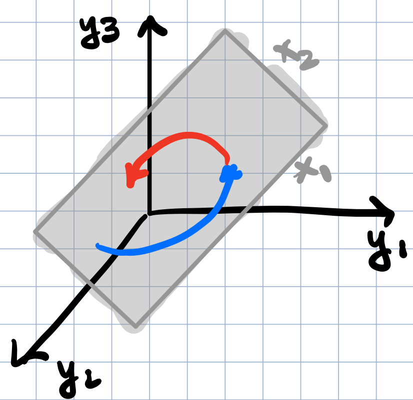
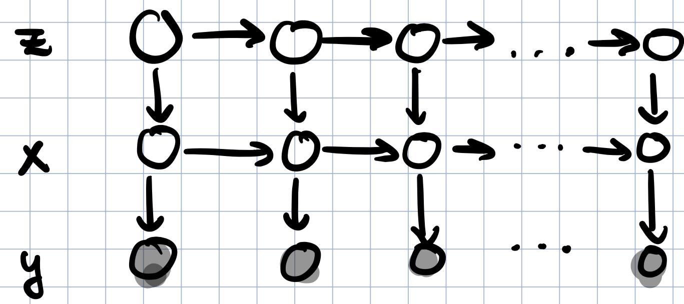
Gradient Based Variational Inference

5/19/25

Recall: SLDS





$$z_t \sim \mathrm{Cat}(\pi_{z_{t-1}})$$
$$x_t \sim N(A_{z_t} x_{t-1} + b_{z_t},\ Q_{z_t})$$
$$y_t \sim N(C x_t + d,\ R)$$

EM for SLDS

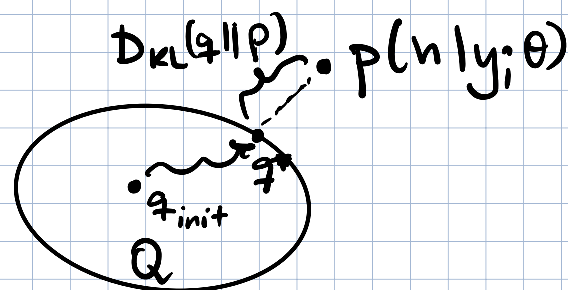
E Step: $q(z,x) \leftarrow p(z,x \mid y;\theta)$ **HARD!**

M step: $\theta \leftarrow \underset{\theta}{\arg\max}\ \mathbb{E}_q[\log p(z,x,y;\theta)]$

IDEA: Approximate q w/ VI

Let $h = (z,x)$

$$D_{KL}(q \| p) = \mathbb{E}_q[\log q(h) - \log p(h \mid y;\theta)]$$



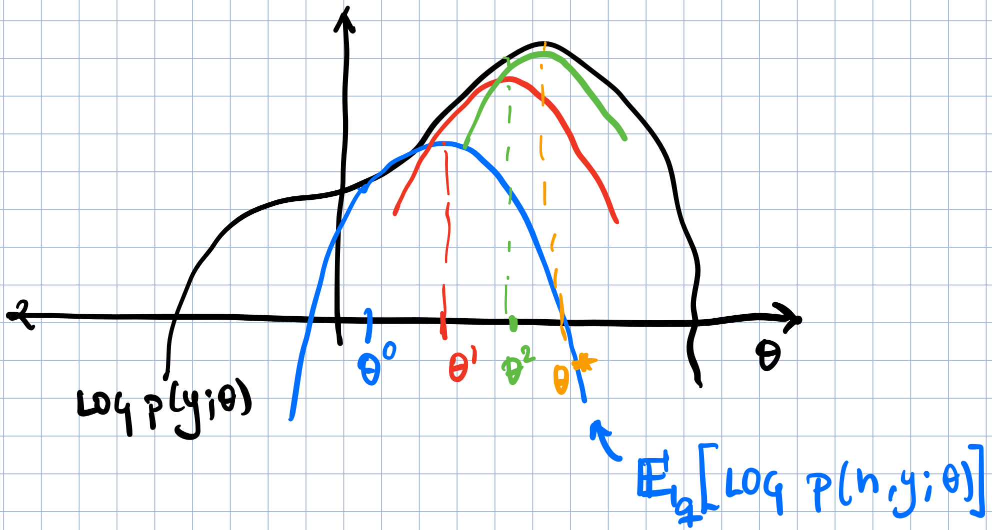
* for certain Q, we can minimize KL via coordinate descent

Variational EM:

E: $q(h)\ \underset{Q}{\arg\min}\ D_{KL}(q(h) \| p(h|y;\theta))$

M: same as above

What is (Variational) EM doing?



Consider minimizing $D_{KL}(q \| p)$

Bayes' Rule

$$p(h \mid y;\theta) = \frac{p(h, y;\theta)}{p(y;\theta)}$$

$$D_{KL}(q(h) \| p(h \mid y;\theta))$$

$$= -\mathbb{E}_q[\log p(h \mid y;\theta) - \log q(h)]$$

$$= -\mathbb{E}_q[\log p(h, y;\theta) - \log p(y;\theta) - \log q(h)]$$

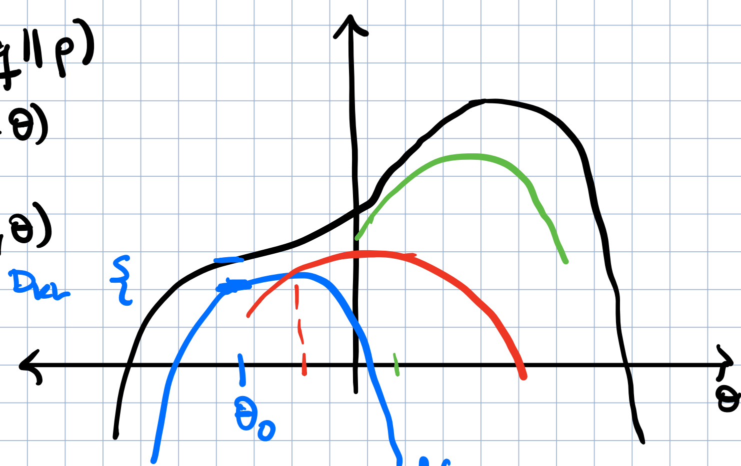
$$= \log p(y;\theta) - \underbrace{\mathbb{E}_q[\log p(h, y;\theta) - \log q(h)]}_{\text{EVIDENCE LOWER BOUND (ELBO)} \quad \mathcal{L}(q,\theta)}$$

$$\mathcal{L}(q,\theta) = \log p(y,\theta) - D_{KL}(q(h) \| p(h \mid y;\theta))$$

Var. E step: $q \leftarrow \underset{Q}{\operatorname{argmin}}\, D_{KL}(q \| p)$

$= \underset{Q}{\operatorname{argmax}}\, \mathcal{L}(q,\theta)$

M step: $\theta \leftarrow \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(q,\theta)$



$\mathcal{L}(q,\theta)$

Fixed form VI

- Last time, we simply assumed $q(z,x)$ factorized

$$Q_{MF} = \{q : q(z,x) = q(z)q(x)\}$$

This is called the <u>mean field</u> assumption

- A stricter assumption is to fix the functional form

eg: $Q = \{q : q(h;\phi) = \text{Cat}(z; \phi_1)\, N(x \mid \phi_2, \phi_3)\}$

$\phi = (\phi_1, \phi_2, \phi_3)$ are the <u>variational params</u>.

Then optimizing over $Q \equiv$ optimizing over params ϕ

$$\mathcal{L}(\phi,\theta) = \mathbb{E}_{q(z,x;\phi)}[\log p(h, y; \theta) - \log q(h;\phi)]$$

<u>IDEA</u>: maximize $\mathcal{L}(\phi,\theta)$ via stochastic gradient ascent

$$\nabla_\theta \mathcal{L}(\phi,\theta) = \mathbb{E}_{q(h;\phi)}[\nabla_\theta \log p(h, y; \theta)]$$

$$\approx \frac{1}{M}\sum_m \nabla_\theta \log p(h^{(m)}, y; \theta) \quad \text{where } h^{(m)} \overset{iid}{\sim} q(h;\phi)$$

$$\nabla_\phi \mathcal{L}(\phi,\theta) \neq \mathbb{E}_{q(h;\phi)}[\nabla_\phi(\log p(h,y;\theta) - \log q(h;\phi))]$$

Reparameterization Trick

noise indep of ϕ

Suppose $\varepsilon \sim N(0,1)$
$h = r(\varepsilon, \phi)$ $\Longrightarrow h \sim q(h;\phi)$

reparam. function

eg $q(h;\phi) = N(h \mid \phi, 1)$ then $h = \phi + \varepsilon = r(\varepsilon, \phi)$

Then

$$\nabla_\phi \mathcal{L}(\theta, \phi) = \mathbb{E}_{r(\varepsilon)}\Big[\nabla_\phi\Big(\log p(r(\varepsilon,\phi), y; \theta) - \log q(r(\varepsilon,\phi);\phi)\Big)\Big]$$

"Law of the unconcious statistician"