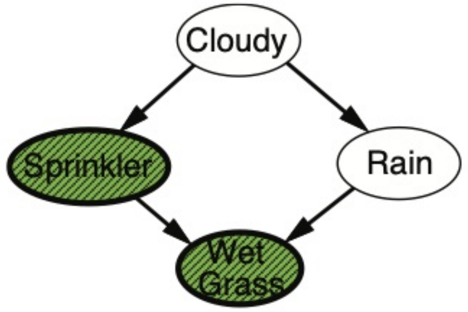
SAMPLING and MARKOV CHAIN MONTECARLO (MCMC)



We want to sample from $p(x \mid y=\hat{y})$

$$p(x \mid y=\hat{y}) = \frac{1}{Z}\, p(x, y=\hat{y})$$

$$\boxed{p(x) = \frac{1}{Z}\, \tilde{p}(x)}$$

$\tilde{p}(x) \geq 0$ but not normalized

$$p(x \mid \hat{y}) = \frac{p(\hat{y} \mid x)\, p(x)}{p(\hat{y})}$$

Z hard to compute

Bayesian inference

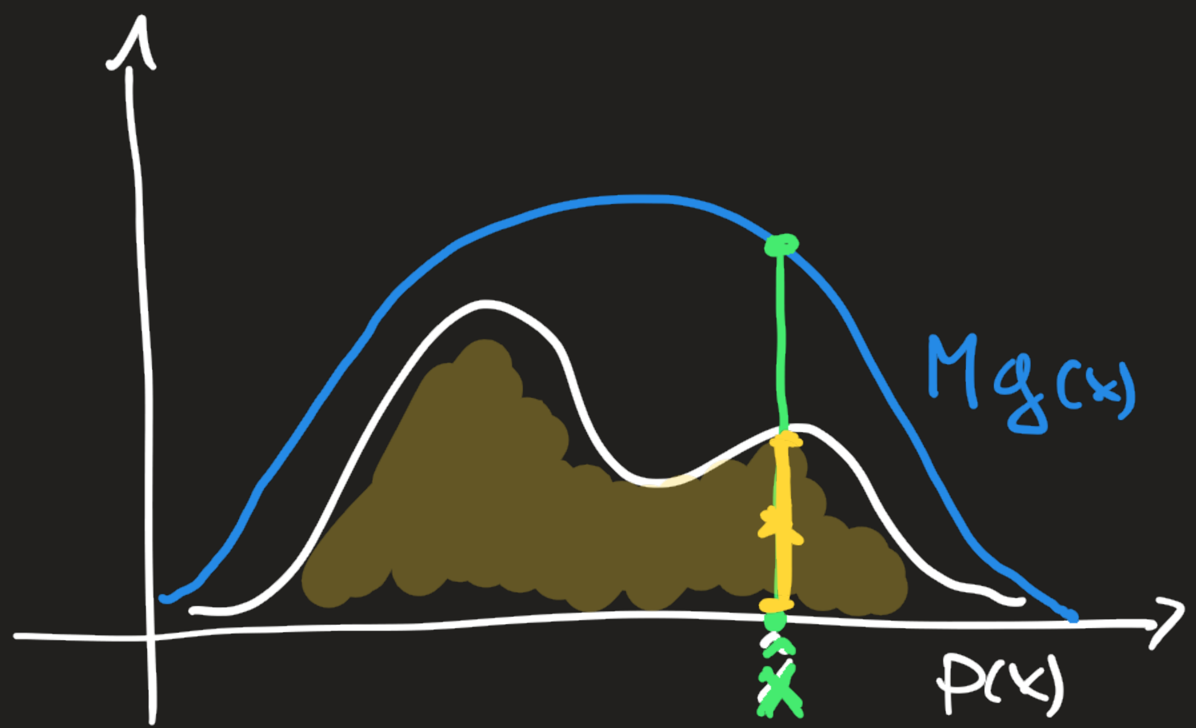
as INDEPENDENT as possible

SAMPLING PROBLEM : generating $x_1, \dots, x_N$ from $p(x)$, knowing $\tilde{p}(x)$

$$\mathbb{E}_x[f(x)] = \int f(x)\, p(x)\, dx \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

- sampling from $q(x) \approx p(x)$ and correct
- MCMC

REJECTION SAMPLING



use $g(x)$ PROPOSAL DISTRIBUTION

$\exists M>0 : Mg(x) \geq p(x), \forall x \in \mathcal{X}$

$\left(\Rightarrow \frac{p(x)}{M g(x)} \leq 1 \right)$

1) SAMPLING $\hat{x}$ FROM $g(x)$

2) ACCEPT $\hat{x}$ WITH PROBABILITY $\alpha = \dfrac{p(\hat{x})}{M g(\hat{x})}$

IF WE REJECT, WE REPEAT UNTIL ACCEPTANCE.

EXPECTED # of SAMPLES FROM g PER p-SAMPLE IS M

INEFFICIENT IN HIGH DIMENSIONS.

IMPORTANCE SAMPLING

$$p(x) = \frac{1}{Z}\tilde{p}(x)$$

$$\mathbb{E}_p[f] = \int f(x)\frac{1}{Z}\tilde{p}(x)\,dx$$

$$= \frac{\int f(x)\,\tilde{p}(x)\,dx}{\int \tilde{p}(x)\,dx}$$

$$\cdot\frac{g(x)}{g(x)} \qquad \frac{g(x)}{g(x)}$$

CONSIDER A PROPOSAL DISTRIBUTION g, which we can sample from

$$\mathbb{E}_p[f] = \int f(x)\,p(x)\,dx = \int \boxed{f(x)\frac{p(x)}{g(x)}}\,g(x)\,dx = \mathbb{E}_g\left[f\frac{p}{g}\right]$$

$$= \frac{\int f(x)\frac{\tilde{p}(x)}{g(x)}\,g(x)\,dx}{\int \frac{\tilde{p}(x)}{g(x)}\,g(x)\,dx}$$

We sample $x_1, \ldots, x_N$ from g

$x_1, \dots, x_N \sim g(x)$

$$\frac{\frac{1}{N} \sum_{i=1}^{N} f(x_i)\, w(x_i)}{\frac{1}{N} \sum_{i=1}^{N} w(x_i)}$$

IMPORTANCE WEIGHTS $w(x_i) := \dfrac{\tilde{p}(x_i)}{g(x_i)}$

IF $\dfrac{f\tilde{p}}{g}$ is approx constant, then estimates can be very good.

IF WEIGHTS vary a lot $\Rightarrow$ large variance, poor estimate

$$\frac{1}{N} \sum w(x_n) \approx Z$$

MARKOV CHAIN

$(X_t)_{t\geq 0}$, $t \in \mathbb{N}$ $\quad X_0, X_1, X_2, \dots \quad X_i \in \mathcal{X}$

$$X_0 \to X_1 \to X_2 \to \dots \to X_n \to \dots$$

- $$\boxed{p(X_n \mid X_{n-1}, \dots, X_0) = p(X_n \mid X_{n-1})}$$

MARKOV or MEMORYLESS PROPERTY

- $p(X_n \mid X_{n-1}) = p(X_1 \mid X_0) \quad , \forall n \geq 1$

TIME HOMOGENEITY

<u>ERGODIC</u> $\quad \forall x, y \in \mathcal{X}, \exists t > 0 : \quad p(X_t = y \mid X_0 = x) > 0$

$p(y \mid x)$ is TRANSITION KERNEL

STATIONARY DISTRIBUTION $\quad \boxed{\pi(y) = \int p(y \mid x)\, \pi(x)\, dx}$ $\quad \pi$ is invariant for the MC dynamics

$p_n(x) = P(X_n \mid X_0) \xrightarrow[n \to \infty]{} \pi(x)$, π is unique

REVERSIBLE MARKOV CHAIN

$(X_t)_{t \geq 0}$ $\quad p(y|x)$ transition kernel

$$\pi(y) = \int p(y|x)\pi(x)\,dx$$

A M.C. IS REVERSIBLE (SATISFIES THE BALANCE CONDITION)

$\exists\, \pi$ distribution s.t.

$$p(x|y)\pi(y) = p(y|x)\pi(x)$$

Then π is STATIONARY

$$\int p(y|x)\pi(x)\,dx = \int p(x|y)\pi(y)\,dx = \pi(y)\int p(x|y)\,dx = \pi(y) \quad \checkmark$$

MCMC

$p(x) = \frac{1}{Z}\tilde{p}(x)$

FIX $q(y|x)$ TRANSITION KERNEL and makes M.C. ERGODIC, easy to sample from

PROPOSAL KERNEL

$\Rightarrow X_t = x$

1) SAMPLE y FROM $q(y|x)$

2) SET $X_{t+1} = \begin{cases} y & \text{with probability } \alpha(y|x) = \min\left\{1, \dfrac{\tilde{p}(y)\cdot q(x|y)}{\tilde{p}(x)\cdot q(y|x)}\right\} \\ x & \end{cases}$

$\frac{p(y)}{p(x)}$ //

METROPOLIS-HASTINGS ACCEPTANCE CRITERION

transition kernel of MC

$$\boxed{p(y|x)\,p(x)} = \alpha(y|x)q(y|x)p(x) \quad y \neq x$$

$$= \min\left\{1, \frac{p(y)}{\cancel{p(x)}}\frac{q(x|y)}{\cancel{q(y|x)}}\right\}\cdot q(y|x)\,\cancel{p(x)}$$

if $q(x|y) = q(y|x)$, this becomes the METROPOLIS criterion

$$= \min\{q(y|x)p(x),\ q(x|y)p(y)\}$$

$$= \min\left\{\frac{q(y|x)}{q(x|y)}\frac{p(x)}{p(y)}, 1\right\}q(x|y)p(y) = \alpha(x|y)\,q(x|y)\,p(y) =$$

$$= \boxed{p(x|y)\,p(y)}$$

detailed balance condition holds.

GIBBS SAMPLING

$p(x) = p(x_1, \dots, x_n)$ $\quad x = (x_1, \dots, x_n)$

ASSUMPTION : we can <u>sample from 1D conditionals</u>

$p(x_i \mid x_{\setminus i})$ $\quad , x_{\setminus i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$

$$\boxed{(x^{(t)})_{t \geq 0}}$$

1) Pick $k \in \{1, \dots, n\}$ — ROUND-ROBIN STRATEGY / UNIFORMLY AT RANDOM

2) Set $x_j^{(t+1)} = x_j^{(t)}$ for $j \neq k$

3) Sample $x_k^{(t+1)} \sim p(x_k \mid x_{\setminus k}^{(t)})$

ISSUES

- ERGODICITY
- IF VARIABLE are STRONGLY CORRELATED, CONVERGENCE IS SLOW

$$q_k(y \mid x) = \begin{cases} p(y_k \mid x_{\setminus k}) & , \text{when } y_{\setminus k} = x_{\setminus k} \\ 0 & , \text{otherwise} \end{cases}$$

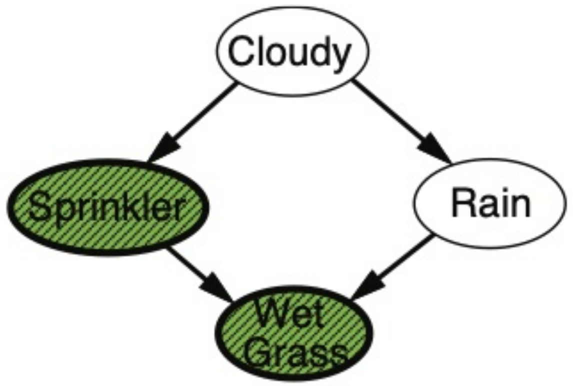
$\Rightarrow \alpha_k(y \mid x) = 1$ acceptance probability

$$\boxed{x_{\setminus k} = y_{\setminus k}}$$

$$\text{MH } \alpha_k: \frac{p(y)\, q_k(x \mid y)}{p(x)\, q_k(y \mid x)} = \frac{p(\cancel{y_k \mid y_{\setminus k}})\, p(\cancel{y_{\setminus k}})}{p(\cancel{x_k \mid x_{\setminus k}})\, p(\cancel{x_{\setminus k}})} \; \frac{p(\cancel{x_k \mid y_{\setminus k}})}{p(\cancel{y_k \mid x_{\setminus k}})} = 1$$

- SAMPLE BLOCKS $x_j, x_k \subseteq x$
- IF $p(x_i \mid x_{\setminus i})$ is NOT KNOWN, then
  - Rejection sampling
  - METROPOLIS_WITHIN_GIBBS

SAMPLING-BASED INFERENCE IN PGM.



x, y : PGM, y is observed, $y = \hat{y}$

make inference on x

(marginal $p(x_j \mid y = \hat{y})$)

SAMPLE from $p(x \mid y = \hat{y})$

we know $\tilde{p}(x) = p(x, y = \hat{y})$, but not $Z = p(y = \hat{y})$

1) Rejection sampling: sample from $p(x, y)$ (ancestral sampling), reject if $y \neq \hat{y}$.

2) MCMC, sampling from $p(x, y = \hat{y})$ or HMC ✓

3) $p(x_i \mid x_{\neg i}, y) = p(x_i \mid mb_i)$ Then GIBBS SAMPLER.

CONVERGENCE DIAGNOSTICS

OUTPUT $\psi : \mathcal{X} \to \mathbb{R}$, $\psi(x)$. we assume that ψ has values in $\mathbb{R}$, and transform it otherwise.

$x_1 \,\_\_\_\, x_N \dots$

$\psi_j = \psi(x_j)$ (NOTATION)

$$\boxed{\bar{\psi} = \frac{1}{N} \sum_j \psi_j}$$ estimate of $\mathbb{E}_p[\psi] = \int \psi(x)\, p(x)\, dx$

- WE NEED MC TO BE STATIONARY
(KEEP SAMPLES ONLY WHEN STATIONARY)

- SAMPLE $\frac{m}{2} \geq 1$ TRAJECTORIES FROM OVER-DISPERSED INITIAL POINTS
- SAMPLE FOR $4n$ STEPS
- THROW AWAY FIRST HALF $\Rightarrow$ we have $2n$ points left (BURN-IN / WARM UP PHASE)
- WE SPLIT THE REMAINED TRAJ IN TWO:

$$\boxed{m \text{ SEQUENCES OF LENGTH } n \text{ EACH}}$$

X_{ij} sample sequence

$1 \le j \le m$

$1 \le i \le w$

$\psi(x_{ij}) = \psi_{ij}$

$$\begin{cases} \bar{\psi}_{\cdot j} = \frac{1}{w}\sum_{i=1}^{w} \psi_{ij} \\ \bar{\psi} = \frac{1}{m}\sum_{j=1}^{m} \bar{\psi}_{\cdot j} \end{cases}$$

$VAR(\psi)$; $VAR(\bar{\psi})$

$[\psi$ is a r.v. $\psi(x)]$

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2 \quad , \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{\cdot j})^2$$

WITHIN VARIANCE

$$W \le VAR(\psi)$$

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{\cdot j} - \bar{\psi})^2$$ - BETWEEN VARIANCE

$$VAR(\psi) \le VAR^{+}(\psi) = \frac{n-1}{n} W + \frac{1}{n} B$$

$$\hat{R} = \sqrt{VAR^{+}(\psi)/W}$$: $\hat{R} > 1$, $\hat{R} \xrightarrow[n \to \infty]{} 1$, when $\hat{R} \le 1.1$ then converged

EFFECTIVE SAMPLE SIZE

$$\mathrm{VAR}[\bar{\Psi}]$$

if ~~m·n samples are independent~~, then $\mathrm{VAR}(\bar{\Psi}) = \dfrac{\mathrm{VAR}(\Psi)}{n \cdot m}$

$$n \cdot m \ \mathrm{VAR}(\bar{\Psi}) \simeq \left(1 + 2\sum_{k=1}^{\infty} \rho_k\right) \mathrm{VAR}(\Psi)$$

ρ_K IS THE AUTOCORRELATION OF LAG K :

$$\rho_K = \mathrm{CORR}\left[\Psi(X_i), \Psi(X_{i+k})\right].$$

$$n_{eff} = \frac{n \cdot m}{\left(1 + 2\sum_{k=1}^{\infty} \rho_k\right)} \qquad \mathrm{VAR}(\bar{\Psi}) = \frac{\mathrm{VAR}(\Psi)}{n_{eff}}$$

$$E\left[(\psi_i - \psi_{i-k})^2\right] = 2(1-\rho_k)\,\mathrm{VAR}(\psi)$$

$$V_{K} = \frac{1}{m(n-k)} \sum_{j=1}^{m} \sum_{i=k+1}^{n} \left(\psi_{i,j} - \psi_{i-k,j}\right)^2$$

VARIOGRAM AT LAG K

$$\boxed{\hat{\rho}_K = 1 - \frac{V_K}{2\,\mathrm{VAR}^+(\psi)}}$$

FOR LARGE K we have few samples => very noisy estimates

$$T = \min\{K \mid K \text{ is odd},\ \hat{\rho}_{K+1} + \hat{\rho}_{K+2} < 0\}$$

$$\sum_{K=1}^{\infty} \rho_K \approx \boxed{\sum_{K=1}^{T} \hat{\rho}_K}$$

$n_{\text{eff}} \simeq 100$

HAMILTONIAN MONTE CARLO

- $p(x) = \frac{1}{Z_x} \tilde{p}(x) = \frac{1}{Z_x} \exp(H_x(x))$

- INTRODUCE MOMENTUM VARIABLES y $|y| = |x|$

$$p(y) = \frac{1}{Z_y} \exp(H_y(y)) \quad , H_y(y) = -\frac{1}{2} y^T y \Rightarrow p(y) \text{ STANDARD GAUSSIAN}$$

$$p(x,y) = p(x)\,p(y) = \frac{1}{Z_x Z_y} \exp(H_x(x) + H_y(y)) = \frac{1}{Z} \exp(\underbrace{H(x,y)}_{\text{HAMILTONIAN}})$$

SAMPLE from $p(x,y)$ and forget y.

—

WE ARE IN POINT x_i

1) SAMPLE $y \sim p(y)$
2) we choose a random direction in time $(\{1,-1\})$
3) WE MOVE ACCORDING TO H.d from (x_i, y) TO A CANDIDATE (x', y') doing L steps
4) M-H ACCEPTANCE: accept if $H(x', y') > H(x,y)$, otherwise accept with prob $\exp(H(x',y') - H(x,y))$

$$H(x', y') = H(x, y)$$

$$(x', y') = (x + \Delta x, y + \Delta y)$$

F.O. TAYLOR EXPANSION

$$H(x+\Delta x, y+\Delta y) \approx H(x,y) + \underbrace{\nabla_x H_x(x)^T \Delta x + \nabla_y H_y(y)^T \Delta y}_{=0}$$

$$\Delta x = \varepsilon \nabla_y H_y(y) = -\varepsilon y$$

$$\Delta y = -\varepsilon \nabla_x H_x(x)$$

($\varepsilon = +\varepsilon_0$ or $-\varepsilon_0$ with prob $\frac{1}{2}$)

We do L steps of this dynamics to get the final (x', y')

There are better integration schemes: LEAP FROG integration

$$\text{M.H. } \alpha = \min \left\{ 1, \frac{P(x', y')}{P(x, y)} \right\}$$

$\hookrightarrow \exp\left(H(x', y') - H(x, y)\right)$