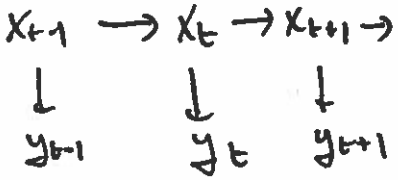
Summary

- EP INFERENCE FOR GENERAL BAYESIAN NETWORKS ← High dimensional integrals of approx posterior replaced by low dimension $P(X_i \mid Pa(X_i))$ integral
- APPROXIMATE MOMENT COMPUTATIONS USING IMPORTANCE SAMPLING ↔ REPLACE
- LEARN TO APPROXIMATE MOMENT COMPUTATIONS USING A LINEAR MODEL ← generalized, ← speed up inference
- USE UNCERTAINTY IN MOMENTS TO DECIDE WHETHER TO REVERT TO IMPORTANCE SAMPLING

EP

- consider chain BNs to avoid bookkeeping



$$p(X_{1:T} \mid y_{1:T}) = \prod_t p(x_t \mid x_{t-1}) p(y_t \mid x_t) = \prod_t f_t(x_t, x_{t-1})$$

approximate affect of f_t on true posterior

$$q(X_{1:T}) = \prod_t \alpha_t(x_t)\beta_t(x_{t-1})$$

(NB: NOT REALLY MEANFIELD
eg. for LGSSMs recovers exact (correlated) Kalman Filter/Smoother equations)

ITERATIVE UPDATE RECIPE

1. FORM CAVITY $\dfrac{q(X_{1:T})}{\alpha_t(x_t)\beta_t(x_{t-1})}$

2. INCLUDE NEW POTENTIAL $\dfrac{q(X_{1:T})}{\alpha_t(x_t)\beta_t(x_{t-1})} \dfrac{f_t(x_t, x_{t-1})}{p(x_t \mid x_{t-1}) p(y_t \mid x_t)} =$

3. MOMENT MATCH
(match mean & variance of $q(X_{1:T})$ for transition & (un-normalized) α_t & β_t)

$$\frac{q(X_{1:T})}{\alpha_t(x_t)\beta_t(x_{t-1})} p(x_t \mid x_{t-1}) p(y_t \mid x_t) \overset{MOM}{=} \frac{q(X_{1:T})}{\alpha_t(x_t)\beta_t(x_{t-1})} \alpha_t^{new}(x_t)\beta_t^{new}(x_{t-1})$$

As $q(x_{1:T}) = \prod_t q(x_t)$ $\Rightarrow$ mean & variance of $\underline{x}_{\neq t,t-1}$ already matched

$\Rightarrow$ moment matching reduces to $\alpha_t(x_t)\beta_t(x_{t-1})\alpha_{t-1}(x_{t-1})\beta_{t-1}(x_t)$

$$\frac{q(x_t,x_{t-1})}{\alpha_{t-1}(x_{t-1})\beta_t(x_{t-1})}\, p(x_t|x_{t-1})p(y_t|x_t) \overset{\text{mom}}{=} \frac{q(x_t,x_{t-1})}{\alpha_{t-1}(x_{t-1})\beta_t(x_{t-1})}\,\alpha_t^{\text{new}}(x_t)\beta_t^{\text{new}}(x_t)$$

$$\Rightarrow \quad \beta_{t+1}(x_t)\alpha_{t-1}(x_{t-1})\,p(x_t|x_{t-1})p(y_t|x_t) \overset{\text{mom}}{=} \beta_{t+1}(x_t)\alpha_{t-1}(x_{t-1}) \times \alpha_t^{\text{new}}(x_t)\beta_t^{\text{new}}(x_t)$$

Now consider update for $\alpha_t(x_t)$ ie $\langle x_t\rangle$ $\langle x_t^2\rangle$ moments

$$\int \beta_{t+1}(x_t)\alpha_{t-1}(x_{t-1})\,p(x_t|x_{t-1})p(y_t|x_t)\,dx_{t-1} \overset{\text{mom}}{=} \int \beta_{t+1}(x_t)\alpha_{t-1}(x_{t-1})\alpha_t^{\text{new}}(x_t)\beta_t^{\text{new}}(x_t)\,dx_{t-1}$$

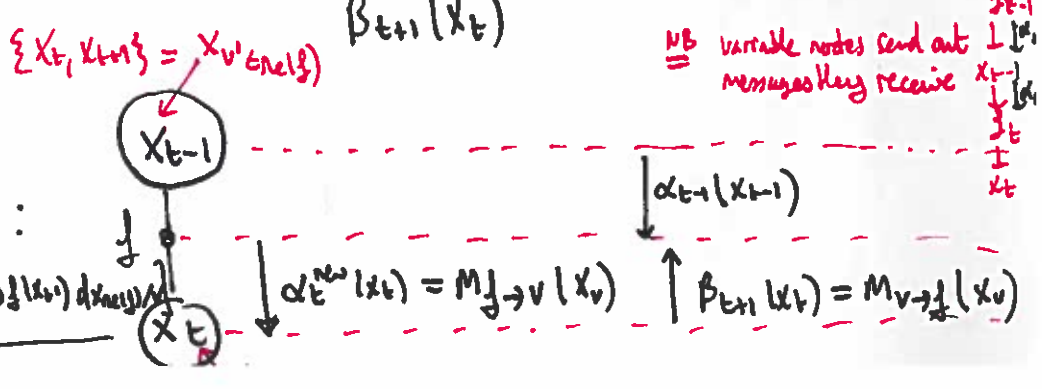
{as $\int \alpha_{t-1}(x_{t-1})\tilde{\beta}_t^{\text{new}}(x_{t-1})\,dx_{t-1} = \text{const.}$ & does not affect 1st & 2nd moments}

$$\int \beta_{t+1}(x_t)\alpha_{t-1}(x_{t-1})\,p(x_t|x_{t-1})p(y_t|x_t)\,dx_{t-1} \overset{\text{mom}}{=} \beta_{t+1}(x_t)\,\alpha_t^{\text{new}}(x_t)$$

project argument onto a Gaussian w/ same 1st & 2nd moments
$\Rightarrow$ ONLY REQUIRES 2D INTEGRALS

$$\Rightarrow \alpha_t^{\text{new}}(x_t) = \frac{\text{proj}\left[\int \beta_{t+1}(x_t)\alpha_{t-1}(x_{t-1})\,p(x_t|x_{t-1})p(y_t|x_t)\,dx_{t-1}\right]}{\beta_{t+1}(x_t)}$$

message passing interpretation:



$$M_{f\to v}(x_v) = \frac{\text{proj}\left[\int f(x_{ne(f)})\prod_{v'\in ne(f)} m_{v'\to f}(x_{v'})\,dx_{ne(f)\setminus v}\right]}{\ldots}$$

IMPORTANCE SAMPLING FOR EP

(Barthelmé & Chopin, 2011)

For approximations $q(x_t)$ that are in the exponential family w/ moments $u(x)$ key challenge is to compute:

$$\emptyset = \int f(x_{ne(f)}) \prod_{v' \in ne(f)} m_{v' \to f}(x_{v'}) \, u(x_v) \, d x_{ne(f)}$$

(for Gaussian $u(x_v) = x_v^2$ & x_v)

IDEA: USE IMPORTANCE SAMPLING

$$q(x_t, x_{t-1}) = p(x_t | x_{t-1}) \Gamma(x_{t-1})$$

$$\emptyset = \int d x_{t:t-1} \; x_t^k \, l(x_t, x_{t-1}) = \int d x_{t:t-1} \; x_t^k \frac{l(x_t, x_{t-1})}{q(x_t, x_{t-1})} q(x_t, x_{t-1})$$

$$= \int d x_{t;t-1} \; x_t^k \, w(x_t, x_{t-1}) \, q(x_t, x_{t-1}) \approx \frac{1}{N} \sum_n \left(x_t^{(n)}\right)^k w(x_t^{(n)}, x_{t-1}^{(n)})$$

where $\{x_{t-1}^{(n)}, x_t^{(n)}\} \sim q(x_t, x_{t-1})$

$$w(x_t, x_{t-1}) = \frac{l(x_t, x_{t-1})}{q(x_t, x_{t-1})} = \frac{\beta_{t+1}(x_t) \, \alpha_{t-1}(x_{t-1}) \, p(y_t | x_t) \, p(x_t | x_{t-1})}{\Gamma(x_{t-1}) \, p(x_t | x_{t-1})}$$

↓ paper does not consider terms like this could incorporate into q.

- Assymptotically unbiased ⇒ consistent
- involves importance sampling each time we need to send a message

LEARNING TO PASS MESSAGES

Hess et al 2013, Eslami 2014, Jitkrittum et al 2015

(neural networks) no uncertainty, initial training step ⇒ not online

(random forests) uncalibrated heuristic uncertainties, not online

(this paper) calibrated uncertainty, online + active

Idea: learn a mapping from $\{\alpha_{t-1}(x_{t-1}), \beta_{t+1}(x_t)\}$ to moments & .:

$\{\alpha_t^{(new)}(x_t), \beta_t^{(new)}(x_t)\}$

Problem

inputs: eg. Gaussian distributions (↑ non-standard) outputs: moments

Idea: design simple (↓ Bayesian) linear in the parameters model w/ clever basis functions by appealing to GPs.

Let $r^{(i)}$ denote incoming messages $r_i = \{\alpha_{t-1}^{(i)}(x_{t-1}), \beta_{t+1}^{(i)}(x_t)\}$ (↙ different examples in training set)

$K_2(r^{(i)}, r^{(j)}) = e^{-\frac{1}{2\gamma^2}\infty}$ Similarity between (sets of) incoming messages

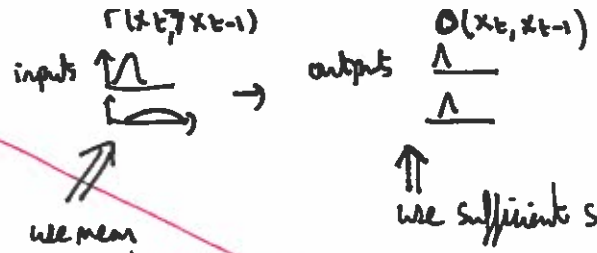
to handle distributions use:

$$K_2(\alpha^{(i)}(x), \alpha^{(j)}(x)) = \exp\left(-\frac{1}{2\gamma^2}\int(\alpha_i^{(i)}(x) - \alpha_j^{(j)}(x))^2 dx\right)$$

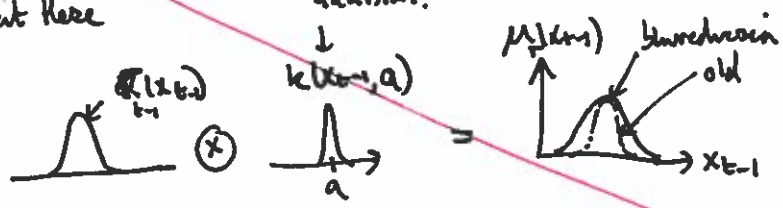
↑ integral over product of Gaussians ⇒ tractable

to be slightly more general (& correct to the authors' previous work) can blur messages first by eg a Gaussian (↓ Gaussian)

$$\mu_r^{(i)}(x) = \int k_1(x, x')\alpha(x')dx'$$



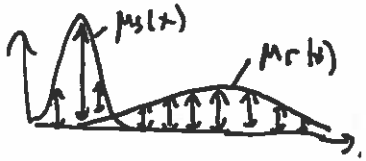
Step 1: blur



$$\mu_{\Gamma}(x_{t-1}) = \int k_1(x_t, x_{t-1}) \, \Gamma(x_t) \, dx$$

Step 2: define kernel between two probability distributions by average square difference between distributions

$$k_2(\Gamma, \Sigma) = e^{-\frac{\|\mu_\Gamma - \mu_\Sigma\|^2_{\mathcal{H}}}{2\sigma^2}} = e^{-\frac{1}{2\sigma^2}\int_{-\infty}^{\infty} (\mu_\Gamma(x) - \mu_\Sigma(x))^2 dx}$$



will be difference between two Gaussians whose means & variances depend on the natural parameters of $[\alpha^{[1]}_{t-1}, \beta^{[1]}_{t+1} \,//\, \alpha^{[2]}_{t-1}, \beta^{[2]}_{t+1}]$

Step 3: do GP regression using the Kernel

BUT Step 3 has $O(N^3)$ cost ⇒ actually don't do GP regression at all, but approximate using sparse GP

Rahimi Recht: $k(x-y) = \int K(\omega) \, e^{j\omega^T(x-y)} \, d\omega$ [Bochner's theorem]

$$= \int p^*(\omega) \, e^{j\omega^T(x-y)} \, d\omega \approx \frac{1}{K}\sum_{k=1}^{K} e^{j\omega_k^T(x-y)} = \underline{\psi}(x)^T \underline{\psi}(y) \quad (\omega_k \sim p^*(\omega))$$

parametric model where $y(x) = \sum_k [\alpha_k \cos(\omega_k^T x) + \beta_k \sin(\omega_k^T x)]$, $\alpha_k, \beta_k \sim \text{Norm}(0,1)$

$\int k_1(x, x') \, k_1(x, x'') \, dx = G(x' - x'')$

NOW USE THIS RESULT TWICE:

Now consider kernel above (main term like $\int \mu_A(x)\mu_B(x) \, dx$):

$$= \int k_1(x, x') A(x') \, k_1(x, x'') B(x'') \, dx' dx'' dx = \mathbb{E}_{A(x)} \mathbb{E}_{B(x)} [G(x' - x'')]$$

approx $G(x'-x'')$ as $\sum_{k=1}^{K} \psi_k(x)\,\psi_k(y)$ & then can analytically integrate over $A(x)$ & $B(x)$

$$\Rightarrow \int \mu_A(x)\,\mu_B(x)\,dx = \sum_{k=1}^{K} \langle \psi_k(x) \rangle_{A(x)} \langle \psi_k(y) \rangle_{A(y)} = \hat{\phi}_A^T \hat{\phi}_B$$

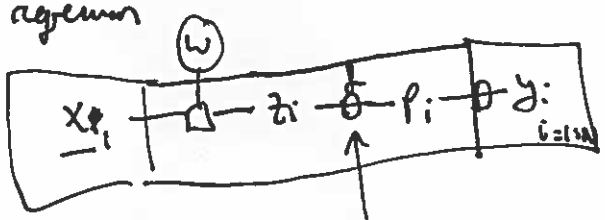
$$\therefore k_2(r,s) \cong e^{-\|\hat{\phi}_r - \hat{\phi}_s\|^2 / 2\sigma^2}$$ (now just a vector norm)

Now apply R&R to this (SE) kernel!

FROM HERE STANDARD ON LINE LEARNING & PREDICTION IN A LINEAR-IN-THE-PARAMETERS REGRESSION MODEL.
[NB. SMART BASIS FUNCTIONS] (whole point of kernel is to design)

RESULTS

Binary logistic regression



look @ this factor (only bit where EP moments non-trivial to compute)

$p(w) = N(0,I)$ ↓ linear so can be handled exactly in EP

$p(z|w,x) = \delta(z_i - w^T x_i)$

$p(p_i|z_i) = \delta\left(p - \frac{1}{1+e^{-z}}\right)$

$m_{p_i \to f} = \text{Beta}(p_i; \alpha, \beta)$

$m_{z_i \to f} = N(z_i; \mu, \sigma^2)$

Bernoulli so can be handled exactly in EP ↓

$p(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}$