gaussian approximation: $q_\theta(z|x)$ could all be Gaussians with mean/covariance based on data x, parameterized by θ.

To find a good approximation, we try to maximise the evidence lower bound (ELBO) $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}(\theta) := \mathbb{E}_q[\log p(x,z) - \log q_\theta(z|x)] = \log p(x) - D_{KL}\{q_\theta(z|x) || p(z|x)\}.$$

Variational Inference

We wish to approximate the true posterior $p(z|x)$ using some 'nice' family of distributions

$$\{q_\theta(z|x) : \theta \in \Theta\}.$$

E.g. Gaussian approximation: $q_\theta(z|x)$ could all be Gaussians with mean/covariance based on data x, parameterized by θ.

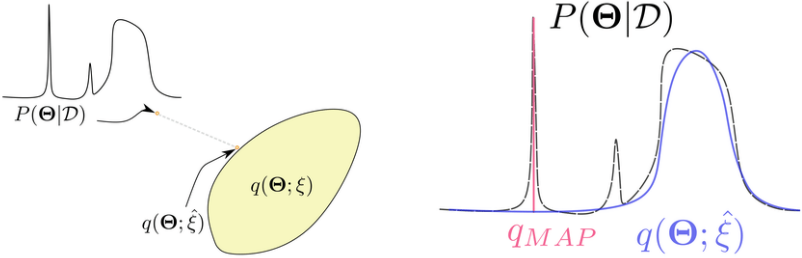
To find a good approximation, we try to maximise the evidence lower bound (ELBO) $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}(\theta) := \mathbb{E}_q[\log p(x, z) - \log q_\theta(z|x)] = \log p(x) - D_{KL}\{q_\theta(z|x)||p(z|x)\}.$$

This is maximised over θ using standard optimization methods.

Variational Inference Cartoon

https://www.researchgate.net/figure/
Left-illustration-of-variational-inference-Right-difference-between-MAP-and
fig3_333678861



VI overview

- Have an explicit objective $\mathcal{L}$.
- Generally fast.
- Conceptually simple deterministic scheme.

VI overview

- Have an explicit objective $\mathcal{L}$.
- Generally fast.
- Conceptually simple deterministic scheme.
- Very parametric: performance depends crucially on choosing a good family $q_\theta(z|x)$.
- Always inexact (no matter how long you run for).

MCMC

Construct a Markov chain $(z_t)_{t=1}^{\infty}$, with transition kernel $q(z_t | z_{t-1}, x)$, whose invariant distribution coincides with the posterior $p(z|x)$.

Then (under conditions) for a large value of T, z_T is (approximately) distributed according to true posterior $p(z|x)$.

MCMC overview

- Asymptotically exact.
- 'Less parametric' than VI (but need to choose kernel sensibly).

MCMC overview

- Asymptotically exact.
- 'Less parametric' than VI (but need to choose kernel sensibly).
- Issues of burn-in; no longer an explicit objective.
- Stochastic algorithm.
- Can be slow to converge.
- It can be difficult to tune the parameters within the algorithm.

VI with auxiliary random variables

Vanilla VI is a deterministic optimization scheme for approximating a posterior distribution.

VI with auxiliary random variables

Vanilla VI is a deterministic optimization scheme for approximating a posterior distribution.

But it is also possible include auxiliary random variables. Suppose we have obtained from MCMC a chain $z_0, \ldots, z_{T-1}, z_T$:

$$q_\theta(z_0, \ldots, z_{T-1}, z_T | x) = q_\theta(z_0 | x) \prod_{t=1}^{T} q_\theta(z_t | z_{t-1}, x).$$

VI with auxiliary random variables

Vanilla VI is a deterministic optimization scheme for approximating a posterior distribution.

But it is also possible include auxiliary random variables. Suppose we have obtained from MCMC a chain $z_0, \ldots, z_{T-1}, z_T$:

$$q_\theta(z_0, \ldots, z_{T-1}, z_T | x) = q_\theta(z_0 | x) \prod_{t=1}^{T} q_\theta(z_t | z_{t-1}, x).$$

We will take as our marginal posterior approximation

$$q_\theta(z_T | x) = \int q_\theta(z_0, \ldots, z_{T-1}, z_T | x) dz_0 \ldots dz_{T-1}.$$

In other words, we see $y := (z_0, \ldots, z_{T-1})$ as a collection of auxiliary random variables. (Rich family of approximate posteriors!)

New objective

Recall $y := (z_0, \ldots, z_{T-1})$. We obtain then a new variational lower bound to optimize:

$$\mathcal{L}_{\text{aux}} := \mathcal{L} - \mathbb{E}_{q(z_T|x)}[D_{KL}\{q(y|z_T, x)||r(y|z_T, x)\}],$$

where $r(y|z_T, x)$ is an auxiliary inference distribution which you can choose freely.

New objective

Recall $y := (z_0, \ldots, z_{T-1})$. We obtain then a new variational lower bound to optimize:

$$\mathcal{L}_{\text{aux}} := \mathcal{L} - \mathbb{E}_{q(z_T|x)}[D_{KL}\{q(y|z_T, x)||r(y|z_T, x)\}],$$

where $r(y|z_T, x)$ is an auxiliary inference distribution which you can choose freely.

In this work they consider $r(y|z_T, x)$ which also has a Markov structure:

$$r(z_0, \ldots, z_{T-1}|x, z_T) = \prod_{t=1}^{T} r_t(z_{t-1}|x, z_t).$$

New objective

Recall $y := (z_0, \dots, z_{T-1})$. We obtain then a new variational lower bound to optimize:

$$\mathcal{L}_{\text{aux}} := \mathcal{L} - \mathbb{E}_{q(z_T|x)}[D_{KL}\{q(y|z_T, x)||r(y|z_T, x)\}],$$

where $r(y|z_T, x)$ is an auxiliary inference distribution which you can choose freely.

In this work they consider $r(y|z_T, x)$ which also has a Markov structure:

$$r(z_0, \dots, z_{T-1}|x, z_T) = \prod_{t=1}^{T} r_t(z_{t-1}|x, z_t).$$

Then with this choice, we have

$$\mathcal{L}_{\text{aux}} = \mathbb{E}_q[\log(p(x, z_T)/q(z_0, x))] + \sum_{t=1}^{T} \log[r_t(z_{t-1}|x, z_t)/q_t(z_t|x, z_{t-1})].$$

New Objective ctd

With auxiliary random variables, now have to perform stochastic gradient variational inference.

New Objective ctd

With auxiliary random variables, now have to perform stochastic gradient variational inference.

It is not difficult to obtain unbiased estimates L of the lower bound $\mathcal{L}_{\mathsf{aux}}$ (Algorithm 1), and so its derivative is an unbiased estimate of the derivative of the lower bound*.

New Objective ctd

With auxiliary random variables, now have to perform stochastic gradient variational inference.

It is not difficult to obtain unbiased estimates L of the lower bound $\mathcal{L}_{\mathrm{aux}}$ (Algorithm 1), and so its derivative is an unbiased estimate of the derivative of the lower bound*.

Thus can use this in a stochastic optimization algorithm (Algorithm 2).

Some things learned...

Some things learned...

- VI can handle auxiliary random variables. The algorithm becomes stochastic optimization, but this is still possible.
- These auxiliary random variables can come from MCMC.
- MCMC itself can provide a rich family of approximate posterior distributions.
- VI can be used in a sense to 'tune' HMC.