Variational Inference ~~#~~ Loopy Belief prop 11/4/13

+ Expectation propagation - HW

- HW + project progress - ambitious
- great GS. ~~HW~~ - not here wednesday

Last week: mean Field variational inference.

Today: LBP, in context of Ising model: Markov Random Field. Review MF in this context

Idea behind variational inference; review for a few minutes.

let $x = \{x_1, \dots, x_n\}$ let $z_{1:m}$ be hidden vars, θ params. $z = \{z_{1:m}, \theta\}$.

as always, we want the posterior distribution of latent vars:

$$p(z \mid x) = \frac{p(z, x)}{\int_z p(z, x)\, dz}.$$

- posterior links data to model. used for downstream analyses: posterior predictive dist, interpretation, etc. via p.t. estimates of z.

- variational inference is one method to compute the approximate posterior dist.

- cant compute the posterior for many models. or - very hard. e.g. truncated gaussian, mean/var complex, unnormalized - easy.

Look again at G Mixture Model.

$$x_i \sim N(\mu_{z_i}, \Sigma) \qquad z_i \sim \text{Mult}(\pi) \qquad \mu_k \sim N(0, \tau^2)$$

chain rule

posterior

$$p(\mu, z \mid x) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i \mid \pi)\, p(x_i \mid z_i, \mu)}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i)\, p(x_i \mid z_i, \mu)\, d\mu_{1:K}}$$

easy for given z

independence.

exponentially hard. K^n terms.

difficult integral.

we are reviewing this because we are about to launch into LDA models — hidden variables also create exponentially difficult denominator in posterior too — + hard integral.

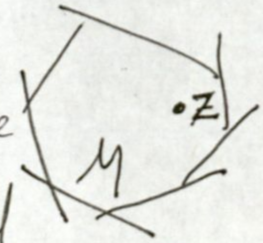
Variational Inference

→ return to the space of parameters M

For many models (discrete, gaussian)

$\exists$ theorem that states, for finite data, the set of parameters (Z) consistent with these data is a convex polytope:

optimization: - Posterior density on this space
- approximate M, optimize within M



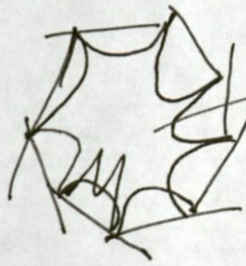
cartoon version

Mean Field: find fully factorized

approximation to posterior: $q(z) = \prod_{i=1}^{m} q(z_i)$.

Compute: $\min_{z} KL(q(z) \| p(z|x))$

each variable is independent



→ M_{MF}

→ based on reverse KL, we know M_{MF} is a subset of the space M.

→ very possible z^* not contained within M_{MF}.

In examples we've discussed ~~...~~ + HW we go directly from KL divergence w/ unnormalized posterior. ~~...~~

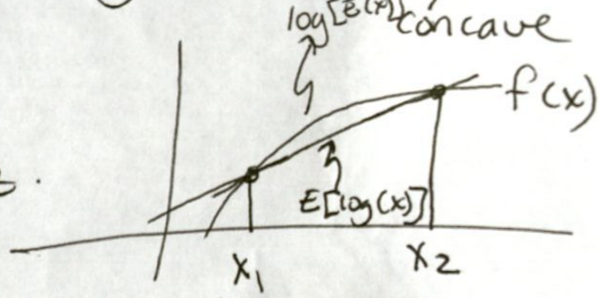
~~...~~

Evidence Lower Bound (ELBO).

Jensen's Inequality

$$\log E(x) \geq E[\log(x)].$$

concave



$$t f(x_1) + (1-t) f(x_2) \leq f(t x_1 + (1-t) x_2)$$

log prob of obs:

$$\log p(x) = \log \int_z p(x,z)\,dz$$

$$= \log \int_z p(x,z) \frac{q(z)}{q(z)}\,dz$$

$$= \log E_q\left[\frac{p(x,z)}{q(z)}\right].$$

$$\geq E_q[\log p(x,z)] - E_q[\log q(z)]$$

T-BE4-V2

3

Recall from last week, variational objective

$$KL(q \| \tilde{p}) = \underbrace{-E_q[\log p(z,x)] + E_q[\log q(z)]}_{\text{negative elbo}} + \log p(x)$$

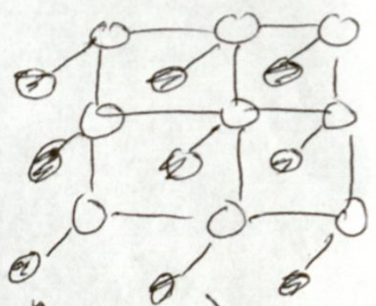
Thus: ELBO is KL divergence wrt unnormalized posterior. Since $p(x)$ indep of q, this is equiv. to minimizing ELBO.

Now: • choose q. s.t. these expectations are computable
• maximize $q(z)$ w.r.t. elbo to get as tight an approximation to $p(z,x)$ as possible.

Ising Model (MRF) Denoising image: $y = \{-1, 1\}$ pixels
$X = \{-1, 1\}$ denoised pixels

$\psi_s(x) = p(y_i|x_i) \triangleq L_i(x_i)$

$\psi_{st}(x_s x_t) = W_{st} x_s x_t$ $W_{st} = 1$ - neighbors want to have similar states



$p(y|x) = \prod_{i=1}^{n} \exp(-L_i(x_i))$ likelihood.

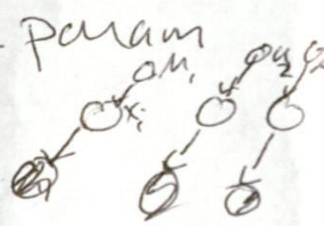
$p(x) = \frac{1}{Z_0} \exp\left(-\sum_{i=1}^{n} \sum_{j \in nei(i)} x_i x_j\right)$ prior

Posterior: $p(x|y) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{n} \sum_{j \in nei(i)} x_i x_j - \sum_{i=1}^{n} L_i(x_i)\right)$

Mean Field $q(x) = \prod_{i=1}^{n} q(x_i)$ let μ_i be "mean value" or variational param

$\log(q_i(x_i)) = E_{\neg q}(\log \tilde{p}(x))$

↳ basic coordinate ascent from last class.



ELBO: for x_i $\log(q_i(x_i)) = E_{\neg q}\left[x_i \sum_{j \in nei(i)} x_j + L_i(x_i) + c\right]$. 21.43 book

$q_i(x_i) \propto \exp\left\{x_i \sum_{j \in nei(i)} \mu_j + L_i(x_i)\right\}$.

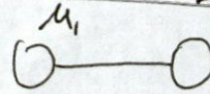
so $q_i(x_i = +1) = \dfrac{\exp\left\{\sum_{j \in nei(i)} x_j + L_i(+1)\right\}}{\sum_{x_i \in \{-1,+1\}} \exp\left\{\sum_{j \in nei(i)} x_j + L_i(x_i)\right\}}$. → logistic function to get μ. Write out for -1, take average

(4) Then 11/4/13

$$\mu_i = +1\left(q_i(x_i=+1)\right) + -1\left(q_i(x_i=-1)\right)$$

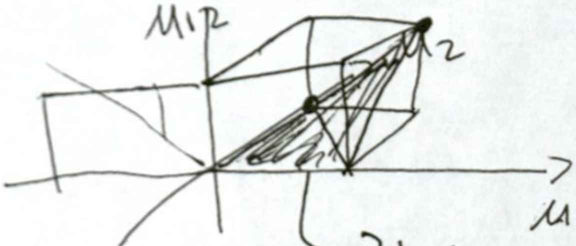
→ MF = GIBBS: $z_i \propto \exp\left[E_{\neg q}\left[\log P(z_i \mid z_{\neg i}, x)\right]\right]$ — looks like conditional

return to drawing



$\{0,1\}$ $\mu_{12} = E[x_1 x_2]$

s.t. $0 \le \mu_{12} \le \mu_i$



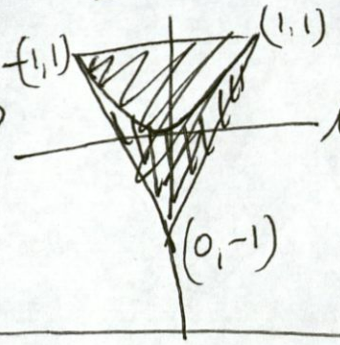
→ marginal polytope $P(x)$

flip over to $-1,-1$ on each axis. μ_1, μ_2

Fully factorized →

$\mu_1^2 = \mu_{12}$.

Slice at $\mu_1 = \mu_2$.



Loopy belief propagation: LBP.

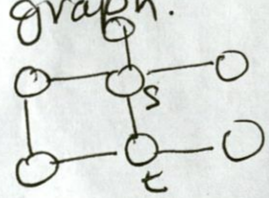
~~recall~~. now: q only locally consistent

not a valid joint probability

LC: $q(x_i) = \sum_{x_j} q(x_i x_j)$ $\forall$ pairs $(i,j) \in E$

Recall Belief ~~approx~~ propagation in a tree is exact. Not true in a graphical model with loops. E.g. Forward backward, peeling (in tree) inside-outside pass. → final answer after

Idea: apply Belief Propagation to loopy graph.



Same ising model. $\psi_s(x_s) = L_s(x_s)$

$\psi_{st}(x_s x_t) = x_s x_t$

Algorithm
- messages $m_{s\to t}(x_t) = 1$ $\forall$ edges.
- beliefs $\mu_s = 1$ $\forall$ nodes s.

repeat: $m_{s\to t}(x_t) = \sum_{x_s} L_s(x_s)\, x_s x_t \prod_{u \in \text{neigh}(s)\setminus t} m_{u\to s}(x_s)$

5

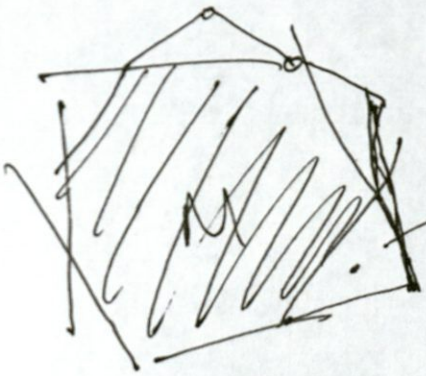
Cont.

11/4/13.

$\mu_s \propto L_s(x_s) \prod_{t \in neighbor(s)} m_{t \to s}(x_s)$.

return μ_s. $\rightarrow$ Synchronous updates.

Will LBP converge? maybe not - no guarantees.
but methods to help - dampening to avoid oseylation
- asynchronous steps.
- scheduling (tree based reparametrization)



M_{LBP}.

$M = M_{LBP}$ when graph is tree

o.w. $M \subset M_{LBP}$ outerpolytope

$O(|V|+|E|)$ constraints (local consistency)

why:

EPropagation:
in HW.
Twist on MF.

MF vs LBP.
- MF not exact in Trees
- MF - node marginals / LBP node + edge marginals.
- MF - more local optima.