Pattern of Scalable Bayesian Inference

Manuel A. Vázquez

Machine Learning Group
Universidad Carlos III de Madrid

June 28, 2017

Índice

Problem statement

Goal

Estimating the posterior distribution of a parameter of interest, θ, given a collection of conditionally independent observations, $\mathbf{x} = \{x_n\}_{n=1}^N$.

...i.e., we aim at approximating

$$p(\theta|\mathbf{x})$$

$\theta \equiv$ parameter (possibly a vector) to be estimated

$\{x_n\}_{n=1}^N = \mathbf{x} \equiv$ available data (observations)

Markov Chain Monte Carlo (MCMC)

MCMC

It aims at simulating a **Markov chain**, $\theta_1, \theta_2, \theta_3 \cdots$, that admits as its *stationary* distribution the posterior distribution of interest. This means that, after some *burn-in* period encompassing, say, M samples we have

$$\theta_{M+j} \sim p(\theta|\mathbf{x}), j > 0$$

Markov property: at iteration t, a sample, θ_t, is drawn conditional on the previous one θ_{t-1}. It usually involves evaluating the joint distribution, $p(\theta, \mathbf{x})$, or the likelihood, $p(\mathbf{x}|\theta)$.

Problem

If the dataset is very large (say, $\mathbf{x}$ has $N = 10^6$ elements), evaluating $p(\theta, \mathbf{x})$ or $p(\mathbf{x}|\theta)$ at each iteration might be very expensive.

Dealing with large datasets

Key idea

at each iteration, operating only on **subsets of data**

We are usually interested in the posterior density, that can be factorized using Bayes rule

$$\pi(\theta \,|\, \mathbf{x}) \propto \pi_0(\theta)\pi(\mathbf{x} \,|\, \theta).$$

Additionally, if the data $\mathbf{x} = \{x_n\}_{n=1}^N$ are **conditionally independent** given the model parameters θ, we have

$$\pi(\theta \,|\, \mathbf{x}) \propto \pi_0(\theta) \underbrace{\prod_{n=1}^{N} \pi(x_n \,|\, \theta)}_{\pi(\mathbf{x} \,|\, \theta)}.$$

For large N this factorization can be exploited to construct MCMC algorithms in which the updates depend only on subsets of data.

Approximating the log likelihood

Since the likelihood is a product, the log likelihood is a sum,

$$\log \pi(\mathbf{x} \mid \theta) = \sum_{n=1}^{N} \log \pi(x_n \mid \theta),$$

This can be approximated using a random subset of $m < N$ terms as

$$\log \pi(\mathbf{x} \mid \theta) \approx \frac{N}{m} \sum_{n=1}^{m} \log \pi(x_n^* \mid \theta),$$

where $\{x_n^*\}_{n=1}^m$ is a uniformly random subset of $\{x_n\}_{n=1}^N$.
This is an **unbiased** estimator of the log likelihood that allows obtaining an unbiased estimate of the posterior (up to a proportionality constant)

$$\log \pi_0(\theta)\pi(\mathbf{x} \mid \theta) \approx \log \pi_0(\theta) + \frac{N}{m} \sum_{n=1}^{m} \log \pi(x_n^* \mid \theta).$$

Metropolis-Hastings

Metropolis-Hastings

...is just **one** (among many) **way of building a Markov chain** tailored for a posterior distribution of interest

From now on, let

$$\theta = \theta_{t-1} \text{ (last sample in the chain)}$$
$$\theta' = \theta_t \text{ (new sample)}$$

Then, a new sample, θ', is drawn from a proposal distribution, $q(\theta' \mid \theta)$, conditional on the last sample, θ, and accepted with probability

$$\alpha(\theta, \theta') = \min\left(1, \frac{p(\theta', \mathbf{x}) q(\theta \mid \theta')}{p(\theta, \mathbf{x}) q(\theta' \mid \theta)}\right).$$

Notice the above expression entails evaluating (usually through the likelihood) the joint distribution, $p(\theta', \mathbf{x})$.

Adaptive subsampling for Metropolis-Hastings

The above acceptance condition yields the

MH test

$$\frac{\pi(\theta' \mid \mathbf{x}) q(\theta \mid \theta')}{\pi(\theta \mid \mathbf{x}) q(\theta' \mid \theta)} > u,$$

where $u \sim \text{Unif}(0, 1)$.

Key idea

to approximate the MH test using a **subset** of the full dataset.

we make up for this by

- modeling the probability that the outcome of such an *approximate* MH test differs from the exact MH test
- setting some *tolerance* parameter on the above metric that allows checking how good we are doing, so that the size of the subset is increased if required

Rewriting the Metropolis-Hastings test

Back to the MH test, rearranging and using log probabilities yields

$$\log\left[\frac{\pi(\mathbf{x}\,|\,\theta')}{\pi(\mathbf{x}\,|\,\theta)}\right] > \log\left[u\frac{q(\theta'\,|\,\theta)\pi_0(\theta)}{q(\theta\,|\,\theta')\pi_0(\theta')}\right],$$

taking advantage of the conditional independence of the data,

$$\sum_{n=1}^{N}\log\left[\frac{\pi(x_n\,|\,\theta')}{\pi(x_n\,|\,\theta)}\right] > \log\left[u\frac{q(\theta'\,|\,\theta)\pi_0(\theta)}{q(\theta\,|\,\theta')\pi_0(\theta')}\right],$$

and dividing both sides by N gives an equivalent threshold,

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}\underbrace{\log\left[\frac{\pi(x_n\,|\,\theta')}{\pi(x_n\,|\,\theta)}\right]}_{\ell_n}}_{\Lambda(\theta,\theta')} > \underbrace{\frac{1}{N}\log\left[u\frac{q(\theta'\,|\,\theta)\pi_0(\theta)}{q(\theta\,|\,\theta')\pi_0(\theta')}\right]}_{\psi(u,\theta,\theta')},$$

and $\Lambda(\theta,\theta') = \frac{1}{N}\sum_{n=1}^{N}\ell_n$ with $\ell_n = \log\pi(x_n\,|\,\theta') - \log\pi(x_n\,|\,\theta)$

Adaptive subsampling for Metropolis-Hastings

Then, the MH test amounts to

$$\Lambda(\theta, \theta') > \psi(u, \theta, \theta') \tag{1}$$

where $\psi(u, \theta, \theta')$ does **not** depend on the data and

$$\Lambda(\theta, \theta') = \frac{1}{N} \sum_{n=1}^{N} \ell_n \quad \text{(average of the log likelihood ratio)}$$

We can approximate $\Lambda(\theta, \theta')$ by averaging m (instead of N) log likelihood ratios $\{\ell_n^*\}_{n=1}^m$,

$$\hat{\Lambda}_m(\theta, \theta') = \frac{1}{m} \sum_{n=1}^{m} \ell_n^* = \frac{1}{m} \sum_{n=1}^{m} \log \left[\frac{\pi(x_n^* \mid \theta')}{\pi(x_n^* \mid \theta)} \right].$$

This *unbiased* estimate of $\Lambda(\theta, \theta')$ yields the approximate MH test

$$\hat{\Lambda}_m(\theta, \theta') > \psi(u, \theta, \theta'), \tag{2}$$

Approximate MH with adaptive stopping rule

There is a mismatch between the true MH test in (1), and the approximation in (2).

Key idea

- model the error of the approximation
- incrementally read more data until an **adaptive stopping rule** informs us that our error is less than some user-specified tolerance

The error model should provide a way to approximate or bound the probability that the approximate outcome disagrees with the full-data outcome:

$$\mathbb{P}\left[\underbrace{((\hat{\Lambda}_m(\theta,\theta') > \psi(u,\theta,\theta'))}_{\text{approximate test}} \neq \underbrace{((\Lambda(\theta,\theta') > \psi(u,\theta,\theta'))}_{\textit{real}\text{ test}}\right].$$

t-statistic for stopping rule

Hypothesis test

$$\Lambda(\theta, \theta') = \psi(u, \theta, \theta')$$

Let us assume the log likelihood ratios are Gaussian, i.e.,

$$\ell_n \sim \mathcal{N}(\mu, \sigma^2) \ .$$

An estimate of its mean based on a subset of size m, is

$$\hat{\mu}_m = \hat{\Lambda}_m(\theta, \theta') = \frac{1}{m} \sum_{n=1}^{m} \ell_n^*,$$

and the empirical standard deviation is given by

$$s_m = \sqrt{\frac{m}{m-1} \left(\hat{\Lambda}_m^2(\theta, \theta') - \hat{\Lambda}_m(\theta, \theta')^2 \right)},$$

where $\hat{\Lambda}_m^2(\theta, \theta') = \frac{1}{m} \sum_{n=1}^{m} (\ell_n^*)^2$.

t-statistic for stopping rule

The standard error of $\hat{\mu}_m$ is $\frac{s_m}{\sqrt{m}}$ and since

- samples are **not** selected from a infinite population,
- **nor** are they selected with replacement

a *finite population correction* is applied to get

$$\hat{\sigma}_m = \frac{s_m}{\sqrt{m}} \sqrt{\frac{N-m}{N-1}}.$$

Then, if m is large enough for the CLT to hold, the test statistic

$$t = \frac{\hat{\Lambda}_m(\theta, \theta') - \psi(u, \theta, \theta')}{\hat{\sigma}_m}$$

follows a Student's t-distribution with $m-1$ degrees of freedom when $\Lambda(\theta, \theta') = \psi(u, \theta, \theta')$.

t-statistic for stopping rule

Then, the probability of the approximate and actual outcomes agree is

$$\rho = 1 - \phi_{m-1}(|t|)$$

is the probability that they disagree, where $\phi_{m-1}(\cdot)$ is the CDF of the Student's t-distribution with $m-1$ degrees of freedom.

Adaptive stopping rule

For any user-provided tolerance $\epsilon \geq 0$, we can incrementally increase m until $\rho \leq \epsilon$.

Concentration inequalities

It provides

- bound on the error of the approximate acceptance probability
- bound on *total variation* (TV) between the approximate and true stationary distributions

$$\|p - q\|_{\mathrm{TV}} = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \, dx.$$

Just like before, an estimate, $\hat{\Lambda}_m(\theta, \theta')$, of the average log likelihood ratio is computed...**but** now, in order to characterize its quality

$$\mathbb{P}\left(\left| \hat{\Lambda}_m(\theta, \theta') - \Lambda(\theta, \theta') \right| \leq c_m \right) \geq 1 - \delta_m$$

for $\delta_m \in (0, 1)$ and some constant c_m.

Concentration inequalities: bounds

Several choices for the bound c_m

Hoeffding's inequality

$$c_m = C_{\theta,\theta'} \sqrt{\frac{2}{m}\left(1 - \frac{m-1}{N}\right) \log\left(\frac{2}{\delta_m}\right)}$$

Empirical Bernstein bound

$$c_m = s_m \sqrt{\frac{2\log(3/\delta_m)}{m}} + \frac{6 C_{\theta,\theta'} \log(3/\delta_m)}{m},$$

with

$$C_{\theta,\theta'} = \max_{1 \le n \le N} \left| \log \pi(x_n \,|\, \theta') - \log \pi(x_n \,|\, \theta) \right| = \max_{1 \le n \le N} |\ell_n|.$$

Problem

$C_{\theta,\theta'}$ requires computing every ℓ_n!! It must be provided as an input to the algorithm

Concentration inequalities: key idea

$$\mathbb{P}\left(\left|\hat{\Lambda}_m(\theta,\theta') - \Lambda(\theta,\theta')\right| \leq c_m\right) \geq 1 - \delta_m$$

means that, with probability $1-\delta_m$,

- $\Lambda(\theta,\theta')$ is **less than** c_m units away from $\hat{\Lambda}_m(\theta,\theta')$

If, additionally,

- $\psi(u,\theta,\theta')$ **is more** than c_m units way from $\hat{\Lambda}_m(\theta,\theta')$,

$$|\hat{\Lambda}_m(\theta,\theta') - \psi(u,\theta,\theta')| > c_m$$

then, the decision threshold, $\psi(u,\theta,\theta')$, is never between the approximate, $\hat{\Lambda}_m(\theta,\theta')$, and true, $\Lambda(\theta,\theta')$, values of the statistic $\Rightarrow$ the decisions match with probability $1-\delta_m$

Adaptive stopping rule

If $|\hat{\Lambda}_m(\theta,\theta') - \psi(u,\theta,\theta')| \leq c_m$, then we want to increase m until this is no longer the case.

Concentration inequalities: bound on the error of the approximate acceptance probability

We can choose a tolerance parameter, ϵ, and set δ_m in a certain way to achieve a global bound. Let $p > 1$ and set

$$\delta_m = \frac{p-1}{pm^p}\epsilon, \quad \text{thus} \quad \sum_{m\geq 1}\delta_m \leq \epsilon.$$

Then, the probability that, for every m (choice of the subset size), the first condition above holds is

$$\mathbb{P}\left(\bigcap_{m\geq 1}\left\{\left|\hat{\Lambda}_m(\theta,\theta') - \Lambda(\theta,\theta')\right| \leq c_m\right\}\right) = 1 - \mathbb{P}\left(\bigcup_{m\geq 1}\left\{\left|\hat{\Lambda}_m(\theta,\theta') - \Lambda(\theta,\theta')\right| > c_m\right\}\right)$$

$$= 1 - \underbrace{\sum_{m\geq 1}\delta_m}_{\leq \epsilon} \geq 1 - \epsilon$$

where it has been used that

$$\mathbb{P}\left(\left|\hat{\Lambda}_m(\theta,\theta') - \Lambda(\theta,\theta')\right| \leq c_m\right) \geq 1-\delta_m \Rightarrow \mathbb{P}\left(\left|\hat{\Lambda}_m(\theta,\theta') - \Lambda(\theta,\theta')\right| \geq c_m\right) \leq \delta_m$$

Error bounds on the stationary distribution

Theorem: Convergence

If <insert some ugly stuff here> then we have

$$\|\pi - \tilde{\pi}\|_{\mathsf{TV}} \leq \frac{Ah\mathcal{E}_{\max}}{1-\lambda}.$$

where A and h are ugly and

$$\mathcal{E}_{\max} = \sup_{\theta,\theta'} |\mathcal{E}(\theta,\theta')|$$

with

$$\mathcal{E}(\theta,\theta') = \tilde{\alpha}(\theta,\theta') - \alpha(\theta,\theta')$$

Subselecting via a lower bound on the likelihood

Firefly Monte Carlo (FlyMC)

Auxiliary variable MCMC sampling procedure that operates on only subsets of data in each iteration...

...by selecting what data to evaluate based on the random indicators (included in the Markov chain state, i.e. sampled).

- it generates samples from the **exact** target posterior 
- requires a lower bound on the likelihood (at every datum) 

FlyMC

Let

$$L_n(\theta) = p(x_n \mid \theta),$$

and $B_n(\theta)$ be a strictly positive lower bound on $L_n(\theta)$, i.e.,

$$0 < B_n(\theta) \leq L_n(\theta).$$

For each datum, a binary auxiliary r.v., $z_n \in \{0, 1\}$, with Bernoulli distribution is introduced

$$p(z_n \mid x_n, \theta) = \left[\underbrace{\frac{L_n(\theta) - B_n(\theta)}{L_n(\theta)}}_{p}\right]^{z_n} \left[\underbrace{\frac{B_n(\theta)}{L_n(\theta)}}_{1-p}\right]^{1-z_n}$$

(z_n are independent for different n (data))

FlyMC: computing the likelihood

The auxiliary variable is included in the posterior ($\Rightarrow$ it must be sampled!!)

$$\tilde{\pi}(\theta, \mathbf{z} \,|\, \mathbf{x}) = \pi(\theta \,|\, \mathbf{x}) p(\mathbf{z} \,|\, \mathbf{x}, \theta) \propto \pi_0(\theta) \prod_{n=1}^{N} \pi(x_n \,|\, \theta) p(z_n \,|\, x_n, \theta).$$

Using the equations on the previous slide

$$\begin{aligned}
\tilde{\pi}(\theta, \mathbf{z} \,|\, \mathbf{x}) &\propto \pi_0(\theta) \prod_{n=1}^{N} L_n(\theta) \left[\frac{L_n(\theta) - B_n(\theta)}{L_n(\theta)}\right]^{z_n} \left[\frac{B_n(\theta)}{L_n(\theta)}\right]^{1-z_n} \\
&= \pi_0(\theta) \prod_{n=1}^{N} (L_n(\theta) - B_n(\theta))^{z_n} B_n(\theta)^{1-z_n} \\
&= \pi_0(\theta) \underbrace{\prod_{n:z_n=1} (L_n(\theta) - B_n(\theta))}_{\text{data required}} \underbrace{\prod_{n:z_n=0} B_n(\theta)}_{\text{data } \textbf{not} \text{ required}}.
\end{aligned}$$

FlyMC: challenges

- constructing a collapsible lower bound that is tight
  - tight so that we only need to evaluate a small amount of data, and
  - collapsible so that the product of bounds for the rest of the data can be evaluated efficiently
- efficient implementation
  - the auxiliary variable might slow down the mixing of the Markov chain (samples might be more correlated)

Stochastic gradients of the log joint density

Inspired by stochastic optimization techniques...it relies on

- Gradient Descent

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\left(\nabla \log \pi(\theta_t, \mathbf{x})\right), \text{ but}$$

$$\nabla \log \pi(\theta_t, \mathbf{x}) = \nabla \log \pi_0(\theta_t) + \sum_{n=1}^{N} \nabla \log \pi(x_n \,|\, \theta_t)$$

depends on all the data... We can alleviate this by using

- Stochastic Gradient Descent (SGD)

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\left(\nabla \log \pi_0(\theta_t) + \frac{N}{m}\sum_{n=1}^{m} \nabla \log \pi(x_n \,|\, \theta_t)\right).$$

It converges to a local extreme if

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty.$$

Stochastic Gradient Langevin Dynamics (SGLD)

SGLD

Approximate MCMC (MH-based) procedure that combines SGD with a simple kind of Langevin dynamics

$$\theta' = \theta_t + \frac{\epsilon_t}{2}\left(\nabla \log \pi_0(\theta_t) + \frac{N}{m}\sum_{n=1}^{m} \nabla \log \pi(x_n \mid \theta_t)\right) + \eta_t,$$

with ϵ_t satisfying (**??**) and $\eta_t \sim \mathcal{N}(0, \epsilon_t)$. Now, we would be computing the acceptance probability (involving the full likelihood)

Key idea

As $\epsilon_t \to 0$, $\theta' \to \theta_t$ and the probability of acceptance converges to 1 $\Rightarrow$ every sample is accepted and there is no need to compute the probability of acceptance

Stochastic Gradient Langevin Dynamics (SGLD): challenges

As $\epsilon_t \to 0$, $\theta' \to \theta_t$ the **chain stops** completely!! Authors suggest that a trade-off can be found so that θ

- is large enough for efficient sampling, but
- small enough so that acceptance probability is *essentially* 1

$$\epsilon_t > \epsilon_\infty > 0$$

One problem remains, though...

Without the stochastic MH acceptance step, asymptotic samples are no longer guaranteed to represent the target distribution