MCMC & HMC

Assume BNN model: $p(w) = \mathcal{N}(0, I)$
$p(y|X,w) = \prod_n \mathcal{N}(y_n | f(x_n, w), \sigma^2)$

Goal: Draw samples $w^s \sim p(w|X,y)$
for our BNN

Using these samples can compute $f(x_*, w^s)$
~~for~~ each sample,

goal

$$\left.\begin{array}{l} \mathbb{E}_{p(w|X,y)}[f(x_*, w)] \\ \mathrm{Var}[f(x_*, w)] \end{array}\right\}$$ estimate via empirical methods on S samples

---

Common problem in Bayesian inference:

sample from posterior distribution
given model and data
- likelihood density
- prior density

we only know posterior density up to constant

$$p(w|X,y) \underset{\underset{\text{unknown}}{\frac{1}{p(y|x)}}}{\propto} \underbrace{p(y|X,w)}_{\text{lik}} \underbrace{p(w)}_{\text{prior}}$$

Idea: w_0 starting value
Apply stochastic transition repeatedly

$$w_0 \xrightarrow{T} w_1 \xrightarrow{T} w_2 \xrightarrow{T} w_3 \ldots w_S, w_{S+1}, \ldots$$

if we choose transition well, can we say these match posterior samples

What is MCMC?

This is a <u>Markov Chain</u> b/c w_t only depends on w_{t-1}

knowing $w_{t-2}, \ldots w_1$ doesn't tell us any more about $p(w_t)$

This is <u>Monte Carlo</u> b/c transitions are <u>random/stochastic</u> not deterministic.

Markov chain theory says that if a transition operator is

(1) irreducible : good at exploring
pos. proba. of visiting all states

(2) aperiodic : no cycles possible $A \rightarrow B \rightarrow C \rightarrow A$

then after many applications of T, we will reach a <u>stationary</u> distrib.

Goal of MCMC: construct T so stationary distr. is the posterior!

How? One way: make T satisfy detailed balance

$$p(w_t)\, T(w_{t+1} \mid w_t) = p(w_{t+1})\, T(w_t \mid w_{t+1})$$

can show this joint distr. leaves $p(w_{next})$ equal to $p(w_{current})$. transitions produce draw from <u>stationary</u>

Metropolis algorithm $\quad T: \text{Normal}(w_t \mid \overset{\text{mean}}{w_{t-1}}, \overset{\text{cov}}{\sigma I})$

<u>symmetric</u> transition <u>proposals</u>

Given w_0
for t in $1, 2, \ldots$

$\quad w^{prop} \sim N(\overset{\text{mean}}{w_{t-1}}, \sigma^2)$

$\quad \text{accept\_proba} \leftarrow \min\left(1, \dfrac{p(w^{prop})}{p(w_{t-1})}\right)$

$\quad$ if rand() < accept_proba:
$\qquad w^t \leftarrow w^{prop}$
$\quad$ else
$\qquad w^t \leftarrow w^{t-1}$

Exercise: write out accept ratio for our <u>BNN</u> model

HMC overview

Want: Transition proposal that explores "typical set" efficiently

Idea: Use gradients of target distr. in proposal

One idea:

$$T(w^{t+1} / w^{t}) = \text{Normal}\left(w^{t} + \epsilon \nabla_{w} p(w^{t}) \,\middle|\, \sigma^2 I \right)$$

but this would be <u>mode</u> <u>seeking</u> not a good sampler

We need better ideas.

Motivation from Physics

add "momentum" r.v. $p \sim N(0, I)$ same size as w

$$\text{Hamitonian}(p, w) = \overbrace{-\log p(w)}^{\text{potential}} + \overbrace{-\log p(p)}^{\text{kinetic}}$$

Sample from joint (p,w) then drop p to get a sample of w

Can (approximately) evolve position & momentum while keeping the Hamitonian (energy) conserved

$$\frac{\partial w}{\partial t} = \frac{\partial}{\partial p}\left[-\log p(p) \right]$$

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial w}\left[-\log p(w) \right]$$

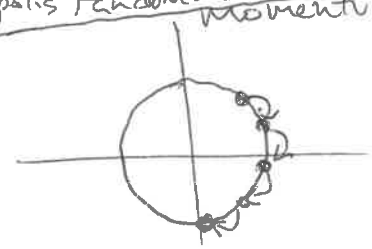
gradient of our ~~target~~ posterior

If we can run a simulation of these eqs: $p_0\ p_1 \ldots p_T$, $w_0\ w_1 \ldots w_T$

Cool way to explore level sets of joint distr over p, w

Accept just like Metropolis random walk

Betancourt Fig. 22 & Neal Fig 1:

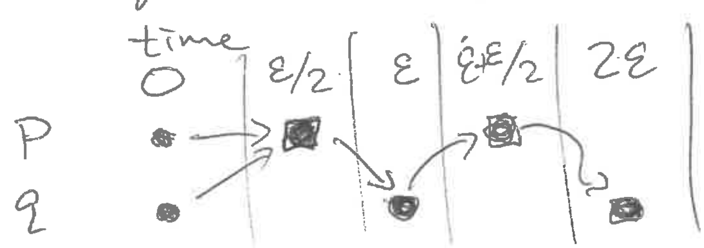


$$p(t+\tfrac{\epsilon}{2}) = p(t) - \tfrac{\epsilon}{2} \nabla_w \log p(w)$$

$$w_{t+\epsilon} \leftarrow w_t + \epsilon\, p(t+\tfrac{\epsilon}{2})$$

$$p_{t+\epsilon+\frac{\epsilon}{2}} \leftarrow p_{t+\frac{\epsilon}{2}} - \epsilon \nabla_w \log p(w)$$

Challenge: numerical approx of the <u>differential</u> eqns. (Not a goal of this course.)



each step only <u>one</u> var changes using delta from other var

Demos + For Next Time

Challenge: with partner, Write calc_potential_energy function for HMC for BNN model

Demos: (((while working, open projector)))

interactive random walk vs. HMC

"Harlem Shake" with MCMC

HW2:

Work with peers! Get help early (at office hours)

use code from class on Tues (autograd + NNs)

break down into parts

simplify! could I make a random walk sampler?

could I do this sampler for linear regression?

reading: black box VI + blog post

Can I give high level punchline for each figure?

Could I use Alg 1 pseudocode to train BNN?

— Skim: lots of extensions (ctrl variates, etc)