Review: VI has us pick a density family $q(w|v)$ w/ parameters $v \in \mathcal{V}$
We then optimize an objective $\mathcal{L}(v)$

	BBVI ~~"score function trick"~~	BbB ~~"log derivative trick"~~
how to compute $\mathcal{L}(v)$? (what identity/technique)	monte carlo	monte carlo
how to compute $\nabla_v \mathcal{L}(v)$	score function trick	~~log derivative~~ reparameterization trick
write formula for it	$\nabla_v \mathcal{L}(v) \approx \frac{1}{S}\sum_{s=1}^{S}[\nabla_v \log q(v)](\log p(w^s, y\|x) - \log q(w^s))$	
do we need log prior to be differentiable	X	✓
log lik	X	✓
log q_v	~~X~~ ✓	✓

$v = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$ w/ mean μ, variance σ

Idea: if $q(w|v)$ is Gaussian, consider two ways to sample from it

(1) draw $w \sim \text{Normal}(\mu, \sigma^2)$

(2) draw $\varepsilon \sim N(0, 1)$ — standardized r.v.
$w \leftarrow \sigma\varepsilon + \mu$ — plus a transform

Can prove these two procedures produce same distribution.

Key idea: define $w \triangleq t(v, \varepsilon)$
deterministic mapping of parameters v and a std. r.v. ε w/ distr. $p(\varepsilon)$

Now, for any function $f(w, v)$ that depends on w and v,

we can rewrite expectations

$$\mathbb{E}_{q(w|v)}[f(w,v)] = \mathbb{E}_{p(\varepsilon)}[f(t(v,\varepsilon), v)]$$

why? now gradient w.r.t. v can go inside

$$\nabla_v \mathbb{E}_{q(w|v)}[f(w,v)] = \nabla_v \mathbb{E}_{p(\varepsilon)}[f(t(v,\varepsilon), v)]$$
$$= \mathbb{E}_{p(\varepsilon)}[\nabla_v f(t(v,\varepsilon), v)]$$
chain rule →

chain rule ↘

$$\nabla_v \mathbb{E}_{q(w|v)}[f(w,v)] = \mathbb{E}_{p(\varepsilon)}\left[\frac{\partial f(w,v)}{\partial w}\frac{\partial w}{\partial v} + \frac{\partial f(w,v)}{\partial v}\right]$$

Given S samples, can compute as

$$= \frac{1}{S}\sum_s \left.\frac{\partial f(w^s,v)}{\partial w}\right|_{w=w^s} \left.\frac{\partial w}{\partial v}\right|_{w=w^s} + \frac{\partial f(w(\varepsilon))}{\partial v}$$

‡Illustration:

What is gradient wrt μ of expected value of 1D normal?

Very easy w/ reparameterization

First way:

$$\nabla_\mu \mathbb{E}_{w\sim N(\mu,\tau^2)}[w]$$

$$= \nabla_\mu \int p(w|\mu,\tau^2)\, w\, dw$$

$$= \nabla_\mu \int \frac{1}{\sqrt{2\pi}}\frac{1}{\tau} e^{-\frac{1}{2\tau^2}(w-\mu)^2} w\, dw$$

HARD!

Second way:

w can be parameterized as $\varepsilon\sim N(0,1)$, $w = \mu + \tau\varepsilon$

$$\nabla_\mu \mathbb{E}_{\varepsilon\sim N(0,1)}[\mu+\tau\varepsilon]$$

$$= \nabla_\mu \mu + \nabla_\mu \underbrace{\mathbb{E}[\tau\varepsilon]}_{\text{const wrt } \mu} \to 0 = 1$$

Algorithm: Applying BbB to BNNs

What we want to optimize:

$$\max_{v \in V} \mathcal{L}(v), \quad \mathcal{L}(v) = \frac{1}{S} \sum_s \log p(y, w^s) - \log q_v(w^s | v)$$

$q_v(w|v) = \prod_j N(w_j | \mu_j, \tau_j)$

$V_j = \begin{bmatrix} \mu_j \\ \tau_j \end{bmatrix}$ mean can be any real; stddev must be positive

wait! will this work for SGD?

want unconstrained parameter — could do $s \triangleq \log(\tau)$, $\rho \triangleq \log(1+e^\tau)$

Using S samples, can compute gradient as

$$\nabla_{\mu_j} \mathbb{E}\left[\log p(y,w) - \log q(w|\mu,\tau) \right]$$

$$= \mathbb{E}_{p(\varepsilon)}\left[\nabla_{\mu_j} \log p(y, w(\varepsilon)|x) - \log q(w(\varepsilon)|\mu,\tau) \right]$$

f

$$= \frac{1}{S} \sum_{s=1}^{S} \frac{\partial f}{\partial \mu} + \frac{\partial f}{\partial w} \frac{\partial w}{\partial \mu}$$

$$\nabla_\rho \mathbb{E}\left[\log p(y,w) - \log q(w|\mu,\rho) \right]$$

can be computed in same way

Alg: Init: means μ, log stddevs s

for each iter:

$\mu \leftarrow \mu - \alpha \nabla_\mu \mathcal{L}$

$\rho \leftarrow \rho - \alpha \nabla_\rho \mathcal{L}$

need many iters

need to tune step size

Methods:

All these can produce samples from approx. posterior which we can use for prediction

$$p(y_* | x_*) = \int p(y_* | w, x_*) p(w | x, y) dw$$

requires	HMC	BBVI	Bayes By Backprop
diff'able log prior PDF	✓		✓
diff'able log lik PDF	✓		✓
a chosen "q" approx posterior density	N/A	✓	✓
diff'able log PDF for q		✓	✓
q must be Normal (at least, for purposes of this class)			✓
estimated gradients		BIG variance	smaller variance

defined by param v,

For any $q(w|v)$, we draw S samples $w^1, w^2, \dots w^S$

BBVI gradient estimator

$$\nabla_v \mathcal{L}(v) = \frac{1}{S} \sum_{s=1}^{S} \left(\nabla_v \log q(w^s | v) \right) \left(\log p(y, w^s | x) - \log q(w^s | v) \right)$$

BbB gradient estimator

Assume $q(w|v)$ is Normal(μ_v, σ_v). Draw S samples $\varepsilon^1 \dots \varepsilon^S$ from $N(0,1)$.

$$\nabla_v \mathcal{L}(v) = \frac{1}{S} \sum_{s=1}^{S} \nabla_v \left[\log p(y, w(\varepsilon^s, \mu_v, \sigma_v)) - \log q(w(\varepsilon^s, \mu_v, \sigma_v) | v) \right]$$