Monte Carlo Methods

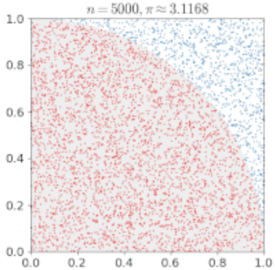
Václav Šmídl

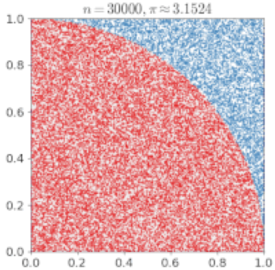
April 7, 2020

Monte Carlo

Used e.g. in numerical integration

$$\int_0^1 f(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i), \quad x_i \sim U(0,1).$$





Monte Carlo

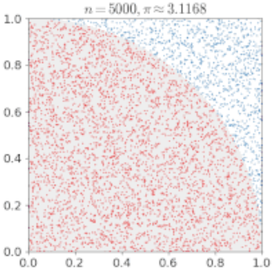
Used e.g. in numerical integration

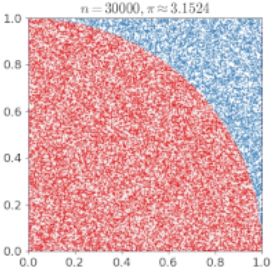
$$\int_0^1 f(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i), \quad x_i \sim U(0,1).$$

In probability calculus, we integrate a lot (expectations)

$$E_{x\sim U(0,1)}(f(x)) = \int_0^1 f(x)U(0,1)dx$$

Many operations can be simplified using the idea of empirical distribution function.





Monte Carlo

Used e.g. in numerical integration

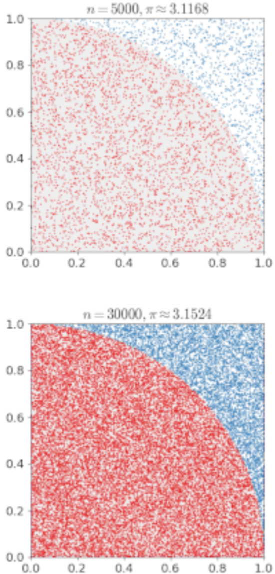
$$\int_0^1 f(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i), \quad x_i \sim U(0,1).$$

In probability calculus, we integrate a lot (expectations)

$$E_{x\sim U(0,1)}(f(x)) = \int_0^1 f(x)U(0,1)dx$$

Many operations can be simplified using the idea of empirical distribution function.
Empirical probability "density" function

$$p(x) \approx \frac{1}{N}\sum \delta(x - x^{(i)}), \quad x^{(i)} \sim p(x),$$



Empirical distribution function

The empirical distribution function is

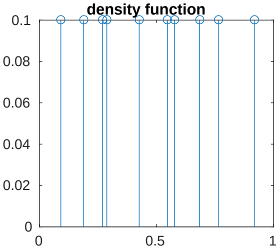
$$\hat{F}(x) = \frac{\text{number of } x^{(i)} < t}{N}$$

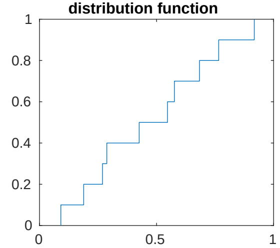
Consider "density" function:

$$p_x(x) \approx \frac{1}{N} \sum \delta(x - x^{(i)}), \quad x^{(i)} \sim p(x),$$

then

$$\hat{F}(x) = \int_{-\infty}^{x} p_x(t) dt$$





Empirical distribution function

The empirical distribution function is

$$\hat{F}(x) = \frac{\text{number of } x^{(i)} < t}{N}$$

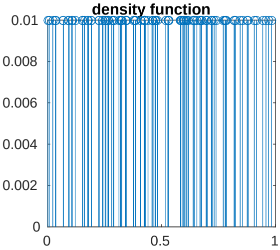
Consider "density" function:

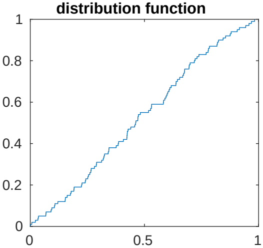
$$p_x(x) \approx \frac{1}{N} \sum \delta(x - x^{(i)}), \quad x^{(i)} \sim p(x),$$

then

$$\hat{F}(x) = \int_{-\infty}^{x} p_x(t)dt$$

Converges to $F_{U(0,1)}$ with $N \to \infty$.





Tricks for Monte Carlo

Replace probability $p(x)$ by its estimate

$$U(0,1) \approx \frac{1}{N} \sum \delta(x - x^{(i)}),$$

then expectations are

$$\begin{aligned}
E_{x \sim U(0,1)}(f(x)) &= \int_0^1 f(x) U(0,1) dx \\
&= \int_0^1 f(x) \left(\frac{1}{N} \sum_i \delta(x - x^{(i)}) \right) dx \\
&= \frac{1}{N} \sum_i \int_0^1 f(x) \left(\delta(x - x^{(i)}) \right) dx \\
&= \frac{1}{N} \sum_i f(x^{(i)})
\end{aligned}$$

Marginalization in empirical pdf

Replace probability $p(x_1, x_2)$ by its estimate

$$p(x_1, x_2) \approx \frac{1}{N} \sum_i \delta(x_1 - x_1^{(i)}) \delta(x_2 - x_2^{(i)}),$$

The marginal

$$\begin{aligned} p(x_1) &= \int_{x_2} p(x_1, x_2) dx_2 \\ &= \int \frac{1}{N} \sum_i \delta(x_1 - x_1^{(i)}) \delta(x_2 - x_2^{(i)}) dx_2, \\ &= \frac{1}{N} \sum_i \delta(x_1 - x_1^{(i)}) \int \delta(x_2 - x_2^{(i)}) dx_2, \\ &= \frac{1}{N} \sum_i \delta(x_1 - x_1^{(i)}) \end{aligned}$$

Monte Carlo for Bayesian inference

The aim of Monte Carlo methods is to approximate the posterior distribution by an empirical distribution

$$p(\theta|D) \approx \frac{1}{N} \sum \delta(\theta - \theta^{(i)}),$$

Problem: we can not sample from unknown $p(\theta|D)$

Monte Carlo for Bayesian inference

The aim of Monte Carlo methods is to approximate the posterior distribution by an empirical distribution

$$p(\theta|D) \approx \frac{1}{N}\sum \delta(\theta - \theta^{(i)}),$$

Problem: we can not sample from unknown $p(\theta|D)$

Many strategies with different properties.

1. Monte Carlo Markov Chain (MCMC)
   1.1 Metropolis-Hastings (Gibbs sampler)
   1.2 Hybrid MC (Hamiltonian Monte Carlo)
2. Importance sampling,
   2.1 Adaptive importance sapling
   2.2 Population Monte Carlo

Convergence assured under mild conditions, different convergence rate.

MCMC: Metropolis Hastings

Instead of fixed distribution, we define a Markov chain that converges to the true distribution.

1. choose transition kernel $q(\theta|\theta^{(i)})$,
2. generate sample $\theta^* \sim q(\theta|\theta^{(i)})$,
3. With probability

$$\min\left(1, \frac{p(\theta^*)q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)})q(\theta^*|\theta^{(i)})}\right)$$

   accept $(i = i + 1, \theta^{(i)} = \theta^*)$, else reject; goto 2.

MCMC: Metropolis Hastings

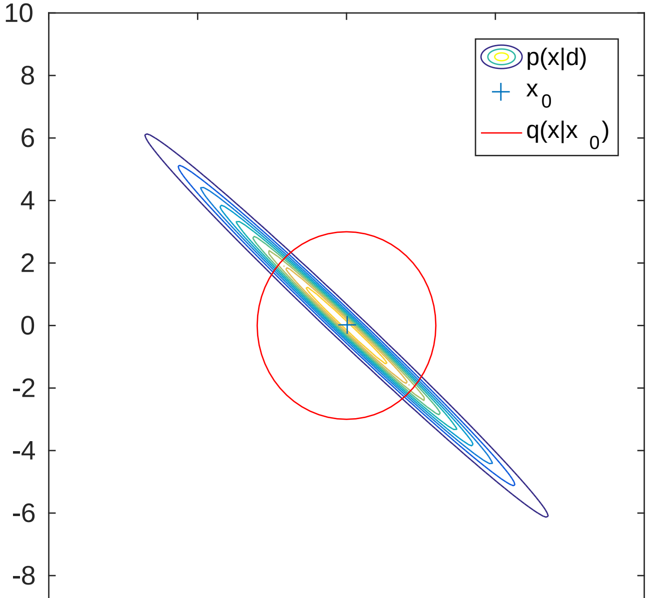
Instead of fixed distribution, we define a Markov chain that converges to the true distribution.

1. choose transition kernel $q(\theta|\theta^{(i)})$,
2. generate sample $\theta^* \sim q(\theta|\theta^{(i)})$,
3. With probability
$$\min\left(1, \frac{p(\theta^*)q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)})q(\theta^*|\theta^{(i)})}\right)$$
accept $(i = i + 1, \theta^{(i)} = \theta^*)$, else reject; goto 2.

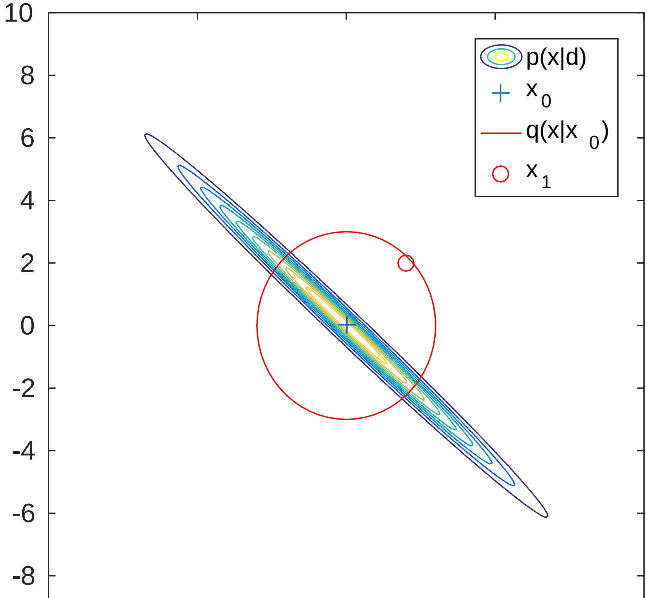
How to choose the kernel:

- Random walk (Gaussian), with parameters ϕ

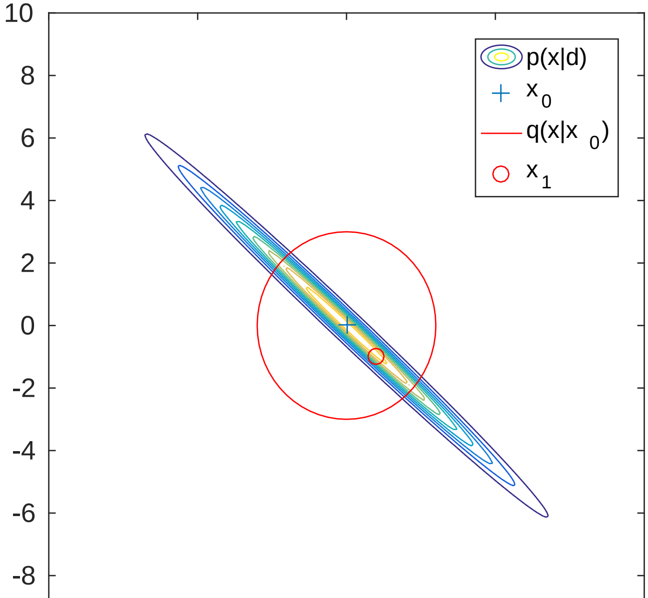
Kernel selection – essential



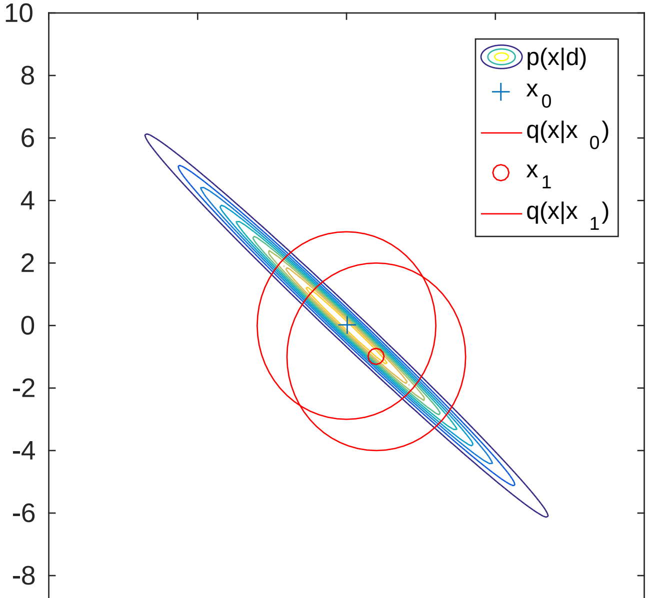
Kernel selection – essential



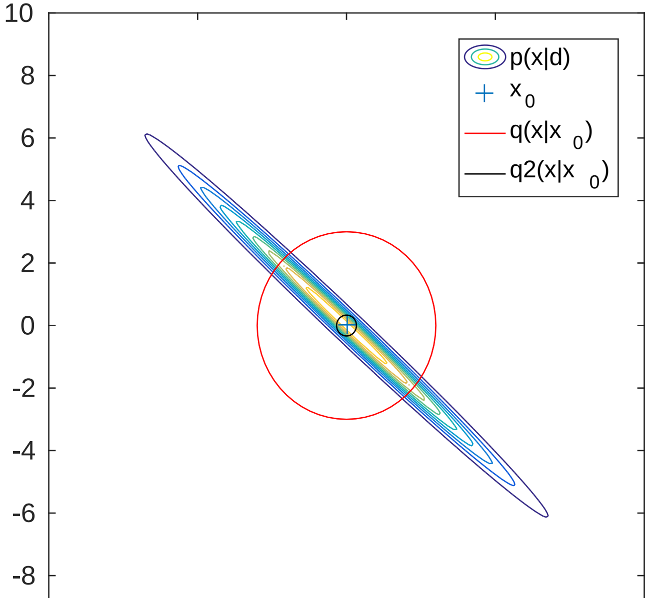
Kernel selection – essential



Kernel selection – essential



Kernel selection – essential



MCMC: Gibbs sampler

Special case of MH for mutidimensional distributions.

$$p(\theta_1, \theta_2, \ldots, \theta_k)$$

sample as follows:

1. generate sample $\theta_1^{(i+1)} \sim p(\theta_1|\theta_2^{(i)}, \ldots \theta_k^{(i)})$,
2. generate sample $\theta_2^{(i+1)} \sim p(\theta_2|\theta_1^{(i+1)}, \ldots \theta_k^{(i)})$,

   $\vdots$

3. generate sample $\theta_k^{(i+1)} \sim p(\theta_2|\theta_1^{(i+1)}, \ldots \theta_{k-1}^{(i+1)})$,

Suitable when these distributions are tractable.

- MH probability of acceptance equal to **one**.

MCMC: Gibbs sampler

Special case of MH for mutidimensional distributions.

$$p(\theta_1, \theta_2, \ldots, \theta_k)$$

sample as follows:

1. generate sample $\theta_1^{(i+1)} \sim p(\theta_1 | \theta_2^{(i)}, \ldots \theta_k^{(i)})$,
2. generate sample $\theta_2^{(i+1)} \sim p(\theta_2 | \theta_1^{(i+1)}, \ldots \theta_k^{(i)})$,

   $\vdots$
3. generate sample $\theta_k^{(i+1)} \sim p(\theta_2 | \theta_1^{(i+1)}, \ldots \theta_{k-1}^{(i+1)})$,

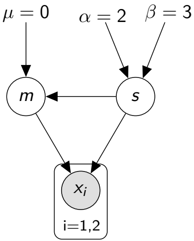
Suitable when these distributions are tractable.

- MH probability of acceptance equal to **one**.
- not suitable for parallel computing

Toy problem from Lecture 1&2

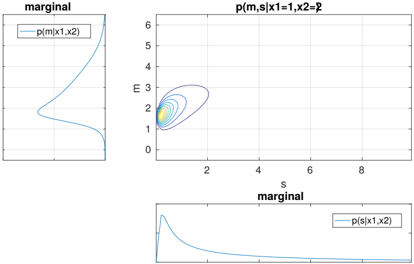
Model

$$\begin{aligned} p(s) &= iG(\alpha, \beta) \\ p(m|s) &= \mathcal{N}(\mu, s) \\ p(x_i|m, s) &= \mathcal{N}(m, s) \end{aligned}$$



- Observations x_i are sampled from a Gaussian with unknown mean and variance.
- We have some prior information about the mean and variance
- Seek

$$p(m, s|m, \alpha, \beta, \mu) \equiv p(\theta|D)$$



Toy: Gibbs sampler

$$
\begin{aligned}
p(m, s|x_1, x_2, \mu, \alpha, \beta, \phi)& \\
&\propto \frac{1}{s}\frac{1}{s^{\alpha+1}} \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta}{s}\right)
\end{aligned}
$$

$$
\begin{aligned}
p(m|s, x_1, x_2, \mu, \alpha, \beta, \phi)& \\
&\propto \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi}\right) \\
&\propto \exp\left(-\frac{1}{2}\left[m^2(\frac{1}{\phi} + \frac{2}{s}) - 2m(\frac{\mu}{\phi} + \frac{x_1+x_2}{s})\right]\right) \\
&= \mathcal{N}\left(m; \left(\frac{1}{\phi} + \frac{2}{s}\right)^{-1}\left(\frac{\mu}{\phi} + \frac{x_1+x_2}{s}\right), \left(\frac{1}{\phi} + \frac{2}{s}\right)^{-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
p(s|m, x_1, x_2, \mu, \alpha, \beta, \phi)& \\
&\propto \frac{1}{s^{\alpha+2}} \exp\left(-\frac{1}{2}\frac{(m-x_1)^2}{s} - \frac{1}{2}\frac{(m-x_2)^2}{s} - \frac{\beta}{s}\right) \\
&= i\mathcal{G}(\alpha+1, 0.5(m-x_1)^2 + 0.5(m-x_2)^2 + \beta)
\end{aligned}
$$

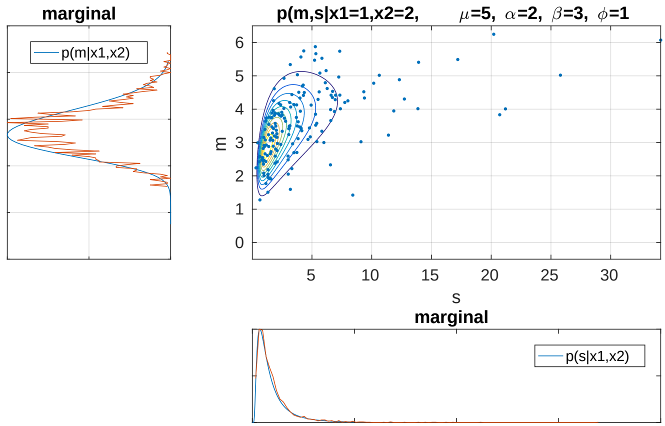
Toy: Gibbs sampler

Repeat:

$$m^{(i+1)} \sim \mathcal{N}\left(m; \left(\frac{1}{\phi} + \frac{2}{s^{(i)}}\right)^{-1} \left(\frac{\mu}{\phi} + \frac{x_1 + x_2}{s^{(i)}}\right), \left(\frac{1}{\phi} + \frac{2}{s^{(i)}}\right)^{-1}\right)$$

$$s^{(i+1)} \sim i\mathcal{G}(\alpha + 1, 0.5(m^{(i+1)} - x_1)^2 + 0.5(m^{(i+1)} - x_2)^2 + \beta)$$

Toy: Gibbs sampler



Hamiltonian(Hybrid) Monte Carlo

- view log-probability as potential energy

$$U(\theta) = -\log p(\theta) = \frac{\theta^2}{2},$$

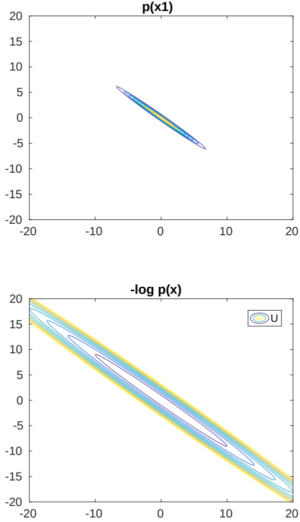
- add kinetic energy in variable p

$$K(p) = \frac{p^2}{2},$$

- define Hamiltonian

$$\frac{d\theta}{dt} = p, \quad \frac{dp}{dt} = -\theta$$

- simulate *differential equation* for selected $t = 0 \ldots t_s$
- resulting sample θ' is independent of p'
- Asymptotically converges to samples from true density.



Hamiltonian(Hybrid) Monte Carlo

- view log-probability as potential energy

$$U(\theta) = -\log p(\theta) = \frac{\theta^2}{2},$$

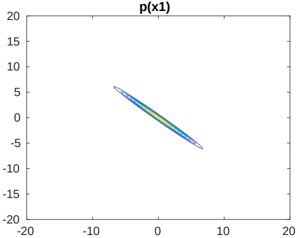
- add kinetic energy in variable p

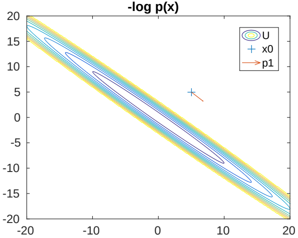
$$K(p) = \frac{p^2}{2},$$

- define Hamiltonian

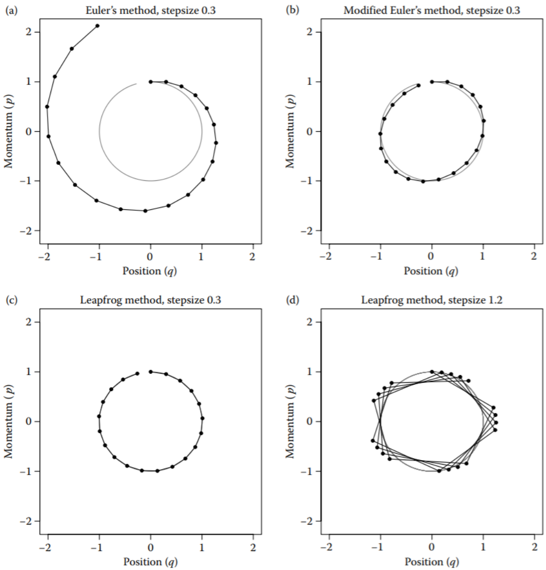
$$\frac{d\theta}{dt} = p, \quad \frac{dp}{dt} = -\theta$$

- simulate *differential equation* for selected $t = 0 \ldots t_s$
- resulting sample θ' is independent of p'
- Asymptotically converges to samples from true density.





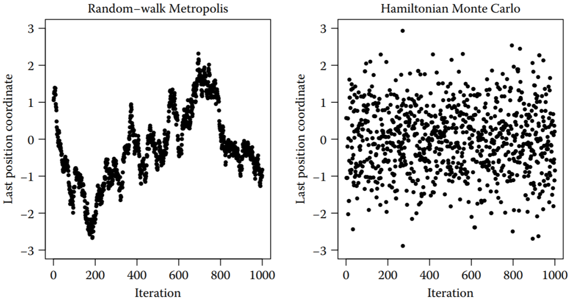
Numerical issues: leapfrog algorithm



[Neal, 2011]

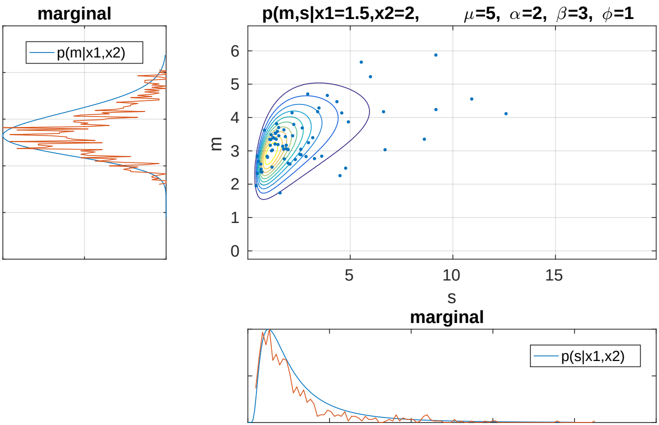
HMC advantages and disadvantages

- Able to use information about gradient
  - troubles with discrete variables
- Generated samples are not excessively correlated (check autocorrelation)



- Much faster exploraton of the space
  - at computational cost (doubles the number of variables)
- How to choose stepsize and number of leapfrogs
  - NUTS, DynamicHMC, etc.

Results



marginal
p(m|x1,x2)
p(m,s|x1=1.5,x2=2, μ=5, α=2, β=3, φ=1
m
0
1
2
3
4
5
6
5
10
15
s
marginal
p(s|x1,x2)

Probabilistic programming

Universal nature of HMC gave rise to automatic tools:

STAN: https://mc-stan.org/

- HMC, NUTS
- Variational inference
- Matlab, R, Mathematica, Python, ...

Turing.jl: https://github.com/TuringLang/Turing.jl

- HMC, NUTS, SMC, ...
- Julia

PyMC3:

- Python

Model development almost too easy

```
@model gdemo(x) = begin
  s ~ InverseGamma(2,3)
  m ~ Normal(0, sqrt(s))
  x[1] ~ Normal(m, sqrt(s))
  x[2] ~ Normal(m, sqrt(s))
  return s, m
end

chain = sample(gdemo([1.5, 2.0]), SGLD(10000, 0.5))
```

- Non-conjugate priors
  - log-normal instead of inverse gamma
- automatic chain rule, differentiation
- Hard part: analyze results
- High dimensions?

Importance Sampling

To represent

$$p(\theta|\cdot) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\theta - \theta^{(i)}). \quad (1)$$

an ideal sampler should sample $\theta^{(i)} \sim p(\theta|\cdot)$, which is not available. Using

$$p(\theta|D) = p(\theta|D)\frac{q(\theta)}{q(\theta)},$$

we can approximate $q(\theta)$ by (1) by sampling $\theta^{(i)} \sim q(\theta)$.

$$\begin{aligned}
p(\theta) &\propto \frac{p(\theta)}{q(\theta)}\frac{1}{N}\sum_{i=1}^{N}\delta(\theta - \theta^{(i)}), \\
&\propto \sum_{i=1}^{N}\tilde{w}_i\delta(\theta - \theta^{(i)}), & \tilde{w}_i &= \frac{p(\theta^{(i)})}{q(\theta^{(i)})} \\
&= \sum_{i=1}^{N}w_i\delta(\theta - \theta^{(i)}) & w_i &= \frac{\tilde{w}_i}{\sum_{i=1}^{N}\tilde{w}_i}
\end{aligned}$$

Algebra of weighted empirical distribution

Moments:

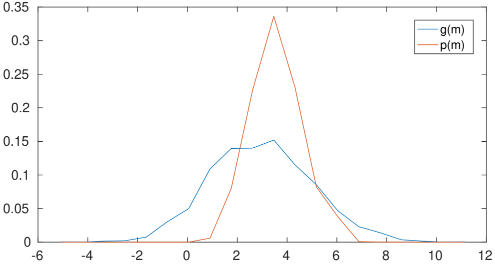
$$E(f(\theta)) = \sum_{i=1}^{N} w_i f(\theta^{(i)})$$

Histogram:

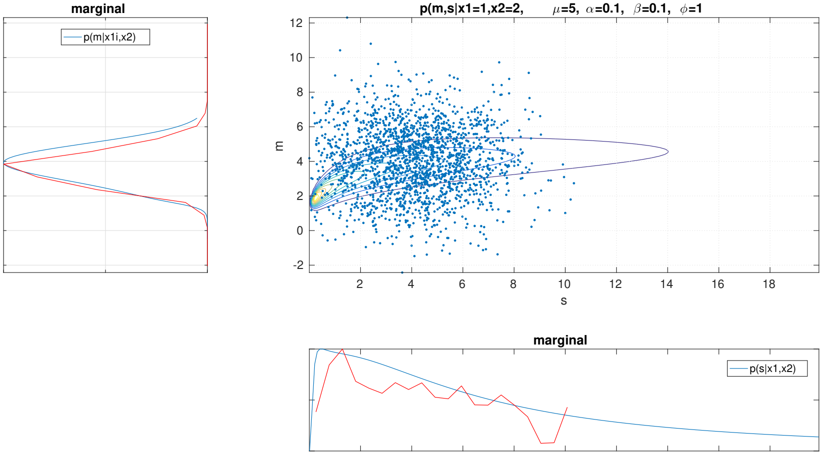
$$c_i = \sum_{i:x_i \in (l_i, u_i]} 1$$

Weighted histogram:

$$c_i = \sum_{i:x_i \in (l_i, u_i]} w_i$$



Toy: Importance sampling $N = 2000$



- Sample from heavy tailed distributions...

Adaptive Importance sampling

What if $q(\theta)$ is too far from $p(\theta)$?

Adaptive Importance sampling

What if $q(\theta)$ is too far from $p(\theta)$? Move it. Choose $q(\theta|\phi)$ and find $\hat{\phi}$.

Adaptive Importance sampling

What if $q(\theta)$ is too far from $p(\theta)$? Move it. Choose $q(\theta|\phi)$ and find $\hat{\phi}$.

Population MC: [Cappé, O., Guillin, A., Marin, J. M., & Robert, C. P. (2004).]

- Sample one generation
- compute weights, estimate parameter
- Sample next generation

Adaptive Importance sampling

What if $q(\theta)$ is too far from $p(\theta)$? Move it. Choose $q(\theta|\phi)$ and find $\hat{\phi}$.

Population MC: [Cappé, O., Guillin, A., Marin, J. M., & Robert, C. P. (2004).]

- Sample one generation
- compute weights, estimate parameter
- Sample next generation

AMIS: [CORNUET, J. M., MARIN, J. M., Mira, A., & Robert, C. P. (2012)]

- Consider each generation to be a component in deterministic mixture

$$q(\theta) = \sum_{g=1}^{G} q_g(\theta)$$

IMIS: [Steele, R. J., Raftery, A. E., & Emond, M. J. (2006).]

- add component centered at sample with high weight

Assignment

Gibbs sampler	
– toy problem	10
– linear regression	20
– mixture model	30
HMC (probabilistic programming)	
– toy problem	10
– linear regression	20
– mixture model	30