Optimisation with SMC Samplers

Adam M. Johansen

adam.johansen@bristol.ac.uk - University of Bristol Statistics Group

Bridging the Gaps: Probabilistic Models for Optimisation,
28th November, 2007

Outline

- Optimisation, Rare Events and Approximate Counting.
- Sequential Monte Carlo Samplers
- SMC for Optimisation
- Some Open Problems

Optimisation, Rare Events and Counting

Problem Formulation

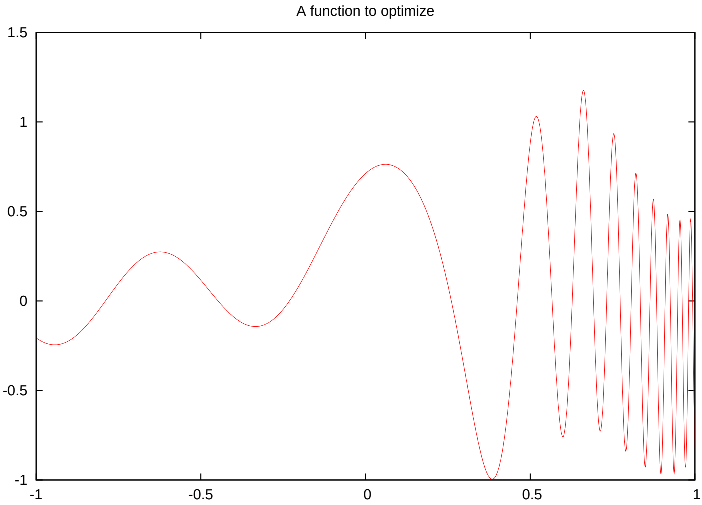
- Given some function $\varphi : E \to \mathbb{R}$ find

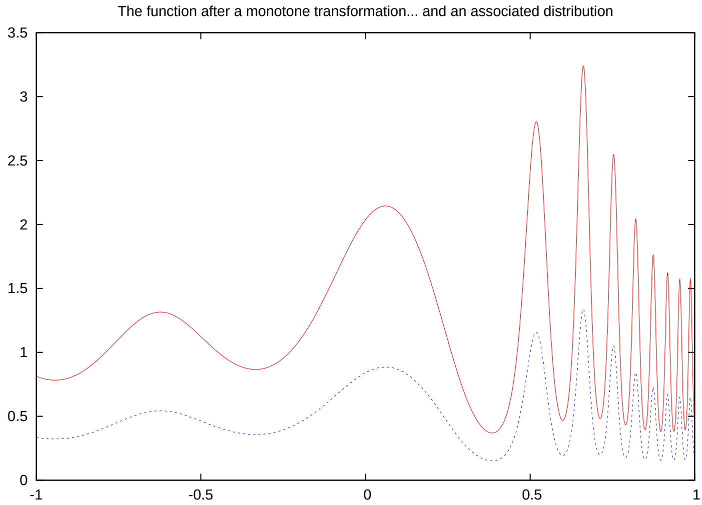
$$e^\star := \arg\max \varphi(e)$$
$$\text{or } E^\star := \left\{e \in E : \varphi(e) \geq \varphi(e')\forall e' \in E\right\}$$

- If $\Phi : \mathbb{R} \to \mathbb{R}$ is any increasing monotone function, then:

$$e^\star = \arg\max \Phi(\varphi(e))$$
$$\text{or } E^\star = \left\{e \in E : \Phi(\varphi(e)) \geq \Phi(\varphi(e'))\forall e' \in E\right\}$$

- We will assume that
  - $\varphi \geq 0$
  - There exists a σ-finite measure, μ, such that φ has finite exponential moments under μ, which does not vanish on $B(e^\star, \epsilon)$ for any $\epsilon > 0$.


A function to optimize
1.5
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1


The function after a monotone transformation... and an associated distribution
3.5
3
2.5
2
1.5
1
0.5
0
-1
-0.5
0
0.5
1

The Basis of Simulated Annealing

- If we wish to obtain minimisers[1] of a function H, we can consider the density

$$\pi_1 \propto \exp\left(-H(x)\right).$$

- In fact, we can make use of a family of related densities

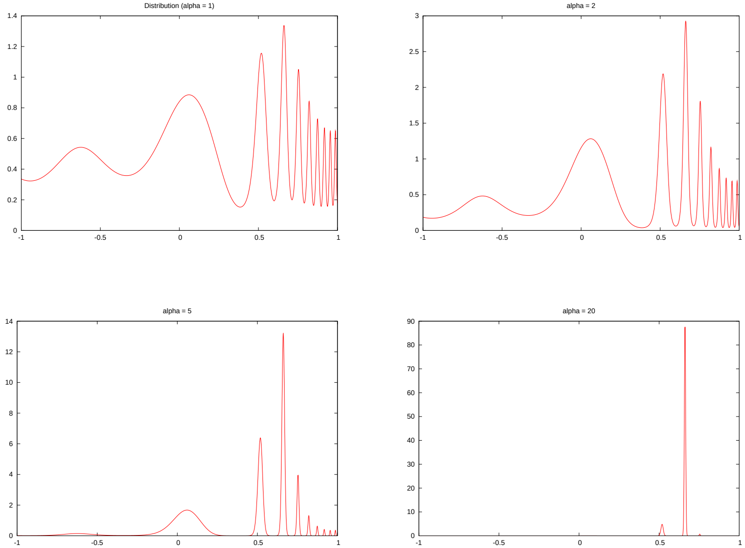
$$\pi_\alpha(x) \propto \exp\left(-\alpha H(x)\right).$$

- As $\alpha \to \infty$ we obtain a distribution concentrated on the modes of π_1.
- One can also consider the definition of these densities to be:

$$\pi_\alpha(x) \propto \pi_1(x)^\alpha.$$

[1] or maximisers of $-H$

Returning to the example



Asymptotically...

If:

- λ is absolutely continuous wrt Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$,
- $\pi_\alpha(dx) := \frac{\lambda(dx)\varphi(x)^\alpha}{\int \lambda(dy)\varphi(y)^\alpha}$,
- and mild regularity conditions apply to $\log \varphi$,

then

$$\lim_{\alpha \to \infty} \pi_\alpha(dt) \to \pi_\infty(dt) \propto \sum_{e \in E^\star} w(e)\delta_e(dt), \tag{1}$$

$$w(e) = \det\left[- \left. \frac{\partial^2 \log \varphi}{\partial \theta_m \partial \theta_n} \right|_{\theta = e} \right]^{-1/2} \tag{2}$$

This follows by a straightforward adaptation of the result of (Hwang, 1980, Theorem 2.1).

Rare Events

Given some set $\mathcal{R}$, such that $\mathbb{P}(\mathcal{R}) \ll 1$:

- Estimate $\mathbb{P}(\mathcal{R})$
- Obtain a collection of samples from $\mathbb{P}$ conditioned upon $\mathcal{R}$: sample from $\mu(\cdot) \propto \mathbb{P}(\cdot)\mathbb{I}_{\mathcal{R}}(\cdot)$

Simple Monte Carlo isn't the answer.

If $\{X_i\}_{i \geq 1} \sim \mathbb{P}$ and $\hat{p}(\mathcal{R}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\mathcal{R}}(X_i)$:

$$\begin{aligned}
\mathbb{E}\left[\hat{p}(\mathcal{R})\right] &= \mathbb{P}(\mathcal{R}) \\
\mathsf{Var}\left[\hat{p}(\mathcal{R})\right] &= \mathbb{P}(\mathcal{R})(1 - \mathbb{P}(\mathcal{R}))/N \\
\frac{\mathsf{Var}\left[\hat{p}(\mathcal{R})\right]}{\mathbb{E}\left[\hat{p}(\mathcal{R})\right]^2} &= \frac{\mathbb{P}(\mathcal{R})(1 - \mathbb{P}(\mathcal{R}))}{\mathbb{P}(\mathcal{R})^2 N} \approx \frac{1}{N\mathbb{P}(\mathcal{R})}
\end{aligned}$$

Importance sampling is suggested...

Importance Sampling for Rare Event Estimation

- We wish to integrate $\mathbb{I}_{\mathcal{R}}(\cdot)$ with respect to $\mathbb{P}$.
- Ideally, sample from $\mu(dx) = \mathbb{P}(dx)\mathbb{I}_{\mathcal{R}}(x)/\mathcal{Z}$,
- and weight the samples according to $W = \frac{\mathbb{P}(X)}{\mu(X)} = \mathcal{Z}^{-1}$.
- $\mathcal{Z}$ is precisely the quantity which we wish to estimate.
- Sampling efficiently from μ is typically impossible.
- $\mathcal{R}$ is usually characterised as the level set of a potential function.
- Could we use distributions $\propto \mathbb{P}(\cdot)(1 + \exp(-\alpha(n)[G(\cdot) - G^{\star}]))^{-1}$, say?

Approximate Counting

Counting problems fit into the same framework:

- Given $E = \{0,1\}^p$, and $S : E \to \mathbb{R}\dots$
- what is $K = |\{x \in E : S(x) \leq s^\star\}|$?

Which could alternatively be reformulated as

- Given $\mathbb{P}(A) = 2^{-p}|A| \; \forall A \subset E\dots$
- what is $K = 2^p \mathbb{P}(B)$ with $B = \{x : S(x) \leq s^\star\}$?

The story so far…

- Three common problems:
  - optimization
  - rare event simulation
  - approximate counting

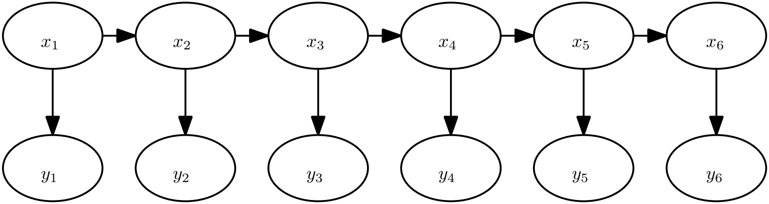
  can be transformed into the problem of obtaining samples from a complex distribution.
- These distributions can be characterised as the restriction of a probability distribution to the level sets of a potential function.
- But simulating from complex distributions is, itself, hard.

Sequential Monte Carlo Samplers

Sequential Monte Carlo

SMC is widely used for the approximate solution of the optimal filtering problem.

- Given an observation sequence $\{y_n\}$,
- associated with a latent Markov chain $\{x_n\}$ termed the *state* process,
- and the conditional independence structure:



- one wishes to estimate $p(x_n|y_{1:n})$ sequentially as observations become available.

- By conditional independence,

$$p(x_n|x_{1:n-1}, y_{1:n-1}) = p(x_n|x_{n-1}),$$
$$p(y_n|x_{1:n}, y_{1:n-1}) = p(y_n|x_n).$$

- Thus: $p(x_{1:n}|y_{1:n-1}) = p(x_n|x_{n-1})p(x_{1:n-1}|y_{1:n-1})$.
- And applying Bayes' rule:

$$p(x_{1:n}|y_{1:n}) = \frac{p(x_{1:n}|y_{1:n-1})p(y_n|x_n)}{\int p(x'_n|y_{1:n-1})p(y_n|x'_n)dx'_n}.$$

- Unfortunately, this integral is intractable.

- Making use of the recursion:

$$p(x_{1:n}|y_{1:n}) \propto p(x_{1:n-1}|y_{1:n-1}) \times p(x_n|x_{n-1})p(y_n|x_n),$$

- A set of weighted samples from $p(x_{1:n-1}|y_{1:n-1})$, $\{W_{n-1}^{(i)}, X_{1:n-1}^{(i)}\}$ is extended according to some proposal $q_n(\cdot|x_{n-1})$ and re-weighted,

$$W_n^{(i)} \propto W_{n-1}^{(i)} \frac{p(X_n^{(i)}|X_{n-1}^{(i)})}{q(X_n^{(i)}|X_{n-1}^{(i)})} p(y_n|X_n^{(i)}).$$

- This process continues iteratively, with resampling applied when the sample diversity falls too low.
- The empirical measure associated with the particle set may be used to approximate the filtering measure at each time-step.

SMC Samplers

Actually, these techniques can be used to sample from *any* sequence of distributions (Del Moral et al., 2006):

- Given a sequence of *target* distributions, $\{\eta_n\}$, on measurable spaces $(E_n, \mathcal{E}_n)$
- Construct a synthetic sequence $\{\tilde{\eta}_n\}$ on the product spaces $\bigotimes_{p=1}^{n} (E_p, \mathcal{E}_p)$ by introducing *arbitrary* auxiliary Markov kernels, $L_p : E_{p+1} \otimes \mathcal{E}_p \rightarrow [0, 1]$:

$$\tilde{\eta}_n(dx_{1:n}) = \eta_n(dx_n) \prod_{p=1}^{n-1} L_p\left(x_{p+1}, dx_p\right),$$

  which each admit one of the target distributions as their final time marginal.

SMC Outline

- Given a sample $\{X_{1:n-1}^{(i)}\}_{i=1}^{N}$ targeting $\tilde{\eta}_{n-1}$,
- sample $X_n^{(i)} \sim K_n(X_{n-1}^{(i)}, \cdot)$,
- calculate

$$W_n(X_{1:n}^{(i)}) = \frac{\tilde{\eta}_n(X_{1:n}^{(i)})}{\tilde{\eta}_{n-1}(X_{1:n-1})K_n(X_{n-1}^{(i)}, X_n^{(i)})}$$

$$= \frac{\eta_n(X_n^{(i)}) \prod_{p=1}^{n-1} L_p(X_{p+1}^{(i)}, X_p^{(i)})}{\eta_{n-1}(X_{n-1}^{(i)}) \prod_{p=1}^{n-2} L_p(X_{p+1}^{(i)}, X_p^{(i)}) K_n(X_{n-1}^{(i)}, X_n^{(i)})}$$

$$= \frac{\eta_n(X_n^{(i)}) L_{n-1}(X_n^{(i)}, X_{n-1}^{(i)})}{\eta_{n-1}(X_{n-1}^{(i)}) K_n(X_{n-1}^{(i)}, X_n^{(i)})}.$$

- Resample, yielding: $\{X_{1:n}^{(i)}\}_{i=1}^{N}$ targeting $\tilde{\eta}_n$.

Another SMC Summary

At each iteration, given a set of weighted samples $\{X_{n-1}^{(i)}, W_{n-1}^{(i)}\}_{i=1}^{N}$ targeting η_{n-1}:

- Sample $X_n^{(i)} \sim K_n(X_{n-1}^{(i)}, \cdot)$.
- $\left\{(X_{n-1}^{(i)}, X_n^{(i)}), W_{n-1}^{(i)}\right\}_{i=1}^{N} \sim \eta_{n-1}(X_{n-1})K_n(X_{n-1}, X_n)$.
- Set weights $W_n^{(i)} = W_{n-1}^{(i)} \frac{\eta_n(X_n)L_{n-1}(X_n,X_{n-1})}{\eta_{n-1}(X_{n-1})K_n(X_{n-1},X_n)}$.
- $\left\{(X_{n-1}, X_n), W_n^{(i)}\right\}_{i=1}^{N} \sim \eta_n(X_n)L_{n-1}(X_n, X_{n-1})$ and, marginally, $\left\{X_n^{(i)}, W_n^{(i)}\right\}_{i=1}^{(i)} \sim \eta_n$.
- Resample to obtain an unweighted particle set.
- Hints that we'd like $L_{n-1}(x_n, x_{n-1}) = \frac{\eta_{n-1}(x_{n-1})K_n(x_{n-1},x_n)}{\int \eta_{n-1}(x'_{n-1})K_n(x'_{n-1},x_n)}$.

SMC for Optimisation

Answers and Questions

How can we use SMC?

Using SMC for optimisation has been proposed in a number of places (for example (Neal, 1998; Del Moral et al., 2006)). The idea is simple:

- Identify η_n with $\pi_{\alpha(n)}$ where α is some monotone function.
- Use SMC to sample iteratively from each η_n.
- For large enough $\alpha(n)$ we obtain a solution.

As yet only one or two special cases and related areas appear to have been considered in detail:

- Maximum likelihood / a posteriori estimation in latent variable models (Johansen et al., 2007).
- Rare event estimation (Johansen et al., 2006).

Why use SMC?

- Extremely flexible: choice of distributions, annealing schedule, proposal kernels. . .
- The population provides robustness.
- Faster "annealing" seems possible.
- It doesn't require a parametric family of proposals matched to the target.
- The iterative framework simplifies the design of proposal distributions.

Open Problems

A number of methodological questions need to be addressed:

- How can the balance between the number of particles and the number of distributions be dealt with?
- How should the proposal kernels be chosen?
- Adaptative methods: how can the sequence of distributions, and the proposal distributions be varied adaptively?

A number of interesting theoretical questions also remain to be addressed:

- Can we obtain meaningful bounds upon the variance and bias of the estimator?
- Are there "better" sequences of distributions?

References

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society B*, 63(3): 411–436, 2006.

C.-R. Hwang. Laplace's method revisited: Weak convergence of probability measures. *The Annals of Probability*, 8(6): 1177–1182, December 1980.

A. M. Johansen, P. Del Moral, and A. Doucet. Sequential Monte Carlo samplers for rare events. In *Proceedings of the 6th International Workshop on Rare Event Simulation*, Bamberg, Germany, October 2006.

A. M. Johansen, A. Doucet, and M. Davy. Particle methods for maximum likelihood parameter estimation in latent variable models. *Statistics and Computing*, 2007. To appear.

R. M. Neal. Annealed importance sampling. Technical Report 9805, University of Toronto, Department of Statistics, 1998. URL `ftp://ftp.cs.toronto.edu/pub/radford/ais.ps`.