Why samples?

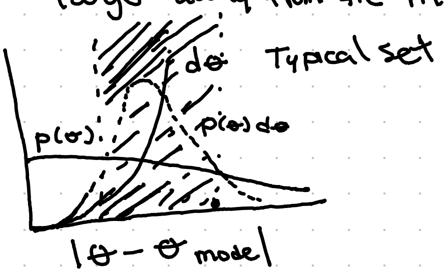
Maximum likelihood is not enough for many applications
or maximum a posteriori
Especially in high dimensions

The probability contained in a given region of param. space V is

$$\int_V p(\theta)\, d\theta$$

prob. density × volume

$p(\theta)$ peaks at the mode, but $d\theta$ is much larger away from the mode in high-dimensions



samples allow us to compute expectation values

consider $X = f(\theta)$

$$\mathbb{E}[X] = \int_V f(\theta)\, p(\theta)\, d\theta$$

$$\sigma^2[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

If we have samples $\theta \sim p(\theta)$, then we can estimate

$$\int_V f(\theta) p(\theta)\, d\theta \approx \langle f(\theta) \rangle_{\theta \sim p(\theta)} = \frac{1}{N} \sum_{i=1}^{N} f(\theta_i)$$

This Monte Carlo estimate of the integral is unbiased:

$$\mathbb{E}[\langle f \rangle] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[f(\theta_i)] = \frac{1}{N} \sum_{i=1}^{N} \int_V f(\theta) p(\theta)\, d\theta$$

$$= \int_V f(\theta) p(\theta)\, d\theta \quad ✓$$

and converges with $\sqrt{N}$:

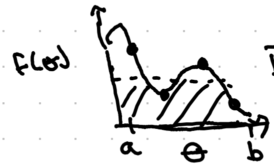
$$\sigma^2[\langle f \rangle] = \sigma^2\left[\frac{1}{N} \sum_{i=1}^{N} f(\theta_i)\right]$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sigma^2[f(\theta_i)]$$ if samples are independent, variance of sum is sum of variance

$$= \frac{1}{N} \sigma^2[f(\theta)]$$

if samples are not independent, replace N with "effective sample size"

Ex: if $p(\theta)$ is uniform,



$$\frac{1}{b-a} \int_a^b f(\theta)\, d\theta \approx \langle f(\theta) \rangle \cdot \frac{(b-a)}{b-a}$$

How can we get $\{\vec{\theta}_i\}$ samples from $p(\vec{\theta})$?

- inverse transform sampling from target with pdf $p(x)$

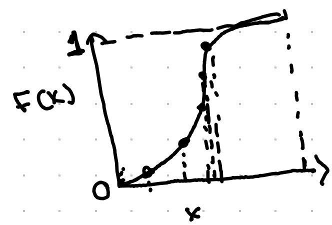
In 1-d, CDF $F(x) = \int_0^x \underbrace{p(x')}_{\text{pdf}} dx'$

$0 < F(x) < 1$

draw $y_i = F(x_i)$ from uniform distribution

$y \sim U(0,1)$

$F^{-1}(y_i) = x_i$



- rejection sampling from target with pdf $p(x)$

draw samples from some easy-to-sample distribution

$x_i \sim g(x)$

Keep x_i with probability $\dfrac{p(x)}{K g(x)}$ where K

is a constant s.t. $p(x) \leq K g(x) \; \forall x$

Markov Chain Monte Carlo

random sampling

sequence of random variables where next step in the sequence depends only on the previous

Efficient, multi-dim. sampling that preferentially samples the "typical set."

Target distribution $\pi(\theta) = p(d|\theta, H)\, p(\theta|H)$

(switching notation a bit, π is not a pdf because it's not normalized)

proposal distribution $q(\theta^{(n+1)}|\theta^{(n)})$

acceptance prob $\alpha(\theta^{(n+1)}|\theta^{(n)})$

Transitition prob. $P(\theta^{(n+1)}|\theta^{(n)})$ given by q and α

We want our chain to respect detailed balance

$$\pi(\theta^{(n+1)})\, P(\theta^{(n)}|\theta^{(n+1)}) = \pi(\theta^{(n)})\, P(\theta^{(n+1)}|\theta^{(n)})$$

i.e. probability of being in state $(n+1)$ and transitioning to (n) is the same as the reverse.

This yields a stationary distribution, i.e. if $\theta^{(n)}$ drawn from π, $\theta^{(n+1)}$ also drawn from π and we will "eventually" converge to the target distribution

Metropolis Hastings algorithm

$$\alpha(\theta^{(n+1)} | \theta^{(n)}) = \min\left[1, \frac{\pi(\theta^{(n+1)})\, q(\theta^{(n)} | \theta^{(n+1)})}{\pi(\theta^{(n)})\, q(\theta^{(n+1)} | \theta^{(n)})}\right]$$

In this case, transition prob is

$$p(\theta^{(n+1)} | \theta^{(n)}) = q(\theta^{(n+1)} | \theta^{(n)})\,\alpha(\theta^{(n+1)} | \theta^{(n)}) + (1-b)\,\delta(\theta^{(n)} - \theta^{(n+1)})$$

$$b = \int d\theta^{(n+1)}\, q(\theta^{(n+1)} | \theta^{(n)})\,\alpha(\theta^{(n+1)} | \theta^{(n)})$$

This choice satisfies detailed balance.

Proof: If $\theta^{(n+1)} = \theta^{(n)}$, trivial

If $\theta^{(n+1)} \neq \theta^{(n)}$,

$$\alpha(\theta^{n+1} | \theta^{n}) = \min\{1, r\}$$

If $r > 1$,

LHS: $\pi(\theta^{n})\, q(\theta^{n+1} | \theta^{n}) \cdot 1$

RHS: $\cancel{\pi(\theta^{n+1})}\, \cancel{q(\theta^{n} | \theta^{n+1})} \cdot \frac{1}{r}$

$$= \;\; '' \;\; '' \cdot \frac{\pi(\theta^{n})\, q(\theta^{n+1} | \theta^{n})}{\cancel{\pi(\theta^{n+1})}\, \cancel{q(\theta^{n} | \theta^{n+1})}} \quad \checkmark$$

If $r < 1$,

LHS: $\pi(\theta^{n})\, q(\theta^{n+1} | \theta^{n}) \cdot r$ similar cancellations

RHS: $\pi(\theta^{n+1})\, q(\theta^{n} | \theta^{n+1}) \cdot 1$

Note that often proposal q is taken to be a Gaussian, so that

$$q(\theta^{n+1} | \theta^{n}) = q(\theta^{n} | \theta^{n+1})$$

and $r = \dfrac{\pi(\theta^{n+1})}{\pi(\theta^{n})}$

Challenges:

- correlations

good jump in one direction isn't necessarily a good jump in another direction

- multimodal

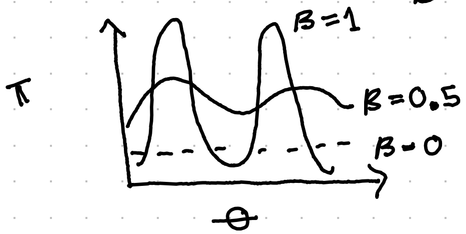
how do you discover second modes?

One technique is Parallel Tempering MCMC

Run multiple chains with different target distributions

$$\pi_\beta(\theta) = [p(d|\theta, H)]^\beta \, p(\theta|H)$$

$\beta = \frac{1}{T}$ ← temperature



$\beta=1$ (cold chain) is our posterior

$\beta \to 0$ (hot chain) is our prior

hot chain can hop around much more easily.

Progress the chains together
every N iterations, propose a swap between neighboring chains i and j

accept the swap with probability

$$A_{ij} = \min\left[1, \frac{\pi_{\beta_i}(\theta_j)}{\pi_{\beta_j}(\theta_i)}\right]$$

$$= \min\left[1, \frac{p(d|\theta_j)}{p(d|\theta_i)}\right]^{\beta_i-\beta_j}$$

helps explore parameter space (find new modes)

also used to calculate evidences
("thermodynamic integration")

Define an evidence for each temperature chain

$$Z_\beta \equiv \int d\theta \, \pi_\beta(\theta) \quad \text{so that} \quad P_\beta(\theta) = \frac{\pi_\beta(\theta)}{Z_\beta}$$

[The evidence we want is $\beta = 1$]

consider $$\frac{\partial}{\partial\beta}\ln(Z_\beta) = \frac{1}{Z_\beta}\frac{\partial}{\partial\beta}Z_\beta$$

$$= \frac{1}{Z_\beta}\int d\theta \frac{\partial}{\partial\beta}\pi_\beta(\theta)$$

$$= \int d\theta \frac{1}{Z_\beta}\frac{\partial \pi_\beta(\theta)}{\partial\beta} \cdot \frac{\pi_\beta(\theta)}{\pi_\beta(\theta)}$$

$$= \int d\theta \, P_\beta(\theta)\frac{\partial \ln \pi_\beta(\theta)}{\partial\beta}$$

$$\frac{\partial \ln \pi_\beta(\theta)}{\partial\beta} = \frac{\partial}{\partial\beta}\left[\ln[(p(d|\theta,H))^\beta] + \ln p(\theta|H)\right]$$

$$= \frac{\partial}{\partial\beta}\left[\beta \ln p(d|\theta,H)\right]$$

$$= \ln p(d|\theta,H)$$ which is just the log-likelihood

so $$\frac{\partial}{\partial\beta}\ln(Z_\beta) = \int d\theta \, P_\beta(\theta)\ln p(d|\theta,H)$$

$$= \mathbb{E}_\beta[\ln p(d|\theta,H)]$$

Easy to estimate for each chain β by taking average over points in the chain

Once we have $\mathbb{E}_\beta[\ln p(d|\theta,H)]$ computed for each chain, we can numerically integrate

$$\int_0^1 \frac{\partial}{\partial\beta}\ln(Z_\beta)\,d\beta = \ln Z_1 - \underbrace{\ln Z_0}$$

$$= Z_1$$ "1 if prior is normalized"