Auto-encoding Variational Bayes
Kingma (2013)

how to perform EFFICIENT inference & learning in directed probabilistic model?
→ whilst in the presence of:
- continuous latent variables w/ intractable posterior distributions
- large datasets

notes gem

A intractable posterior refers to a sctn where the post prob dist cannot be computed in a simple or comp viable way

Cannot:
- exact prob of specific para values
- mean, var & sum stats
- Samples from the dist

recall Bayes: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$

($P(\theta|D)$: Posterior distribution; $P(D|\theta)$: likelihood; $P(\theta)$: Prior; $P(D)$: evidence; Sum over poss Parameter values)

the main reason these problem become impossible is because of the Denominator/evidence $P(D)$

(1) no exact evidence to normalize the numerator and obtain posterior
(Den is constant & num is prop to This)

(2) high-Dims → many paras/latent vars either impossible or comp exhausting to calc/APPOX

Issues w/ intractability:

1. cannot sample from dist
2. calc point estimates
3. Compare between models (need $p(D)$)

Back to paper...

Paper introduces a stochastic variational inference & learning Algo

- ↳ Scales to large datasets
- ↳ works w/ intractable case

this approach is for models w/ cont latent variables and/or parameters

optimized w/ stochastic gradient ascent ~~descs~~

I.I.d = independent & identically distributed

- ↳ RNNs & LSTMs handle non-I.I.d

~~this~~

this paper introduces the "Stochastic Gradient variational Bayes" (SGVB)

then a specific application called "Auto-encoding variational Bayes" for scenarios where dataset is I.I.D but w/ continuous latent variables + (common & important scenario)

Variational Bayes optimiz of a Approx to the intractable posterior (not poss)

Paper shows how re-para of variational lower Bounds yields a simple unbias estimator of the lower Bound (SGVB)

↳ Allows eff approx posterior inference in almost any model w/ cont latent varcables and/or paras

For i.i.d uses Auto-encode VB (AEVB)
→ uses SGVB to make inference eff
→ Approx post inference use simple Ancestral Sampl to learn model paras
→ Avoid expensive methods such as Monte carlo

the learned approx posterior INF Model can be used for tasks: recognition, denoising representation & VIS

~~Very~~ when NN is used for recognition = variational Auto encoder

---

SGVB is the foundational trick introduced

Purpose is to solve prob of intractable exp in variational inference

Using re-para trick to get Differentiable unbiased estimator of variational lower Bound

SGVB is appliable to all models w/ continuous latent varcabls

→ regarless of whether data is i.i.d
→ & regarless of how latent vars are structed accross datapoints

AEVB

- Specalized SGVB for i.i.d dataset & continuous latent varcabls
- Key innov of AEVB = intro of recognition model
- Recong model often implemented by NN
- learns to map datapoints (x) to the paras of their approx posterior latent DRVs $q(z|x)$
- uses ancestral sample per datapoint to acheive — avoids using Monte Carlo
- if recognition NN = VAE architecture

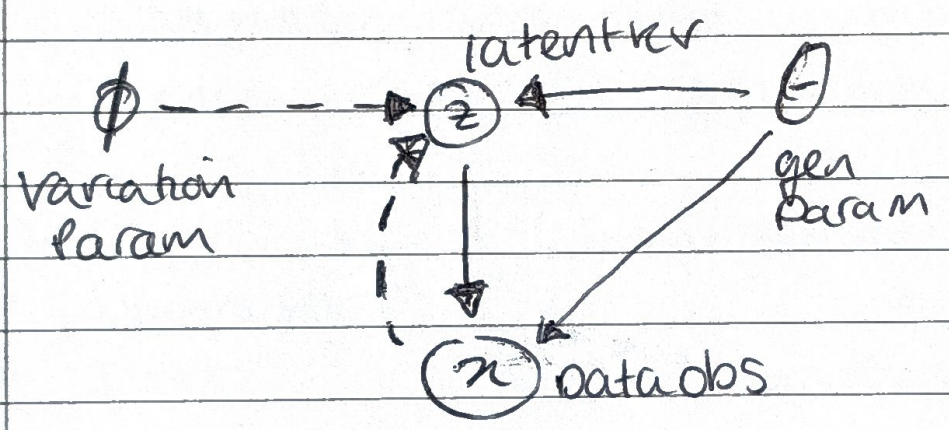
Back to paper ...

Method

Strategy presented can be used to derive a lower bound estimator (Stochastic objective fun) for directed graphical model w/ cont latent vars

paper restricts section to common case w/ i.i.d dataset w/ latent vars

↳ Perform ML or Max like (ML) or Max a posterior (MAP) inference on the global params
↳ & Variational inference of latent vars



Solid line = gen model
$p_\theta(z)p_\theta(x|z)$

Dash = var Approx
$q_\theta(z|x)$ to intract post $p_\theta(z|x)$

var & gen params learnt jointly

Problem Scenario

Dataset $X = \{x^{(i)}\}_{i=1}^{N}$ i.i.d sample

data is generated by random process involving an unobserved cont random variable z

Process of two steps:

(1) a value $z^{(i)}$ is generated from some prior distribution $p_{\theta^*}(z)$ (needs Params) θ defines shapes of distribs

(2) a value $x^{(i)}$ is generated from some conditional distribution $p_{\theta^*}(x|z)$

z is an (latent) hidden, unobserved variable that fundamentally influences how x is generated

θ true param, $z^{(i)}$ latent variable

- (weights & biases)
  - fixed
  - underlying Distribs

- unobserved RVs
- Diff values for diff data points
- often a specific latent var (or set) per data point
- Inferee to estimate (les) latents

Assume that prior $p_\theta(z)$ & likelihood $p_\theta(x|z)$ come from parametric distributions & that their PDFs are DIFF w.r.t both θ & z

both params θ & latents z are unknown to use

simplifs
important Assumps are not made about the marginal or posterior probs

Only interested in general algo that works Efficient in a case of
① intractability
② large dataset - even batch is to costly
prefer small minibatch or single sample

Authors goal is to propose solutions to 3 probs:

① Efficient Approx ML or MAP est for the params θ. these allow us to mimic the hidden random process & generat artificial data

② Eff approx of posterior inference of latent variable z given obs x & for choice of params θ

③ Eff approx marginal inference of the var x

to Acheive all 3 the Authors introduce the concept of a recognition model

$q_\phi(z|x)$

this is an appox to the intractable true posterior $p_\theta(z|x)$

later A method is introduced to learn ϕ & θ jointly

ϕ = recognition model params
θ = generative model params

From a theory point, unobserved vars z are interpreted as latent repres or "code"

the recognition model $q_\phi(z|x)$ is a probabalistic <u>encoder</u>

given a datapoint x it produces a distribution over the possible values of the latent z

↳ from which x could have been generated

Also, $p_\theta(x|z)$ is a probabalistic Decoder as it takes z & Produces x

2.2 Variational Bound

the marginal likelihood (Denominator)

↳ is composed of a sum over the marginal likelihoods of individual datapoints

$$\log p_\theta(x^{(1)}, \ldots, x^{(n)}) = \sum_{i=1}^{n} \log p_\theta(x^{(i)})$$

rewrite = $\underbrace{D_{KL}(q_\phi(z|x^{(i)}) \| p_\theta(z|x^{(i)}))}_{\text{Divergance from true Post}} + \underbrace{L(\theta, \phi, x^{(i)})}_{\text{variational lower Bound on marg lik of } x^{(i)}}$

we want to diff & optimize the lower Bound w.r.t variational params ϕ & generative Params θ

gradient of lower wrt ϕ is problematic

w/ monte carlo gradient exhibits very high variance & is impractal for use

2.3 SGVB estimator & AEVB algo

Section introduces a practical estimator of the lower bound & its derivatives w.r.t the parameters

Assume posterior in the form $q_\phi(z|x)$

↳ technique can be applied to $q_\phi(z)$

* Bunch of stuff rearranging algos & using monte carlo

2.4 Reparamerizing trick

Let z be a continuous random var

& $z \sim q_\phi(z|x)$ be a conditional distrib

often able to express the random variable z as a deterministic variable $z = g_\phi(\epsilon, x)$

ϵ = Auxilary variable w/ indep marginal $p(\epsilon)$

$g_\phi(.)$ is some func parameterized by ϕ

repara is useful here as it can be used to rewrite an expectation w.r.t $q_\phi(z|x)$

such that the MC estimate is Diff w.r.t ϕ

3. Variational Auto-encoder

Example uses as NN for the prob encoder $q_\phi(z|x)$

and where ϕ & θ are optimized jointly w/ the AEVB Algo

→ complicated

4. Related work

generative models are named as Decods

Recognition models are labelled as encodes

hidden units number based on intuitions

3 latent variables

↳ says for high Dim space results became unreliable