MONTE CARLO INFERENCE

$x^s \sim p(x|D)$

GENERATE UNWEIGHED SAMPLES FROM THE POSTERIOR ⟶ USE THEM TO COMPUTE STUFF EG MARGINALS, LIKELIHOOD OR POSTERIOR PREDICTIVE

STANDARD SAMPLING

- **INVERSE PROBABILITY TRANSFORM:** SAMPLE FROM UNIFORM $U \sim U(0,1)$, MAP TO DISTRIBUTION OF INTEREST WITH $F^{-1}(U)$
  - ⟶ **EXPONENTIAL:** $F^{-1}(u) = -\ln(u)/\lambda$
  - ⟶ **2D GAUSSIAN:** SAMPLE FROM $z_1, z_2 \in (-1,1)$. REJECT $z_1^2 + z_2^2 > 1$. TRANSFORM TO $x_1 = z_1\left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$, CHANGE OF VARIABLE FORMULA $p(x_1,x_2) = p(z_1,z_2)\left|\frac{\partial(z_1,z_2)}{\partial(x_1,x_2)}\right|$
  FOR MVG: $\Sigma = LL^T$ CHOLESKY DECOMP. L IS LOWER TRIANGULAR. SAMPLE x_s AS ABOVE ⟶ $y = L \cdot x + \mu$

REJECTION SAMPLING

- CREATE PROPOSAL DISTRIBUTION q WHERE $Mq(x) \geq \tilde{p}(x)$, M CONSTANT, $\tilde{p}(x)$ IS UNNORMALIZED p. SAMPLE $x \sim q(x)$, x LOCATION. SAMPLE $u \sim U(0,1)$ y LOCATION.
- ACCEPT IF $u > \frac{\tilde{p}(x)}{Mq(x)}$
- **ACCEPTANCE RATE** $p(\text{accept}) = \frac{1}{M}\int \tilde{p}(x)\,dx$ ⟶ PICK M AS SMALL AS POSSIBLE
- CAN USE WITH M = MLE; MARGINAL TO DRAW SAMPLE FROM POSTERIOR. OK IF PRIOR IS INFORMATIVE
- **PICKING q, p:** BOUND LOG DENSITY WITH PIECEWISE LINEAR FCN; AT FIXED GRID POINTS. ⟶ ENVELOPE IS PIECEWISE EXPONENTIAL. SAMPLE REJECTED ⟶ MAKE GRID TIGHTER.
- **IN HIGH DIM:** $M = \left(\frac{\sigma_q}{\sigma_p}\right)^D$ IS OPTIMUM., ACC. RATE IS $\frac{1}{M}$ CURSE OF DIM! ⟶ MCMC YO

IMPORTANCE SAMPLING

- APPROXIMATES INTEGRALS OF THE FORM $I = E[f] = \int f(x)p(x)\,dx$. SAMPLE FROM HIGH PROBABILITY REGIONS $p(x)$, WHERE $|f(x)|$ IS LARGE.
- **IS SUPER-EFFICIENT:** NEEDS LESS SAMPLES THAN EXACT DISTRIBUTION BECAUSE FOCUS ON RELEVANT PARTS OF THE SPACE
- SAMPLES FROM PROPOSAL; THEN $E[f] = \int f(x)\frac{p(x)}{q(x)}q(x)\,dx = \frac{1}{S}\sum_s^S w_s f(x^s) = \hat{I}$ ⟶ $\frac{p(x^s)}{q(x^s)} = w_s$ **IMPORTANCE WEIGHT**
- PICK q TO MINIMIZE VARIANCE OF ESTIMATE $\hat{I}$ ⟶ $q^*(x) = \frac{|f(x)|p(x)}{\int |f(x')|p(x')dx'}$. TRIAL & ERROR WHEN WE HAVE NO f IN MIX
- **ANCESTRAL SAMPLING:** FOR DAGS
  - **NO EVIDENCE** ⟶ ROOT, CHILDREN, CHILDREN OF CHILDREN, ... OK BECAUSE WE HAVE ORDERING.
  - **EVIDENCE:** EV. NODES ARE CLAMPED TO OBSERVED VALUES. DO ANCESTRAL BUT REJECT ALL SAMPLING IF INCONSISTENT VALUES. LONG SAMPLING
  - ⟶ **MORE EFFICIENT:** ONLY USE OBSERVED VALUES FOR OBS. VARS, NO SAMPLING $q(x) = \prod_t p(x_t | x_{pa(t)}) \prod_t \delta_{x_t^s}(x_t)$
    $w(x) = \prod_t p(x_t | x_{pa(t)})$
- **SAMPLE IMPORTANCE RESAMPLING:** DRAW ~~WEIGHTS~~ SAMPLES FROM $p(x)$ WITH E.S. $p(x) \approx \sum_s^S w_s \delta_{x^s}(x)$. THEN SAMPLE WITH REPLACEMENT FROM p PICKING WITH PROB w_s ⟶ $p(x) \approx \frac{1}{S'}\sum_{s'}^{S'} \delta_{x^{s'}}(x)$ • CAN USE FOR BAYESIAN INFERENCE IN LOW-DIM

PARTICLE FILTERING

DOES RECURSIVE BAYESIAN INFERENCE FOR DYNAMIC BAYESIAN NETWORKS; NONLINEARITY, NON-STATIONARITY

APPROXIMATE STATE TRAJECTORY: $p(z_{1:T}|y_{1:T}) \approx \sum_s \hat{w}_T^s \delta_{z_{1:T}^s}(z_{1:T})$, w_s ARE NORMALIZED WEIGHTS. BELIEF STATE UPDATED WITH IMPORTANCE SAMPLING. $w_T^s \propto \frac{p(z_{1:T}^s|y_{1:T})}{q(z_{1:T}^s|y_{1:T})}$, USUAL MARKOV ASSUMPTIONS ⟶ $w_t^s \propto w_{T-1}^s \frac{p(y_T|z_T^s)p(z_T^s|z_{T-1}^s)}{q(z_T^s|z_{T-1}^s, y_T)}$, CAN NOW APPROX POSTERIOR DENSITY $p(z_T|y_{1:T}) \approx \sum_s w_T^s \delta_{z_T^s}(z_T)$

DEGENERACY ⟶ AS t IS PF FAILS AFTER A FEW STEPS BECAUSE MOST PARTICLES HAVE NEGLIGIBLE WEIGHT. BECAUSE WE SAMPLE IN INCREASINGLY HIGH DIM SPACE. ESTIMATE USING VARIANCE OF WEIGHTS OVER STEPS. HIGH VAR ⟶ NO GOOD

FIXES:
- **RESAMPLING:** DROP PARTICLES WITH LOW WEIGHT AND REPLACE WITH GOOD ONES ACCORDING TO THEIR WEIGHTS. $O(S)$
  BUT NOW WE LOSE DIVERSITY ⟶ BOOTSTRAP, MCMC, OR SAMPLE NEW PARTS FROM A KDE ESTIMATE FOR SMOOTHING REGULARIZED PF
- **PROPOSAL DISTRIBUTION:** SAMPLE FROM PRIOR ⟶ $w_T^s \propto w_{T-1}^s p(y_T|z_T^s)$ **CONDENSATION ALGO.** NOT GOOD IF LIKELIHOOD IS NARROWER THAN PRIOR (SENSOR MORE INFORMATIVE THAN MOTION MODEL) ⟶ **SAMPLE FROM DATA**
- **OPTIMAL PROPOSAL:** $w_T^s \propto w_{T-1}^s \int p(y_T|z_T)p(z_T|z_{T-1}^s)\,dz_T$ CONDITIONAL ON OBS VALUES WEIGHT VAR IS 0
  GENERALLY UNTRACTABLE BUT OK FOR DISCRETE STATE SPACES OR GAUSSIAN.
  NOT GAUSSIAN ⟶ USE GAUSSIAN APPROXIMATION VIA UNSCENTED TRANSFORM

- SAMPLE FROM PROPOSAL, USUALLY PRIOR/PREV TIMESTEP
- NORMALIZE SAMPLE WEIGHTS $w_t = w_{t-1} \cdot$ NEW EVIDENCE WEIGHT
  $w_T^s = \frac{w_T^s}{\sum_s w_T^s}$
- $S = \frac{1}{\text{VAR}(w_i)}$
- RESAMPLE THOSE WITH LOW S

PF APPLICATIONS

ROBOT LOCALIZATION: DISCRETE OCCUPANCY GRID. COULD USE HMM BUT $O(K^2)$ SPACE IS LARGE. PF AS SPARSE APPROXIMATION.
VISUAL TRACKING: OK BUT SENSITIVE TO COLOR → MORE PARTICLES OR FRAMESKIP
TIME SERIES: NEUROSCIENCE STUFF

RAO-BLACKWELLIZED PF

TWO TYPES OF PARTICLE. Q_T; Z_T. WHENEVER Q_T, WE INTEGRATE Z_T OUT. WE SAMPLE Q_T AND $P(Z_T|Q_T)$ IS A DISTRIBUTION **DISTRIBUTIONAL PARTICLE**

REDUCES SAMPLING SPACE → REDUCES VARIANCE. • M/E FOR EACH PARTICLE • EXAMPLE: SLDS MODEL, USE KALMAN FOR PREDICTION → MIXTURE OF KALMANS

• **LOOK-AHEAD:** NEW WEIGHTS ARE INDEPENDENT FROM NEW Q_T VALUES. COMPUTE THEM, ASSES S FOR PARTICLES TO REPLACE. INITIALIZE RESAMPLES WITH ALREADY COMPUTED ~~WEIGHTS~~ POSTERIOR. $O(MS)$

• **APPLICATIONS:**
- MANEUVERING TARGET TRACKING · MARKOV CHAIN DISCRETE STATE. IE MISSILES.
- FASTSLAM · STANDARD KALMAN IS $O(L^3)$. SAMPLE ROBOT TRAJECTORY Q AND RUN L KALMAN 2D INSIDE EACH PARTICLE; $O(L)$ PER PARTICLE

MARKOV CHAIN MONTE CARLO

IDEA: CONSTRUCT MARKOV CHAIN WHOSE STATIONARY DISTRIBUTION IS TARGET DENSITY $p^*(x)$, PRIOR OR POSTERIOR.
BY PERFORMING RANDOM WALK WE DRAW CORRELATED SAMPLES AND WE CAN INTEGRATE WRT p^*

MCMC	VARIATIONAL METHODS
• EASIER TO IMPLEMENT	• FASTER FOR SMALL/MEDIUM PROBLEMS
• WORKS ON MORE MODELS	• DETERMINISTIC
• FASTER ON REALLY HUGE MODELS	• EASY TO KNOW WHEN TO STOP
	• GIVES LOWER BOUND ON LL

GIBBS SAMPLING

ANALOGUE OF COORDINATE DESCENT

• SAMPLE EACH VARIABLE IN TURN, CONDITIONED ON VALUES OF ALL OTHER VARS ON DISTRIBUTION. $x_1^{s+1} \sim p(x_1 | x_2^s, x_3^s)$; $x_2^{s+1} \sim p(x_2 | x_1^{s+1}, x_3^s)$; $x_3^{s+1} \sim p(x_3 | x_1^{s+1}, x_2^{s+1})$

• $p(x_t | x_{-t})$ IS FULL CONDITIONAL FOR t • IF $p(x)$ IS GRAPHICAL MODEL, FOR t WE ONLY NEED ITS MARKOV BLANKET

• NEED TO BURN-IN THE CHAIN, RUN AND DISCARD SAMPLES, TO MAKE IT ENTER ITS STATIONARY DISTRIBUTION / DELICATE BALANCE

GIBBS ON ISING

$p(x_t | x_{-t}, \theta) \propto \prod_{s \in N(t)} \psi_{st}(x_s, x_t)$ FULL CONDITIONAL: $p(x_t = +1 | x_{-t}, \theta) = \frac{\exp(J\eta_t)}{\exp(J\eta_t) + \exp(-J\eta_t)} = \text{sigm}(2J\eta_t)$ J COUPLING STRENGTH, $\eta = x_t(\underbrace{\lambda_t}_{\text{AGREE}} - \underbrace{D_t}_{\text{DISAGREE}})$

COMBINE W/ LOCAL EVIDENCE: $\psi_t(x_t) = \mathcal{N}(y_t | x_t, \sigma^2) \rightarrow p(x_t = +1 | x_{-t}, y, \theta) = \text{sigm}\left(2J\eta_t - \log \frac{\psi_t(+1)}{\psi_t(-1)}\right)$ • IS MORE OVERCONFIDENT THAN VARINF

GIBBS FOR INFERRING GMM PARAMS

SEMI-CONJUGATE PRIOR, SAME PRIORS FOR EACH MIXTURE COMPONENT

INDICATORS: $p(z_i = u | x_i, \mu, \Sigma, \pi) \propto N(x_i | \mu_u, \Sigma_u) \times \pi_u$ MIXING WEIGHTS: $p(\pi | z) = \text{Dir}(\{\alpha_u + \sum_i \mathbb{I}(z_i = u)\}_{u=1}^K)$

MEANS: $p(\mu_u | \Sigma_u, Z, X) = N(\mu_u | m_u, V_u)$ COVARIANCES $p(\Sigma_u | \mu_u, Z, X) = IW(\Sigma_u | S_u, V_u)$

ISSUE: LABEL SWITCHING: MODEL PARAMS θ AND INDICATORS z ARE UNIDENTIFIABLE. WHAT ONE SAMPLE CONSIDERS C1, ANOTHER CONSIDERS C2.

$O(NKD)$ → CANNOT AVERAGE OVER SAMPLES, NOT AN ISSUE IN EM OR VBEM SINCE THEY LOCK ON A SINGLE MODE

→ BEST NOT TO ASK QUESTIONS WITHOUT IDENTIFIABLE ANSWERS, IE CLUSTER MEMBERSHIP, BUT LIKE CLUSTER SAMENESS INSTEAD

COLLAPSED GIBBS

INTEGRATE SOME UNKNOWNS OUT AND SAMPLE THE REST. WE SAMPLE z WHILE INTEGRATING OUT θ. → WE CAN DRAW I.I.D. SAMPLES $\theta^i \sim p(\theta | z^i, D)$, LOWER VARIANCE THAN JOINT STATE SPACE.

RAO-BLACKWELL THEOREM: z, θ DEPENDENT RV, $f(z, \theta)$ SCALAR FCN; → $\text{VAR}_{z,\theta}[f(z,\theta)] \geq \text{VAR}_z[E_\theta[f(z,\theta)|z]]$ • VARIANCE WILL NEVER BE HIGHER THAN FULL SPACE

FOR A GMM: WE COLLAPSE ON μ_u, Σ_u, π, WE SAMPLE z. z_i, x_i BECOME INTERDEPENDENT BUT IS OK; AT EACH x REMOVE SUFF STATS, COMPUTE, UPDATE

IF CONJUGATE PRIORS → $p(x_i | D_{-i}, u)$ IN CLOSED FORM. • GENERALLY BETTER THAN VANILLA

GIBBS FOR HIERARCHICAL GLM

SO TO BORROW STATISTICAL STRENGTH $y_{ij} = x_{ij}^T w_j + \epsilon_{ij}$ • w_j COME FROM COMMON PRIOR $w_j \sim N(\mu_w, \Sigma_w)$. LEARN COMMON PARENTS

PRIORS FOR SHARED PARAMS $\mu_w \sim N(\mu_0, V_0)$; $\Sigma_w \sim IW(\eta_0, S_0^{-1})$; $\sigma^2 \sim IG(\nu_0/2, \nu_0 \delta_0^2/2)$

FULL CONDITIONALS: $p(w_j | D, \theta) \sim N$; $p(\mu_w | w, \Sigma_w) \sim N$; $p(\Sigma_w | \mu_w, w) \sim IW$; $p(\sigma^2 | D, w) \sim IG$

POSTERIOR PREDICTIVE: $E[y_i | x_i] = x_{ij}^T \hat{w}_j$ $\hat{w}_j = E[w_j | D] = \frac{1}{S}\sum_s w_j^s$ • VERY NICE REGULARIZATION

BUGS / JAGS

GIBBS IS COOL BECAUSE MANY MODELS. OPTIMIZER ONLY NEEDS MODEL SPECIFICATION (A DGM) AND METHODS FOR SAMPLING FROM FULL CONDITIONALS. BUGS/JAGS ARE TOOLS FOR THIS

IMPUTATION POSTERIOR

GIBBS WHERE WE SPLIT z_s AND θ_s. MCMC VERSION OF EM. $E \to I$; $M \to P$ IS DATA AUGMENTATION TECH.

BLOCKING GIBBS

IF VARS ARE SIGNIFICANTLY CORRELATED IT'LL TAKE A WHILE FOR MOVING AWAY FROM SPACE. → SAMPLE GROUPS OF VARS AT SAME TIME. FOR LARGER MOVES

METROPOLIS / HASTINGS

GIBBS IS LIMITED BECAUSE REQUIRES MARKOV STRUCTURE.

IDEA: WE PROPOSE A MOVE FROM x TO x' WITH $Q(x'|x)$ **KERNEL/PROPOSAL DISTRIBUTION**. COMMON CHOICE: $Q(x'|x) = N(x'|x, \Sigma)$ → RANDOM WALK METROPOLIS

- IF $Q(x'|x) = Q(x')$ → INDEPENDENCE SAMPLER
- AFTER MOVE IS PROPOSED, WE ACCEPT IT OR REJECT IT. **SYMMETRIC** $Q(x'|x) = Q(x|x')$, ACCEPT $R = \min\left(1, \frac{\tilde{p}(x')}{\tilde{p}(x)}\right)$
- SAMPLE $u \sim U(0,1)$ AND CHECK AGAINST R
- WE OCCASIONALLY ALLOW MOVES DOWNHILL TO LESS PROBABLE STATES, AND ALWAYS MOVE TO MORE PROBABLE

ASYMMETRIC: $R = \min(1, a)$; $a = \frac{\tilde{p}(x')/Q(x'|x)}{\tilde{p}(x)/Q(x|x')}$ **HASTINGS CORRECTION** ↳ PROPOSAL MIGHT FAVOR STATES INSTEAD OF TARGET

GIBBS IS SPECIAL CASE OF MH WHERE x_i IS FROM FULL CONDITIONAL, x_{-i} UNCHANGED, AND ACCEPTANCE RATE 100%.

- DOES NOT REQUIRE KNOWLEDGE OF Z BECAUSE THEY CANCEL OUT IN a

CHOOSING THE PROPOSAL DEPENDS ON IF NONZERO PROB OF MOVING TO NONZERO PROB TARGET SPACES. IN PRACTICE WE HAVE TO TUNE STUFF BECAUSE VAR TOO LITTLE → WE DON'T EXPLORE; VAR TOO LARGE → WE DON'T GET ENOUGH SAMPLE, CHUNKY HISTOS. PLAY AROUND WITH KERNEL WIDTH. **DO PILOT RUNS**. 25% - 40% ACCEPTANCE RATE IS GOOD

- **GAUSSIAN PROPOSALS!** INDEPENDENCE PROPOSALS OR RW PROPOSALS → $N(w'|w, s^2|H^{-1})$. SET s^2 AT $2.38^2/D$ FOR 23.4% ACC. RATE. CONTINUOUS STATE SPACES, USE HESSIAN TO DEFINE COVARIANCE.
- **MIXTURE PROPOSALS:** $Q(x'|x) = \sum_n^N w_n Q_n(w'|w)$
- **DATA-DRIVEN PROPOSALS** $Q(x'|x, D)$ DEPEND ON DATA AS WELL. SAMPLE (x, D) FROM FULL MODEL. TRAIN DISCRIMINATIVE CLASSIF FOR $p(x|f(D))$ (FEATURES). CAN DO ON JUST PARTS OF DATA STATE SPACE. $Q(x'|x, D) = \pi_0 Q_0(x'|x) + \sum_n^N \pi_n q_k(x'_n|f_n(D))$
  → **GENERATE AND TEST**: D/ PROPOSAL ARE TESTED ON POSTERIOR RATIO.

ADAPTIVE MCMC: ALLOW CHANGES OF PROPOSAL PARAMS AS ALGO RUNS. IE LARGE → SMALL COVARIANCE, XPLORATION → XPLOITATION. WATCH OUT NOT TO VIOLATE MARKOV PROPERTY. NO DEPENDENCY ON WHOLE CHAIN.

INITIALIZATION: IT HAS TO START TO WITH NONZERO PROBABILITY. TRICKY TO FIND IN CLOSED FORM. USUALLY FIND A LOCAL MODE W/ OPTIMIZER AND INITIALIZE THERE. DISCRETE STATE SPACES: MULTIPLE RESTARTS. CONTINUOUS SPACES: DO LOCAL EXPLORATION TO VISIT ENOUGH PROB MASS OF POSTERIOR

AUXILIARY VARIABLE MCMC

IMPROVES SAMPLING EFFICIENCY BY INTRODUCING DUMMY VARS TO REDUCE CORR. BETWEEN ORIGINALS. $\sum_z p(x,z) = p(x)$, $p(x|z)$ EASIER TO SAMPLE

EXAMPLE: LOGISTIC REGRESSION

$$\begin{cases} z_i = w^T x_i + \epsilon_i \\ \epsilon_i \sim N(0,1) \\ p(y=1) = \text{sigm}(w^T x_i) \end{cases}$$

APPROXIMATE LOGISTIC WITH SCALE MIXTURE OF STUDENT'Ts

$\epsilon_i \sim T(0,1,\nu) \rightarrow \epsilon_i \sim N(0, \lambda_i^{-1})$, $\lambda_i \sim Ga(\nu/2, \nu/2)$

- SUCH FORMULATION CAN RANGE BETWEEN PROBIT AND LOGIT, ESTIMATING ν
- CAN BE USED FOR SETTING STRENGTH OF REGULARIZER, 'BAYESIAN CV'

SLICE SAMPLING

UNIDIMENSIONAL, MULTIMODAL DISTRIBUTION. ADD AUX VAR TO MAKE LARGE MOVES. SAMPLE x^t, SAMPLE u^{t+1} UNIF. ON $[0, f(x^t)]$

SAMPLE x^{t+1} ON SLICE WHERE $f(x) \geq u^{t+1}$. **FOR MULTIVARIATE:** AUX VAR FOREACH DIM. DOES NOT NEED FULL CONDITIONALS (GIBBS) OR USER-SPECIFIED PROPOSAL (MH)

SWENDSEN-WANG

FOR ISING MODELS. SPEEDS-UP MIXING WHEN $J > 0$. FRUSTRATES SYSTEM FOR $J < 0$: EXPONENTIALLY MANY MODES.

ONE PER EDGE AUX. VARIABLES → **BOND VARIABLES**. SAMPLE FROM EXPAND MODELS AND THROW z AWAY / KEEPING x. BONDS ARE CI WITH NODES.

EASY FACTORIZATION. IF NODES BETWEEN EDGES ARE IN SAME STATE SET BOND ON WITH $p = 1 - e^{-2J}$, ASSESS CONNECTED COMPONENTS. FORCE NODES IN CC TO ASSUME SAME STATE, RANDOM.

HYBRID / HAMILTONIAN MCMC

CONTINUOUS STATE SPACES. WE CAN COMPUTE GRADIENT OF UNNORMALIZED LOG-POSTERIOR. IE NEURAL NETS.

AUX VARIABLES FOR 'MOMENTUM' OF PARTICLE | PARAMS. UPDATE POSITION/MOMENTUM. SET HOW MANY AND HOW BIG LEAPFROG STEPS.

ANNEALING

TEMPERATURE. IN THE BEGINNING DISTRIBUTION IS SMOOTHER.

SIMULATED ANNEALING

FINDS GLOBAL OPTIMUM OF BLACK-BOX FUNCTION f.

BOLTZMANN DISTRIBUTION: $p(x) \propto \exp(-f(x)/T)$ E IS ENERGY. T IS TEMP. $T \to 0$

- AT HIGH TS FCN IS APPROX.LY FLAT → EASY TO MOVE AROUND, THEN LARGE PEAKS LARGER AND SMALL (LOCAL) ONES DISAPPEAR.
- CONTINUATION METHOD → WE CAN TRACK THE PEAK AND FIND GLOBAL OPTIMUM
- SAMPLE ACCORDING TO PROPOSAL IE RW $x' = x_n + \epsilon_n$ → $\alpha = \exp(1(f(x) - f(x'))/T)$ → **ACCEPT** WITH PROB $\min(1, \alpha)$
- WE MIGHT STILL ACCEPT A LOCALLY WORSE STATE IF TEMP HIGH ENOUGH
- **COOLING SCHEDULE** IS CRITICAL. USUALLY EXPONENTIAL $T_n = T_0 C^n$, $T_0 = 1 \sim C \sim 0,8$

ANNEALED IMPORTANCE SAMPLING

SIMULATED ANNEALING + IMPORTANCE SAMPLING. DRAW SAMPLES FROM 'DIFFICULT' DISTRIBUTIONS.

START FROM EASY DISTRIBUTION AND GO OVER STEPS TO THE DIFFICULT $f_j(x) = f_0(x)^{\beta} f_n(x)^{1-\beta}$, $1 = \beta_0 > \beta_1 > \dots \beta_n = 0$

HAVE MARKOV CHAINS $T_j(x, x')$ LEAVING p_j INVARIANT.

SAMPLE $z_{N-1} \sim f_N \dots z_0 \sim T_1(z_1, \cdot)$; $x = z_0$ AND GIVE WEIGHT $w = \dfrac{f_{N-1}(z_{N-1}) \dots f_0(z_0)}{f_N(z_{N-1}) \dots f_1(z_0)}$

PARALLEL TEMPERING

RUN MULTIPLE CHAINS IN PARALLEL AT DIFFERENT TEMPERATURES. ALLOW CHAINS TO SAMPLE FROM NEIGHBORING ONES.

WHY MH WORKS

BECAUSE IT DEFINES A TRANSITION FUNCTION WHICH SATISFIES DETAILED BALANCE. → P^x IS STATIONARY DISTRIBUTION

CHAIN IS ERGODIC AND IRREDUCIBLE → P^x IS UNIQUE

REVERSIBLE JUMP (TRANSDIMENSIONAL MCMC)

WE SAMPLE IN SPACES OF DIFFERENT DIMENSIONALITY. EG MIXTURE MODELS W/ UNKNOWN NUMBER OF MIXTURES.

TROUBLE WHEN COMPUTING MH ACCEPTANCE RATIO → AUGMENT LOW-DIM SPACE WITH EXTRA RV SO MEASURE IS COMMON.

BURN-IN

MIXING TIME: TIME TO CONVERGE TO STATIONARY DISTRIBUTION. $\tau_\varepsilon = \max \tau_\varepsilon(x_0)$. DETERMINED BY $\gamma = \lambda_1 - \lambda_2$ EIGENVALS OF TRANSITION MATRIX

→ $\tau_\varepsilon \leq O\left(\frac{1}{\gamma} \log \frac{n}{\varepsilon}\right)$. n IS STATES. HARD FOR HIGH DIM/CONTINUOUS SPACES.

→ CONDUCTANCE OF CHAIN: MIN PROB OVER ALL SUBSETS OF STATES TO TRANSITION TO ITS COMPLEMENT. ϕ $\tau_\varepsilon \leq O\left(\frac{1}{\phi^2} \log \frac{n}{\varepsilon}\right)$

TOO ~~HIGH~~ LOW CONDUCTANCE → MCMC NOT GOOD

DIAGNOSTICS: TRACE PLOTS. START FROM OVERDISPERSED STARTING POINT AND ASSESS CONVERGENCE TO SAME DISTRIBUTION OF SOME VARS OF INTEREST

ESPR: ESTIMATED POINT SCALE REDUCTION. VARIANCE WITHIN CHAIN VS VARIANCE ACROSS CHAINS. THINK ANOVA. B, W.

$\hat{V} = \frac{s-1}{s} W + \frac{1}{2} S$ UNBIASED OVER STATIONARITY, OVERESTIMATES IF OVERDISPERSION. $R = \sqrt{\frac{\hat{V}}{W}}$ • MEASURES HOW MUCH POSTERIOR VARIANCE WOULD DECREASE IF SAMPLING CONTINUES.

ANOVA ANALYSIS BETWEEN/WITHIN CHAINS

$R \approx 1$ ✓

ACCURACY

MCMC PRODUCES AUTOCORRELATED SAMPLES USE AUTOCORRELATION FCN $\text{VAR}_{MCMC} = \text{VAR}_{MC}(\bar{f}) + \frac{1}{s^2} \sum E[(f_s - f^x)(f_t - f^x)]$

→ USE THINNING = SAMPLE SKIPPING • WAS THIS GIBBS IS BETTER

IMPORTANCE OF CORR — DEPENDS ON CORR

→ $S_{EFF} = \frac{\text{VAR}_{MC}(f)}{\text{VAR}_{MCMC}(f)}$

CHAINS

RUN 'MEDIUM' NUMBER OF CHAINS OF 'MEDIUM' LENGTH. DISCARD FIRST HALF.

3 100ish

ML APPROXIMATION (MCMC CHAPTER)

$P(D|M) = \int P(D|\theta, M) P(\theta|M)\, d\theta$ IS OFTEN INTRACTABLE IF NON-CONJUGATE PRIOR / HIDDEN VARS.

CANDIDATE METHOD $P(\theta|M) = \dfrac{P(D|\theta,M)P(\theta|M)}{P(\theta|D,M)}$ $\forall \theta$. NUMERATOR IS USUALLY GIVEN. APPROX DENOMINATOR WITH MCMC SHAKY ASSUMPTION THAT DENOM. HAS MARGINALIZED OVER ALL MODES
→ POOR RESULTS

HARMONIC MEAN SAMPLING

$1/P(D) \simeq \frac{1}{S} \sum \frac{1}{P(D|\theta^s)}$; $\theta^s = P(\theta|D)$ HARMONIC MEAN OF DATA LIKELIHOOD UNDER EACH SAMPLE. THEORETICALLY CORRECT BUT **WORST MONTE-CARLO METHOD EVER**
→ ONLY DEPENDS ON POSTERIOR SAMPLE
→ OFTEN INSENSITIVE TO PRIOR.

ANNEALED IMPORTANCE SAMPLING

TO EVALUATE RATIO OF PARTITION FUNCTIONS $Z_0 = \int f_0(x)dx = \int f_1(z)dz$, $Z_N = \int f_N(x)dx = \int g(z)dz$.

$\frac{Z_0}{Z_N} = E_{\hat{q}}\left[\frac{f(z)}{g(z)}\right] \approx \frac{1}{S}\sum_1^S w_s$ • IF F_N PRIOR AND F_0 POSTERIOR $Z_N = P(D)$ GIVEN WE KNOW PRIOR NORM. CONSTANT Z_0

HAMILTONIAN MCMC

- IDEA: USE HAMILTONIAN DYNAMICS TO PRODUCE MCMC PROPOSALS → FASTER STATE SPACE EXPLORATION THAN WITH RANDOM WALKS
  - MOMENTUM VARIABLES - VOLUME PRESERVING NO NEED FOR JACOBIANS

HAMILTONIAN DYNAMICS

- Q POSITION - p MOMENTUM - HAMILTONIAN $H(p,q)$, $\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$, $\frac{dp}{dt} = -\frac{\partial H}{\partial q_i}$ - $z = (q,p)$ VECTOR → $\frac{dz}{dt} = J \nabla H(z)$

ENERGY: $H(q,p) = U(q) + K(p)$ - $U(q) = -\log \pi(q) + K$, q DISTRIBUTION
POTENTIAL KINETIC - $K(p) = p^T M^{-1} p / 2$, M MASS; SYMMETRIC POSDEF, DIAGONAL; OFTEN MULTIPLE OF I, $-\log p$ OF 0-MEAN GAUSSIAN WITH COVAR M IS $N \sim (0,1)$ FOR q

- $\frac{dq}{dt} = M^{-1} p$
- $\frac{dp}{dt} = \frac{\partial U}{\partial q_i}$
- IN 1D: $H(p,q) = U(q) + K(p)$, $U(q) = q^2/2$ $K(p) = p^2/2$; $\frac{dq}{dt} = p$ $\frac{dp}{dt} = -q$; $Q = R\cos(t+a)$, $p = -R\sin(a+t)$

PROPERTIES:
- REVERSIBILITY: CAN GO BACK FROM T TO $T-A$, IS INVERTIBLE. LEAVE DISTRIBUTION INVARIANT (MARKOV)
- CONSERVATION H IS PRESERVED, INVARIANT, IN MH ACCEPTANCE PROB = 1 IF INVARIANT. IN PRACTICE APPROXIMATELY INVARIANT
- VOLUME PRESERVATION IN (q,p) SPACE → GOOD BECAUSE NO NEED TO JACOBIAN. DIVERGENCE = 0 → PRESERVES VOLUME. $\det(J) = 1$ PRESERVES VOLUME
- SYMPLECTICNESS $z = (q,p)$, $B^T J^{-1} B = J^{-1}$ IMPLIES VOL CONSERVATION STRONGER CONDITION

HOW TO DISCRETIZE?

- EULER'S METHOD → POOR CONVERGENCE
- MODIFIED EULER
  - $K(p) = \sum_i^d \frac{p_i^2}{2m_i}$ →
$$\begin{cases} p_i(t+\varepsilon) = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}(q(t)) \\ q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{p_i(t+\varepsilon)}{m_i} \end{cases}$$
  - CAN ALSO DO q BEFORE p
  - BETTER BECAUSE EXACT VOLUME PRESERVATION
  - USE UPDATED p VALUE → À LA GIBBS
- LEAPFROG METHOD → EVEN DEEPER →
$$\begin{cases} p(t+\varepsilon/2) = p_i(t) - (\varepsilon/2) \frac{\partial U}{\partial q_i}(q(t)) \\ q(t+\varepsilon) = q_i(t) + \varepsilon \frac{p(t+\varepsilon/2)}{m_i} \\ p(t+\varepsilon) = p_i(t+\varepsilon/2) - (\varepsilon/2) \frac{\partial U}{\partial q_i}(q(t+\varepsilon)) \end{cases}$$
  - DO HALFWAY STEP INCREMENT FOR p
  - VOL PRESERVING; STABLE TRAJECTORIES

LET'S DO SOME MCMC NOW

- DENSITY → POTENTIAL ENERGY TO SAMPLE
- MOMENTUM FOR ORIGINAL VARIABLES
- $U(q) = -\log[\pi(q) L(q|D)]$ POSTERIOR FROM MODEL PRIOR π; LIKELIHOOD L
- SIMULATE MC RESAMPLING MOMENTUM AT EACH STEP
- MH UPDATE DRAWN FROM HAMILTONIAN PROPOSAL
- CANONICAL DISTRIBUTION $P(x) = \frac{1}{Z} \exp[-E(x)/T]$ → $P(q,p) = \frac{1}{Z} \exp[-H(p,q)/T]$
  E ENERGY FOR STATE x
  → $E(x) = -\log P(x) - \log(Z)$ → $P(q,p) = \frac{1}{Z} \exp[-U(q)/T] \exp[-K(p)/T]$
  JOINTS ARE FAVOURED!
- p, q INDEPENDENT
- ε VARIABLES, p NEEDED TO MAKE IT WORK
  1 MOMENTUM PER DIMENSION
- MOVES ON HYPERSURFACE OF CONSTANT PROBABILITY DENSITY

- G IS DISTRIBUTION OF INTEREST; WE CHOOSE DISTRIBUTION OF p VIA FORM OF K
  → USUALLY A QUADRATIC FCN $K(p) = \sum_i^d \frac{p_i^2}{2m_i}$ → p IS 0 MEAN MVN, m_i IS VARIANCE
- STEP 1: DRAW p_i FROM THEIR MVNS, INDEPENDENTLY OF q. → JOINT NOT CHANGED. p_i INDEPENDENT
- STEP 2: DO A METROPOLIS UPDATE USING $H(q,p)$ AS PROPOSAL. RUN HAMILTONIAN TRAJECTORY FOR L STEPS, STEPSIZE ε HYPERPARAMS LEAPFROG METHOD
  - NEGATE MOMENTUM VARIABLES AT END: L. ← ACCEPT WITH PROB $\min\left[1, \exp\left(-H(q^*,p^*) + H(q,p)\right)\right]$
    ↳ MAKES PROPOSAL SYMMETRICAL; NOT DONE BECAUSE WERE NOT RUNNING THE CHAIN FOR > 1 STEPS. NEXT ITERATION WILL RESAMPLE p_i'S
  - $H(p,q)$ ALMOST UNCHANGED - Q CAN END UP IN DIFFERENT PLACES; RESAMPLE OF MOMENTUMS IS NECESSARY OTHERWISE $U(q)$ WOULD NEVER BE UNDER THE STATIONARY $H(p,q)$
  - REVERSIBILITY → SATISFIES DETAILED BALANCE CONDITIONS — CANONICAL DISTRIBUTION UNCHANGED/INVARIANT - ERGODICITY → WON'T GET STUCK IN REGION OF STATESPACE UNLESS PROBABILITY IN SOLUTIONS → 2 MANY L OR ε

→ BEHAVIOR IS VERY DIFFERENT FROM RANDOM WALK BECAUSE GRADIENTS/MOMENTUMS, TUNE BY MAGNITUDE OF STDDEV.
→ IS GOOD BECAUSE MOVES ON MANIF IN CONSISTENT DIRECTIONS. BETTER ACCEPTANCE PROBABILITY. BETTER EXPLORATION
→ IMPROVEMENT WRT RW IS RATIO OF MAX/MIN VARIANCES OF p_i
- RW → DISTANCE OF ACCEPTED GROWS WITH $\sqrt{N}$ OF ITERATIONS

- CAN USE DIFFERENT K FUNCTIONS/TRANSFORMATION TO MAKE IT EFFICIENT/TRANSFORMATION INVARIANT/...
- DIFFERENT 'RULE OF THUMB' VALUES THAN VANILLA MCMC
- SETTING L AND ε IS TRICKY, TRIAL & ERROR. CAN HAVE DIFFERENT ε FOR DIFFERENT p,q PAIRS/DIMENSIONS
- SCALES BETTER IN INCREASING DIM SPACES
- FOR NEURAL NETS: CAN ALTERNATE HMC STEPS FOR PARAMS, AND STEPS FOR HYPERPARAMS

EXTENSIONS

- DISCRETIZATIONS OTHER THAN LEAPFROG METHOD → H SPLIT IN OTHER WAYS
- LANGEVIN M.C: EQUIVALENT TO HMC WITH 1 LEAPFROG STEP, CAN NOT HAVE EXACT p_i', IMPLY THEM BY MAKING q_i IID AND SAMPLE FROM GAUSSIAN.
- REFRESH MOMENTUM: $p' = \alpha p + (1-\alpha^2)^{1/2} N$. N IS MVN VECTOR. REDUCE MOMENTUM AT EACH ITERATION; BUT SOFT, GRADUAL MOVE
- ACCEPT / REJECT WINDOWS, HAVE ACCEPTANCE DEPEND ON 'WINDOWS' OF STATES TOO CLOSE/OK
- USE APPROXIMATIONS FOR $U(x)$
- DO STUFF TO ϵ ADAPTIVELY
- TEMPERING. EITHER USE EXPLICIT T, CONTINUATION METHOD, OR MULTIPLY/DIVIDE MOMENTUM VARIANCES IN FIRST/SECOND HALF OF LEAPFROG TRAJECTORY. STEP $(t+\epsilon/2)$ PRESERVES VOLUME KEEP IT ABOVE 1