Likelihood-free Inference and Optimization using Stochastic Gradient Approximations

Abstract

In this paper we apply recent techniques from Bayesian inference that use gradient information to sample efficiently from the true posterior distribution to the *likelihood-free* or *approximate Bayesian computation* (ABC) setting. To do this for ABC we adopt a *gradient-free* stochastic approximation algorithm by Spall [**?**]. Together these algorithms provide both optimization and inference for likelihood-free models as the algorithm ABC-SGLD transitions from optimization to sampling. We demonstrate ABC-SGLD on problems where the true gradient information is known and on challenging ABC simulators.

1 INTRODUCTION

- Arguably the two most useful procedures in simulation-based science are optimization and Bayesian inference.
- Optimization can take the form of simple grid-search to sophisticated techniques like factorial design [**?**], Bayesian experiment design [**?**], (mention others).
- Recently, Bayesian optimization techniques have shown success in optimizing very expensive black-box simulators [**?**].
- The main approach of Bayesian inference of simulator parameters is ABC. ABC is largely based a few sampling algorithms: SMC and MCMC.
- Though gradients are not directly computable for simulators, they can be computed analytically by using finite differences (see [**?**]). This quickly becomes infeasible for large p problems. There is however an alternative stochastic approximation algorithm by Spall [**?**] that requires only 2 simulation calls independent of p.
- By using this approximation, gradient-based algorithms can be adopted for simulators: for optimization, using analogous stochastic gradient descent algorithms and for Bayesian inference using Langevin dynamics [**?**].
- Recently, the SGLD [**?**] algorithm has efficiently combined optimization with inference using ideas from SGD with Langevin dynamics.
- This paper describes ABC-SGLD.

2 GRADIENTS FROM FORWARD SIMULATIONS

2.1 Robbin's Monro: SGD

2.2 Spall's method: SPSA

In the gradient-free setting, Spall [**?**] provides a stochastic approximate to the true gradient using only 2 forward simulations (function evaluations). This is in contrast to multivariate finite-difference stochastic approximation FDSA [**?**] requiring $2p$ evaluations.

The gradient estimate is

$$\hat{g}_t\left(\boldsymbol{\theta}_t\right) = \begin{bmatrix} \frac{y_t^+ - y_t^-}{2c_{t1}\Delta_{t1}} \\ \vdots \\ \frac{y_t^+ - y_t^-}{2c_{tp}\Delta_{tp}} \end{bmatrix} \tag{1}$$

where c_{tp} is a step-size that is usually constant for all dimensions p, but can be different (as shown in this case); $\Delta_{tp} \in -1, +1$ is a *perturbation mask* (called symmetric Bernouilli variables by Spall), i.e. $\Delta_{tp} \sim 2 * \text{Bernouilli}(0.5) - 1$; and $y_t^{\pm}$ are function evaluations:

$$y_t^+ = L\left(\boldsymbol{\theta} + c_t\Delta_t\right) \tag{2}$$

$$y_t^- = L\left(\boldsymbol{\theta} - c_t\Delta_t\right) \tag{3}$$

A way of reducing the noise in the gradient is by averaging

over q draws of Δ_t:

$$\hat{g}_t(\boldsymbol{\theta}_t) = \frac{1}{q}\sum_{j=1}^{q}\hat{g}_t^j(\boldsymbol{\theta}_t) \tag{4}$$

where for each j new perturbation masks are drawn.

Another variance reduction technique called the method of *common random numbers* (CRN) [**?**] sets a common seed for the random number generator (RNG) for both calls to the simulator. This technique can remove the effect of the simulator noise in the gradient estimate, leaving on the randomness of the perturbation masks as the source of noise. Consider using SPSA instead of SGD: the CRN technique is equivalent of using the same batch of data vectors to evaluate the log-likelihood which is a sensible approach.

2.2.1 Variations

- Varying c_q: use a different per repeat. Seems to converge faster. What is the correct noise process? Right now trying this: flip coin, if heads perturb $c* = 1 + U(0,1)$, else perturb $c/ = 1 + U(0,1)$.

3 LANGEVIN DYNAMICS

4 STOCHASTIC-GRADIENT LD

$$\begin{aligned}
V_h &= \text{Var}\left(\frac{N}{n}\sum_{i=1}^{n}\nabla\log p(\mathbf{x}_i|\boldsymbol{\theta}) - \sum_{i=1}^{N}\nabla\log p(\mathbf{x}_i|\boldsymbol{\theta})\right) \quad (5)\\
h &= \frac{N}{n}\sum_{i=1}^{n}s_i \quad (6)\\
s_i &= \nabla\log p(\mathbf{x}_i|\boldsymbol{\theta}) + \frac{1}{N}\nabla p(\boldsymbol{\theta}) \quad (7)\\
\text{Var}(h) &= \frac{N^2}{n^2}\sum_{i=1}^{n}\text{Var}(s_i) \quad (8)\\
&= \frac{N^2}{n^2}\sum_{i=1}^{n}\frac{1}{n}\left(\sum_{j=1}(s_i-\bar{s})^2\right) \quad (9)\\
&= \frac{N^2}{n^2}\sum_{i=1}^{n}V_s \quad (10)\\
&= \frac{N^2}{n^2}nV_s \quad (11)\\
&= \frac{N^2}{n}V_s \quad (12)\\
\text{Var}(\boldsymbol{\theta}) &= \frac{\epsilon^2}{4}\text{Var}(h) + \epsilon \quad (13)\\
\frac{\epsilon^2}{4}\text{Var}(h) &<< \epsilon \quad (14)\\
\frac{\epsilon N^2}{4n}V_s &<< 1 \quad (15)
\end{aligned}$$

5 SGLD-ABC

$$\begin{aligned}
V_{\hat{g}_t^j(\boldsymbol{\theta}_t)} &= \frac{1}{q-1}\sum_{j=1}^{q}(\hat{g}_t^j(\boldsymbol{\theta}_t) - \hat{g}_t(\boldsymbol{\theta}_t))^2 \quad (16)\\
V_{\hat{g}_t(\boldsymbol{\theta}_t)} &= \frac{1}{q-1}V_{\hat{g}_t^j(\boldsymbol{\theta}_t)} \quad (17)\\
\frac{\epsilon^2}{4}V_{\hat{g}_t(\boldsymbol{\theta}_t)} &= \frac{\epsilon^2}{4q}V_{\hat{g}_t^j(\boldsymbol{\theta}_t)} \quad (18)\\
&<< \epsilon \quad (19)\\
\frac{\epsilon}{4(q-1)}V_{\hat{g}_t^j(\boldsymbol{\theta}_t)} &<< 1 \quad (20)
\end{aligned}$$

Check for $g(\boldsymbol{\theta}) + h(\boldsymbol{\theta}) + \eta$ that $h(\boldsymbol{\theta}) + \eta \approx \eta$

Also $\theta_{t+1} = \theta_t + \frac{\epsilon}{2}\hat{g}_t(\boldsymbol{\theta}_t) + \sqrt{\epsilon}\mathcal{N}(0,1)$

5.1 Statistical Tests for Gradients

From Byrd:

$$\begin{aligned}
\frac{||V_{\hat{g}_t^j(\boldsymbol{\theta}_t)}||_1}{q} &\leq r^2||\hat{g}_t||_2^2 \quad (21)\\
\hat{q} &= \frac{||V_{\hat{g}_t^j(\boldsymbol{\theta}_t)}||_1}{r^2||\hat{g}_t||_2^2} \quad (22)
\end{aligned}$$

TODO:

- Check that q gradients are normal. Especially as q increases.
- Merge Byrd's test for $\hat{q}$ and condition on SGLD-ABC.

6 EXPERIMENTS

6.1 Logistic Regression with Stochastic Gradients

6.2 Likelihood-free Inference of Blowfly Dynamics

6.3 Other Useful ABC problem

7 DISCUSSION

8 CONCLUSION

References