Generative Models

Learn a generative model for a class of images (eg. MNIST).

$p(s)$, prior $= \mathcal{N}(0,1)$ usually

$\dim(s) \ll \dim(x)$

$p_\theta(x|s)$

Goal 1:

find θ so that $p_\theta(x)$ is the distribution of our class of images. (eg, MNIST, or room interiors, or ...).

ie, find θ to maximize the probability of the training data, $x^1 \dots x^N$.

ie, maximize $\prod_n p_\theta(x^n) = p_\theta(x^1 \dots x^n)$

or equivalently $\sum_n \log p_\theta(x^n) = \log p_\theta(x^1 \dots x^n)$

assume we can compute $\frac{\partial P_\theta(x)}{\partial \theta}$

eg, G_θ is a NN with weights θ
or a "differentiable renderer"
with parameters θ.

Goal 2: compute $p(s|x)$, so we can do inverse graphics.
ie, infer the scene from an image.

Lets work on Goal 1 first

$P_\theta(x)$ is unknown, but

$$P_\theta(x) = \sum_s P_\theta(x, s) = \sum_s P_\theta(x|s) \cdot P(s)$$

$P(s)$ is known, & $P_\theta(x|s)$ can be learned (as we shall see). $\sum_s$ is very large & intractable, but we will see how to deal with it later.

$$\therefore \log P_\theta(x^1, \cdots, x^N) = \sum_n \log P_\theta(x^n)$$

$$= \sum_n \log \left[\sum_s P_\theta(x^n, s) \right]$$

Difficult to maximize because $\sum_s$ is inside log.

Variational Inference

Let $q_n(s)$ be any distribution (but preferably one that approximates $P_\theta(s|x^n)$, which we want to compute.)

Then,

$$\log P_\theta(x^1 \dots x^N)$$

$$= \sum_n \log \left[\sum_s \frac{q_n(s)\, P_\theta(x^n, s)}{q_n(s)} \right]$$

$$\geq \sum_n \sum_s q_n(s) \log \frac{P_\theta(x^n, s)}{q_n(s)} \stackrel{\text{def}}{=} \mathcal{L}(\theta, q)$$

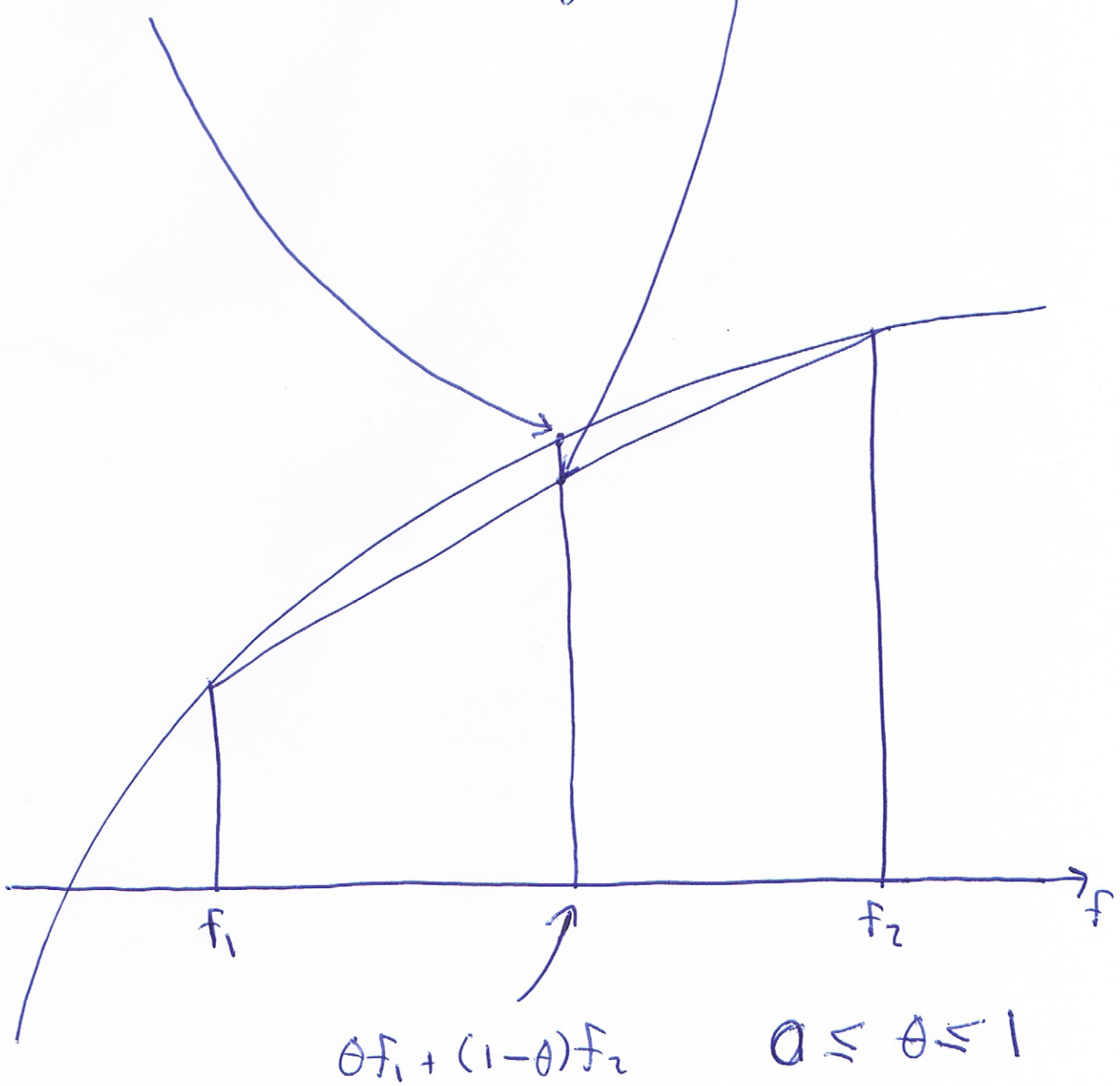
$$q = (q_1 \dots q_N)$$

sum is now out side the log!

Why can we do this?

Because log is concave.

$$\log[\theta f_1 + (1-\theta) f_2] \geq \theta \log f_1 + (1-\theta) \log f_2$$



ie $\log(\theta_1 f_1 + \theta_2 f_2) \geq \theta_1 \log f_1 + \theta_2 \log f_2$

if $\theta_1, \theta_2 \geq 0$ & $\theta_1 + \theta_2 = 1$

more generally,

$$\log \sum_s \theta_s f_s \geq \sum_s \theta_s \log f_s$$

if $\theta_s \geq 0$ & $\sum_s \theta_s = 1$

~~if θ_1, θ_2~~

note: Σ comes outside of log.

let $\theta_s = q_n(s)$ & $f_s = P(X^n, s) / q(s)$

then

$$\log \sum_s q_n(s) \frac{P(X^n, s)}{q_n(s)} \geq \sum_s q_n(s) \log \frac{P(X^n, s)}{q_n(s)}$$

note: $q_n(s) \geq 0$ & $\sum_s q_n(s) = 1$ since it is a prob. distribution.

so $\log P_\theta(x^1 \dots x^n) \geq$

$$\sum_{n,s} q_n(s) \log \frac{P_\theta(x^n, s)}{q_n(s)}$$

when does this become equality?

when $q_n(s) = P_\theta(s|x^n)$ = posterior prob. of s, which is what we also want to ~~compute~~ approximate.

Proof.

$$\sum_{n,s} q_n(s) \log \frac{P_\theta(x^n, s)}{P_\theta(s|x^n)}$$

$$= \sum_{n,s} P_\theta(s|x^n) \cdot \log \frac{P_\theta(x^n, s)}{P_\theta(x^n, s)/P(x^n)}$$

$$= \sum_{n,s} P_\theta(s|x^n) \cdot \log P_\theta(x^n)$$

$$= \sum_{n} \log P_\theta(x^n) \cdot \underbrace{\sum_s P_\theta(s|x^n)}_{1}$$

$$= \sum_n \log P_\theta(x^n)$$

$$= \log P_\theta(x^1 \dots x^n)$$

Instead of maximizing $\log P_\theta(x^1 \dots x^N)$, which is intractable, we shall maximize

$$\mathcal{L} = \sum_{n,s} q_n(s) \log \frac{P_\theta(x^n, s)}{q_n(s)} \quad \text{over } \theta.$$

This is much easier, if we choose $q_n(s)$ wisely.

~~Let's simplify.~~

Note that making $\mathcal{L}$ big forces $\log P_\theta(x^1 \dots x^N)$ to be big too.

~~However, we must choose an appropriate $q_n(s)$.~~

Simplifying

$$\mathcal{L} = \sum_{n,s} q_n(s) \log \frac{p_\theta(x^n, s)}{q_n(s)}$$

$$= \sum_{n,s} q_n(s) \log \frac{p_\theta(x^n|s) \cdot p(s)}{q_n(s)}$$

$$= \sum_{n,s} q_n(s) \log p_\theta(x^n|s) - \sum_n \underbrace{\sum_s q_n(s) \log \frac{q_n(s)}{p(s)}}_{KL(q_n, p)}$$

note: $KL(q_n, p)$ does not depend on θ.

KL divergence

Given two distributions q, p,

$$KL(q,p) = \sum_s q(s) \log \frac{q(s)}{p(s)}.$$

Exercise: $KL(q,p) \geq 0$

$KL(q,p) = 0$ iff $q = p$

interpretation:

KL measures the "distance" between two distributions.

but note: $KL(q,p) \neq KL(p,q)$

∴ When maximizing $\mathcal{L}$, we put downward pressure on $KL(q_n, p)$.

ie, $KL(q_n, p)$ acts like a regularization term, pressuring $q_n(s) \simeq p(s)$, the prior on s.

As we shall see later, there is also pressure to make $q_n(s) \simeq p(s|x^n)$, the posterior of s given x^n.

$\mathcal{L}$ is a function of θ & $q = (q_1 \cdots q_n)$.

ie, $\mathcal{L} = \mathcal{L}(\theta, q)$.

To maximize $\mathcal{L}$, we will maximize over both θ & q. In fact, we will alternate ~~~~ between θ & q.

ie:

~~do until convergence~~

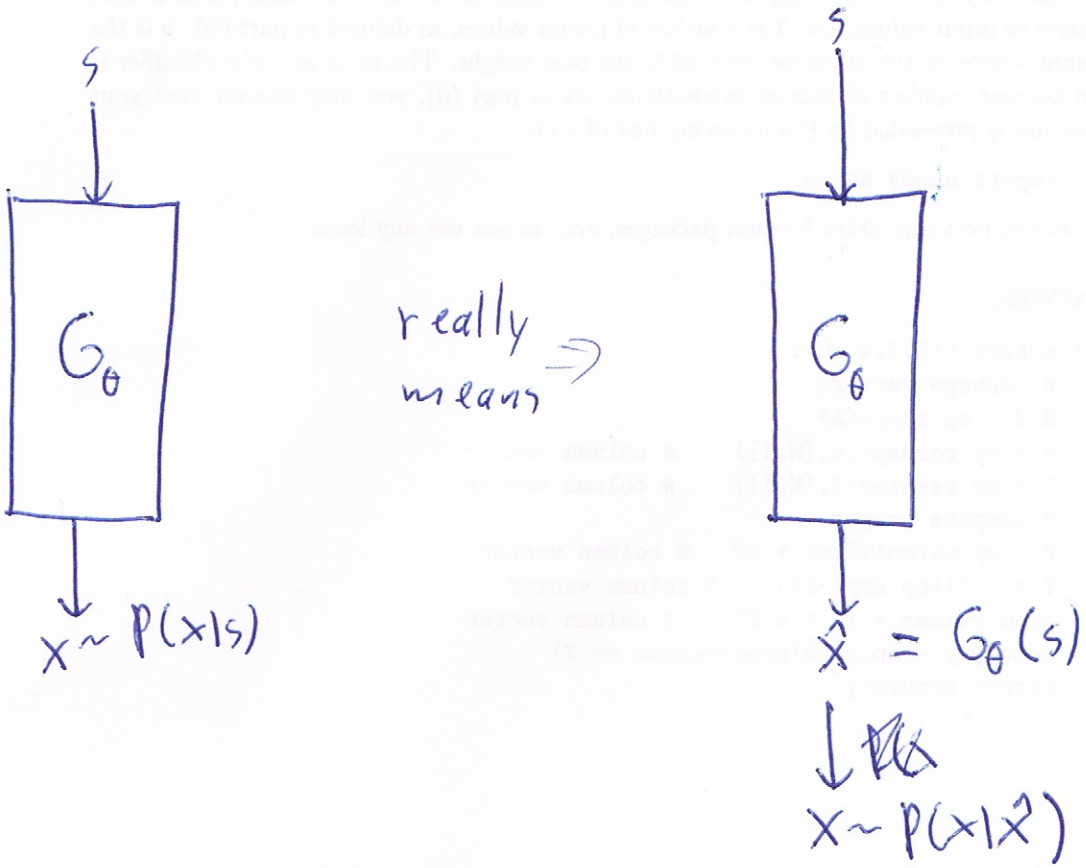
~~maximize $\mathcal{L}$ over θ~~

~~maximize $\mathcal{L}$ over q~~

(1) we can ~~do~~ maximize over θ by doing back-propagation through

$$\sum_{n,s} q_n(s) \log P_\theta(x^n | s)$$

Our Generative Model



Usually, so, $P(x|s) = P(x|\hat{x} = G_\theta(s))$.

P is chosen so that $P(x|\hat{x})$ is high iff $x \approx \hat{x}$

Usually, $P(x|\hat{x}) = \mathcal{N}(x|\hat{x}, \sigma) = \dfrac{e^{-\|x-\hat{x}\|^2/2\sigma^2}}{(\sqrt{2\pi}\sigma)^k}$

$k = \dim(x)$.

$$\therefore \log P_\theta(x|s)$$

$$= -\|x - \hat{x}\|^2/2\sigma^2 - k \log \sqrt{2\pi}\sigma$$

$$= -\|x - G_\theta(s)\|^2/2\sigma^2 - k \log \sqrt{2\pi}\sigma$$

$$\therefore \frac{\partial \log P(x|s)}{\partial \theta} = \frac{1}{\sigma^2} \|x - G_\theta(s)\| \frac{\partial G_\theta(s)}{\partial \theta}$$

so, maximizing $\log P_\theta(x|s)$ over θ

means minimizing $\|x - G_\theta(s)\|^2$ over θ.

we can easily do this by backpropagation if G_θ is a NN.

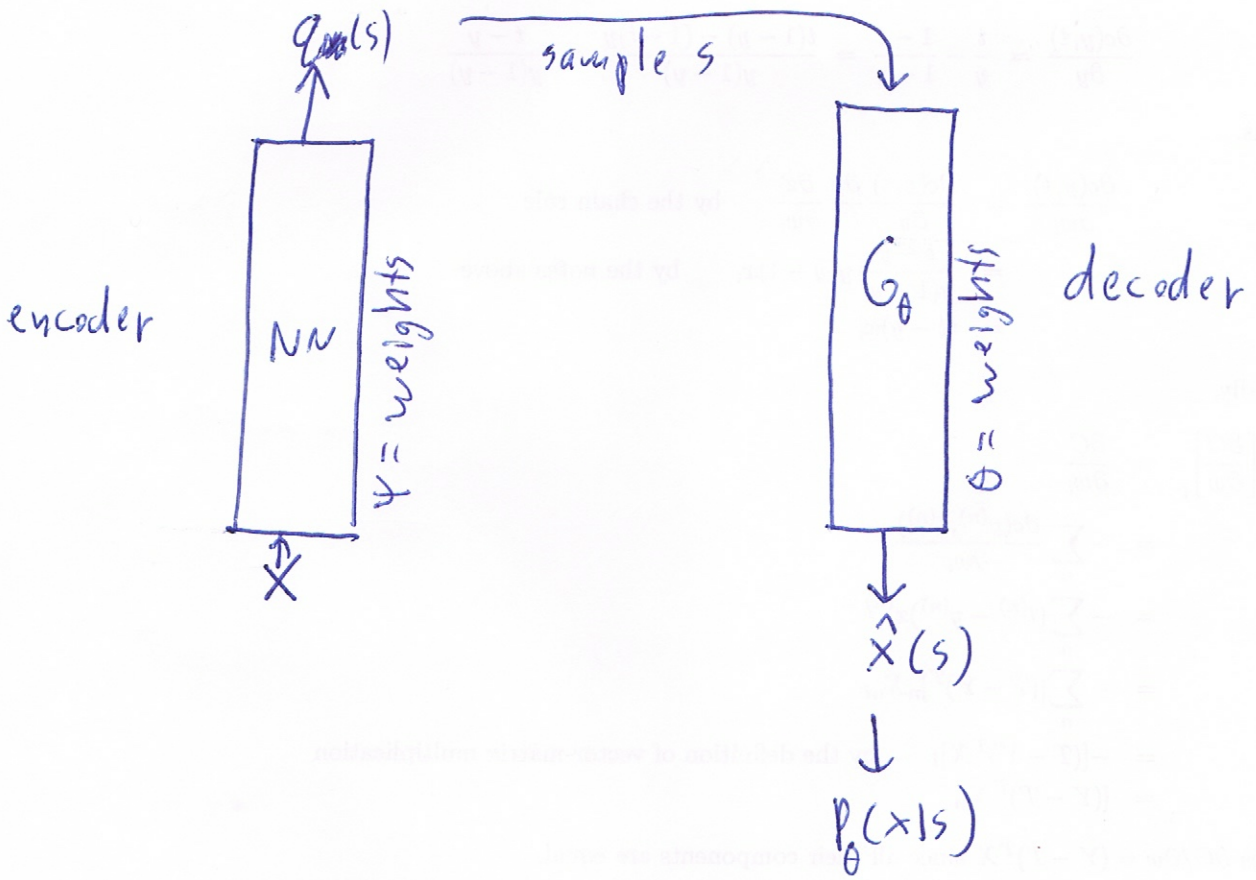
So, we can maximize $\mathcal{L}(\theta, q)$ over θ by backprop.

How can we maximize over q?

The maximum occurs ~~at~~ when $q_n(s) = P(s|x^n)$, as we shall see later. We would love to ~~can~~ compute this, but it is intractable in general. Instead, we ~~assume $q_n(s)$ has a particular~~ limit ourselves to a tractable ~~set~~ class of q_n functions, & maximize over that set. This is called variational approximation

Variational Autoencoders

If we assume that $q_n(s)$ is defined by a neural net with ~~parameters~~ input x^n, then we have a variational autoencoder.



$$q_n(s) = q_\psi(s|x^n)$$

so, $\mathcal{L}(\theta, q) = \mathcal{L}(\theta, \psi)$ i.e., need to maximize $\mathcal{L}$ over ψ.

We can also use back-prop to maximize ~~over~~ $\mathcal{L}$ over ψ. (how?)

But how do we back-propagate through the sampling operation?

(See Kingma & Welling).

In the end, when $\mathcal{L}$ is maximized, $q_n(s)$ will be an approximation of $p_\theta(s|x^n)$, which we wanted.

ie, the prob. dist. of the scene given an image.

To see this, recall that

$$\mathcal{L} = \sum_{n,s} q_n(s) \log \frac{p(x^n, s)}{q_n(s)} \quad \text{by def.}$$

$$= \sum_{n,s} q_n(s) \log \frac{p(s|x^n) \cdot p(x^n)}{q_n(s)}$$

$$= -\sum_{n,s} q_n(s) \log \frac{q_n(s)}{p(s|x^n)} + \sum_{n,s} q_n(s) \log p(x^n)$$

$$= -KL[q_n(s), p(s|x^n)] + \sum_n \log p(x^n) \cdot \underbrace{\sum_s q_n(s)}_{1}$$

$$\therefore \mathcal{L} = -KL[q_n(s), p(s|x^n)] + \sum_n \log p(x^n)$$

so, maximizing $\mathcal{L}$ over q_n is equivalent to minimizing the KL divergence between $q_n(s)$ & $p(s|x^n)$.

ie, making $q_n(s)$ similar to $p(s|x^n)$

Note: by maximizing $\mathcal{L}$ we are both

- making $q_n(s)$ similar to $p(s|x^n)$, &
- maximizing the likelihood of the data, $p(x^1, \cdots, x^N)$.