classical approaches for inverse inference

- **conjugate priors**: choose $p^s(Y)$ and $p^s(X|Y)$ such that $p^s(Y|X)$ can be analytically calculated and is in the same distribution family as $p^s(Y)$ (⇒ incremental Bayesian updating)
  - advantage: efficient and mathematically elegant
  - disadvantage: very unrealistic ⇒ big simulation gap
- **likelihood-based inference**: $p^s(Y)$ and $p^s(X|Y)$ are known (and not just $X=\phi(Y,\eta)$), but $p^s(Y|X)$ is intractable
  ⇒ create a sample $\{Y_k \sim p^s(Y|X=x^{obs})\}_{k>1}^{K}$ using Markov chain Monte Carlo (MCMC) or a variant thereof (HMC)

important relaxation: $p^s(X|Y)$ can be unnormalized, $Z=\int p^s(x|y)\,dx$ intractable

e.g. Gibbs distribution: $p^s(X|Y,T) = \frac{1}{Z}\exp\left(-H(X,Y)/T\right)$

T: temperature
Z: unknown partition function
$H(X,Y)$: energy of state X with parameters Y

basic MCMC algorithm

⓪ - x^{obs} given, $y^{(0)}$ arbitrary initial guess

- pick a "proposal distribution" for transition probs $q(y' \mid y^{(t-1)}, x^{obs})$
  e.g. Gaussian (something easy to work with)

① for $t = 1, \dots, T$

ⓐ $y' \sim q(y' \mid y^{(t-1)}, x^{obs})$ sample proposal

ⓑ calculate acceptance weight $\alpha = \dfrac{p^s(x^{obs} \mid y')\, p^s(y')}{p^s(x^{obs} \mid y^{(t-1)})\, p^s(y^{(t-1)})} = \dfrac{p^s(y' \mid x^{obs})}{p^s(y^{(t-1)} \mid x^{obs})}$

(intractable posteriors)

(evidence $p^s(x^{obs})$ and partition fcl. 's cancel out!)

ⓒ sample $u \sim \text{uniform}(0,1)$ acceptance threshold

ⓓ if $u \le \alpha$: accept $y^{(t)} = y'$

else : reject $y^{(t)} = y^{(t-1)}$

Thm: for $T \to \infty$, $\{y^{(t)}\}_{t=1}^{T} \sim p^s(y \mid x^{obs})$

classical solutions to SBI (contd.)

- analytic: conjugate priors
- likelihood-based inference: $p^s(Y)$ and $p^s(X|Y)$ are known, but posterior $p^s(Y|X)$ is intractable

⇒ MCMC and its variants: create sample $\{Y^{(t)} \sim p^s(Y|X^{obs})\}_{t=1}^{T}$ via a clever Markov chain $Y^{(t)} \sim p(Y|Y^{(t-1)}, X^{obs})$

- advantages:
  - very general, lot of mathematical theory (performance guarantees)
  - implemented in "probabilistic programming languages", e.g. STAN and libraries (pyro, pyMC, Edward?, Turing.jl)
- disadvantages:
  - not amortized, algorithm runs from scratch for each new X^{obs}
  - expensive – often, T must be very large for complete "mixing", i.e. until all modes of $p^s(Y|X^{obs})$ have been covered
    – often, a long "burn-in" phase is need to forget a bad initial guess $Y^{(0)}$ ⇒ throw away $Y^{(0)} \dots Y^{(T_0)}$ ← hyperparameter or. computed
    – samples $Y^{(t)}$ and $Y^{(t-1)}$ are close to each other

(even $y^{(t)} = y^{(t-1)}$ when y' is rejected)
may bias derived statistics of the chain
$\Rightarrow$ skip k samples in the chain $y^{(t)} \rightsquigarrow y^{(t+k)} \rightsquigarrow y^{(t+2k)} \rightsquigarrow \dots$
- often difficult to define proposal distribution $p(y' \mid y^{(t-1)})$ that has low rejection rate

$\Rightarrow$ MCMC only applicable when $\dim(y)$ is not large and $p^s(x|y)$ not too slow
making MCMC faster is hot research

- <u>likelihood-free inference</u>: $p^s(x|y)$ unknown, only implicitly defined as $\phi_\# p^s(y,\eta)$
only samples $x = \phi(y,\eta)$ with $y,\eta \sim p^s(y,\eta)$

Approximate Bayesian Computation (ABC) $\hat{=}$ brute force
- requires a distance $\text{dist}(x, x^{obs})$ for outcomes, usually hand-crafted so far
alg: ⓪ $t=0$, sample $\{\}$ (we later become $\{y^{(t)} \sim p^s(y|x^{obs})\}_{t=1}^{T}$)
① repeat until $t=T$
ⓐ sample $y \sim p^s(y)$ (or $y,\eta \sim p^s(y,\eta)$), simulate $x \sim \phi(y,\eta)$
ⓑ if $\text{dist}(x, x^{obs}) \leq \varepsilon$: sample.add(y), $t = t+1$; else reject

- advantage: • works when MCMC doesn't
  • available in libraries (pyABC)
- disadvantage • very slow: if ε (hyperparameter) is small,
  $\{y^{(t)}\}_{t=1}^{T}$ is a good representation of $p^s(y|x^{obs})$
  but rejection rate is very high and vice versa
  no amortization, wasted simulation budget
  • hard to define $dist(x, x^{obs})$ if x is complicated or $dim(x)$ is high
  ⇒ instead design hand-crafted summary statistics $h(x)$ and compare
  ⇒ only feasible when $dim(y)$ is not too high $dist(h(x), h(x^{obs}))$