We did a var ~~dist~~ ~~tot~~ approx with a factored dist — BUT we could do others:

eg. Q: a tree structured dist?

$$E_Q = -E_Q \log P - H_Q$$

1) We can compute H_Q

recall tree:

$$q(x_1 \cdots x_N) = \left[\prod_{i \in V} q_i\right]\left[\prod_{ij \in E} \frac{q_{ij}}{q_i q_j}\right]$$

$$= \frac{\prod_{ij \in E} q_{ij}}{\prod_{i \in V} q_i^{(d_i - 1)}}$$

degree of edge

$$\text{So } H_Q = -\sum_{\substack{\text{values} \\ \text{of pairs}}} \left[\sum_{ij \in E} q_{ij} \log q_{ij}\right] + \sum_{\text{values}} \left[(d_i - 1) \sum_{i \in V} q_i \log q_i\right]$$

tractable:

We can also compute

$$-E_Q \log P$$

<u>recall</u>:

$$P(H|X) = \frac{1}{Z} \exp\left[-\sum_{ij} \theta_{ij}(H_i, H_j) - \sum_i \theta_i(H_i) \right]$$

$$-\log P = \log Z + \sum_{ij} \theta_{ij}(H_i, H_j) + \sum_i \theta_i(H_i)$$

vars; x_i in other handw notes; sorry!

constant - not a prob, cause we're minimizing!

So

$$-E_Q \log P = \log Z + \sum_{\substack{\text{values}\\ \text{of pairs}}} \left[\sum_{ij \in E} q_{ij} t_{ij} \right] + \sum_{\text{values}} \left[(d_i - 1) \sum_{i \in V} q_i t_i \right]$$

Here I used the change of var

$$t_{ij} = \theta_{ij} + \theta_i + \theta_j$$

$$t_i = \theta_i$$

to get an expression that looks like entropy.

Now: we want to min

$$E_Q \log Q - E_Q \log P$$

for Q some fixed tree. (which we chose)

Recall the q_{ij}, q_i, q_j are marginals

so we must

$$\min \sum_{\text{values}} \left[\sum_{ij \in E} q_{ij} \left[\log q_{ij} + t_{ij} \right] \right] + \sum_{\text{values}} \left[\sum_i (d_i - 1) \cdot q_i \left[\log q_i + t_i \right] \right]$$

s.t. $\sum_i q_{ij} = q_j$; $\sum_j q_{ij} = q_i$; $\sum q_i = 1$

(cause these are marginals)

write

λ_{ε_i} for LM's assoc. with $\sum_j q_{v_{ij}} = q_{v_i}$

λ_{ε_j} " " $\sum_i q_{v_{ij}} = q_{v_j}$

λ_{v_i} LM " $\sum_i q_{v_i} = 1$.

(Notice λ_{ε_i} is a vector
λ_{v_i} scalar

write Lagrangian $\mathcal{L}$.

at stationary point

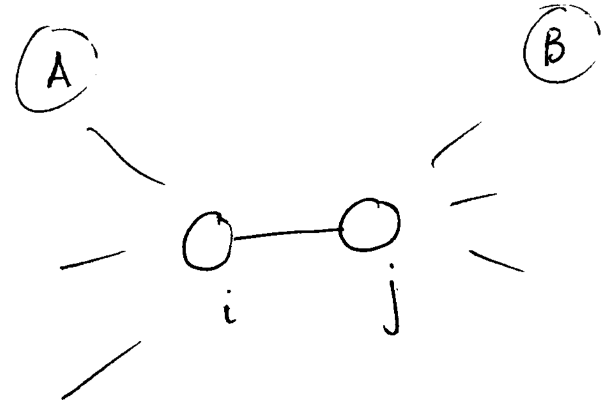
$$\left[\frac{\partial \mathcal{L}}{\partial q_{v_{ij}}}\right]_{uv} = 0 = \left[\log q_{v_{ij}}\right]_{uv} + 1 + \left[\tau_{ij}\right]_{uv} + \underbrace{\sum}_{\text{all incoming edges to } i}\left[\lambda_{\varepsilon_i}\right]_u + \underbrace{\sum}_{\text{all inc. to } j}\left[\lambda_{\varepsilon_j}\right]_v$$

u,v'th entry in table

so

$$\left[q_{v_{ij}}\right]_{uv} \propto \left[e^{-\tau_{ij}}\right]_{uv} \cdot \left[e^{\sum \lambda_{\varepsilon_i}}\right]_u \cdot \left[e^{\sum \lambda_{\varepsilon_j}}\right]_v$$

compare with B.P. eqns



$$\left[q_{ij}\right]_{uv} \propto \left[\psi_{ij}\,\varphi_i\,\varphi_j\right]_{uv} \cdot \left[\prod_{\substack{\text{all inc}\\ \text{to } i}} m_{ia}\right]_u \cdot \left[\prod_{\substack{\text{all inc}\\ \text{to } j}} m_{jb}\right]_v$$

<u>Conclusion</u>

- Messages = log L.M.s

<u>Two outcomes</u> :

1) We can fit a variational model of a single tree (MP as above)

2) Loopy BP "like" fitting var m of tree without worrying about tree

Now what is happening in terms of M.P. ?

→ fix a tree.

→ Interpret MP ≡ convex hull of all states that can arise in this repn of G.M.

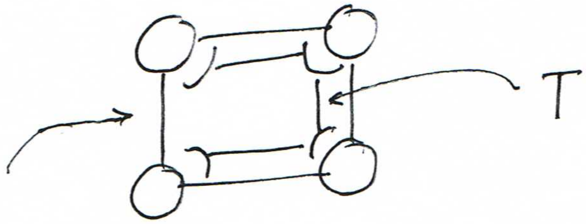
→ ~~We must have that~~ Call the polytope that satisfies $\sum_j q_{ij} = q_i$; $\sum_i q_{ij} = q_j$; $\sum q_i = 1$ the <u>Local Polytope</u> $= \mathcal{L}p$.

→ Notice

$$MP \subset \mathcal{L}p.$$

Notice that for choice of tree T, a lot of $q_{ij} = q_i \cdot q_j$ (cause the vars are ~~indep~~ cond indep given parents

eg.



then in T, these two have $q_{ij} = q_i \cdot q_j$

So we are finding the q_{ij}, q_i, q_j <u>that</u>

- are in $\mathcal{L}P$.
- meet indep constraints implied by T
- minimize

$$-E_Q \log P + E_Q \log Q$$

(then extract info from q).

Extracting info from Q.

- if we're lucky, q_{ij} are integer. (might be a vert of M_p!). → nothing to do
- else, it's a tree; → max product

Idea:

rather than

$$\min \; -E_Q \log P + E_Q \log Q$$

for Q a tree,

do it for $Q \in \mathbb{L}_P$.

→ How do we get $E_Q \log Q$?

Here is one strategy

- drop the tree
- ~~compu~~ fit q by using the expression for $E_Q \log Q$ that came from tree

$$E_Q \log Q \simeq -\sum_{\text{edges}} \underbrace{\sum_{\text{values of pairs}}}_{\text{notice I flipped order}} \left[q_{v_{ij}} \log q_{v_{ij}} \right] + \sum_{\text{verts}} \sum_{\text{values}} \left[(d_i - 1) \sum_j q_{v_i} \log \right.$$

$$= H_Q^?$$

This isn't the true exp. for $E_Q \log Q$, but its easy to eval.

loopy b.p. $\equiv$ min $E_Q \log P - H_Q^?$

s.t. $Q \in \mathcal{L}P$.

Notice the <u>form</u> of the costfunction

$$\underset{\substack{\uparrow \\ \text{linear in } Q.}}{E_Q \log P} \;-\; \underset{\substack{\uparrow \\ \text{Some property of } Q \text{ that approx} \\ \text{entropy.}}}{\overset{?}{H}_Q}.$$

Notice also that we're identifying points in LP (or MP) with probability distributions. It turns out that we can formalize this

The exponential family:

any p.d. that is written as

$$p(x) = \exp\left[\theta^T \phi(x) - A(\theta)\right]$$

(for our purposes - other possible.)

$$\exp\big[\underbrace{\theta^T}_{\text{canonical parameters}} \underbrace{\phi(x)}_{\text{sufficient statistics}} - \underbrace{A(\theta)}_{\text{log partition fn} \equiv \text{log normalizing con}}\big]$$

We will confine attn to case where:

- $\phi(x)$ are linearly independent (no real issue here, just creates a lot of if's, and's, bu

- θ is such that

$$A(\theta) = \log \int \exp[\theta^T \phi(x)] < \infty$$

&

Examples:

1D Normal dist:

$$\exp\left[\cancel{(\alpha, \beta)} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^T (x^2, x) - A(\theta) \right]$$

here $\alpha < 0$; $\text{std} = \frac{-1}{2\alpha}$.

$$\text{mean} = \left(\frac{1}{2\alpha}\right)^2 . \beta .$$

multi D ND:

follows easily.

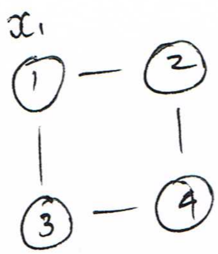
Poisson dist:

- recall this is dist on ~~the~~ non-neg integers
- $p(k) = \lambda^k \frac{e^{-\lambda}}{k!}$ rate/intensity.

$$p(k) = \exp\left[\alpha \cdot k - A(\alpha)\right]$$

$$\alpha = \log \lambda \quad , \text{ etc.}$$

Discrete MRF :



etc. $\mathbb{1}(x_i)$ is 1-hot vector
$\mathbb{1}(x_i, x_j)$ is 1-hot table, straightened in vector

$$p(x) = \exp\left[\theta^T \begin{bmatrix} \mathbb{1}(x_1) \\ \vdots \\ \mathbb{1}(x_i, x_j) \\ \vdots \end{bmatrix} - A(\theta)\right]$$

$A(\theta)$ is extremely interesting

$$\nabla_\theta A = \nabla_\theta \left[\log \int e^{\theta^T \varphi} \cdot dx \right]$$

$$= \frac{1}{\int e^{\theta^T \varphi} dx} \cdot \int \varphi e^{\theta^T \varphi} dx$$

$$= E_p[\phi]$$

(recall - we've seen something like this before when talking about max-likelihood $\equiv$ max entropy.

Now assume we have some ϕ (likely indicator fns in our case).

We can define

$$\Lambda : \Theta \longrightarrow M$$

marginal polytope

$$\Lambda(\theta) = E_\theta[\phi]$$

this is in M, cause M is all possible expects of

Thm:

Λ is 1-1 (assuming ϕ are linearly indep).

(proof in Wainwright - mildly technical).

Now we want to consider dual of $A(\theta)$.

$$A^*(\mu) = \sup_{\theta \in \Theta} \left[\langle \mu, \theta \rangle - A(\theta) \right]$$

Note this is a function of μ.
Known as a conjugate dual

Why is Λ 1-1 ?

(sketch of proof - details in Wainwright)

$A(\theta)$ is convex

we must show for any μ, there is some Θ st $E_{p(x;\theta)}[\phi] = \mu$.

BUT $E_{p(x;\theta)}[\phi] = \nabla_{\theta} A(\theta)$.

under very mild conditions, map

$x \to \frac{df}{dx}$ is 1-1 for x convex

- proof by drawing!

$A(\theta)$ is a <u>convex</u> function of θ.

<u>recall</u> $$\frac{\partial A}{\partial \theta_i} = e^{-A} \cdot \int e^{\theta^T \varphi} \cdot \varphi_i \, dx$$

<u>so</u> $$\frac{\partial^2 A}{\partial \theta_i \partial \theta_j} = e^{-A} \cdot \int e^{\theta^T \varphi} \varphi_i \varphi_j \, dx - \left[e^{-A} \cdot e^{-A} \int e^{\theta^T \varphi_j} \varphi_j \, dx \right] \left[\int e^{\theta^T \varphi} \varphi_j \, dx \right]$$

$$= E_p[\varphi_i \varphi_j] - E_p[\varphi_i] E_p[\varphi_j]$$

$$= \mathrm{cov}(\varphi_i, \varphi_j)$$

so $$H_A = \mathrm{cov\,mat}[\varphi]$$

this is positive definite under our conditions (φ linearly indep.)

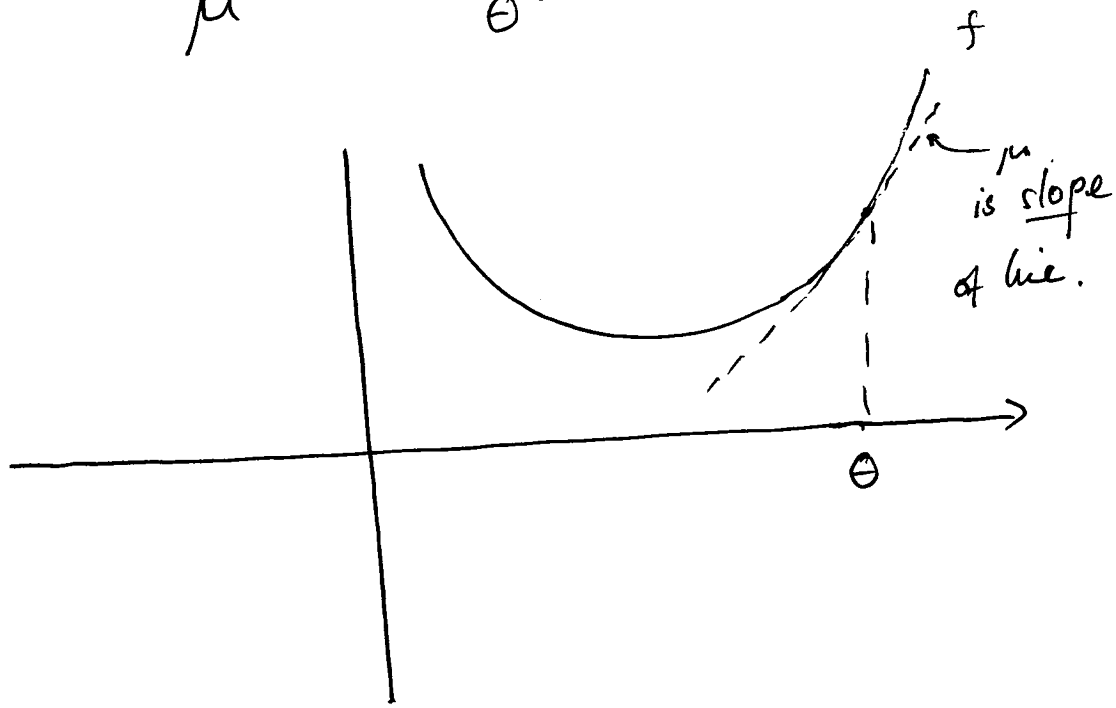
Now consider conjugate dual.

$$f^*(\mu) = \sup_{\theta} \left[\langle \mu, \theta \rangle - f(\theta) \right]$$

for convex f.

- assume f <u>differentiable</u>

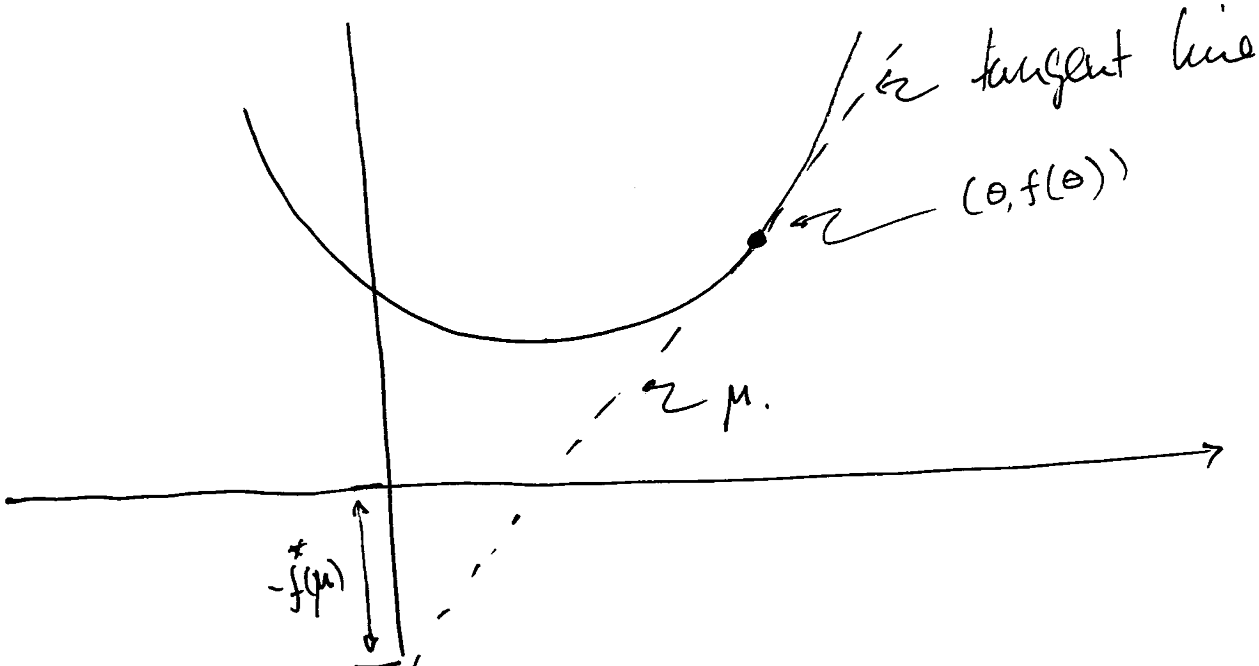
then $\mu = \nabla_\theta f$



Common visualization

- choose $\mu \rightarrow$ what is $f^*(\mu)$?
- consider θ such that

$$\mu = \frac{df}{d\theta}$$



tangent line has slope. μ.

passes through $(\theta, f(\theta))$

$\therefore \quad y = \mu x + (f(\theta) - \mu\theta)$

$\therefore$ at $x = 0$, ~~the~~ $y = -f^*(\mu)$

for f convex, f^* is convex. (19)

Show for $f \in C^2$, but generally true.

<u>Proof</u>: $f^*(p) = \sup_x [xp - f(x)]$

f diff, convex so

$$p = \frac{df}{dx} \quad \text{at sup.}$$

$\frac{df}{dx}$ is a <u>function</u>, and ~~each~~ is 1-1

so g st $g \circ \frac{df}{dx} = \text{Id}$ exists

$$g(p) = x. \quad \text{at sup}$$

so $f^*(p) = p \cdot g(p) - f(g(p))$

$$\frac{df^*}{dp} = p\frac{dg}{dp} + g - f' \cdot \frac{dg}{dp} = g(p)$$

so $\frac{d^2 f^*}{dp^2} = \frac{dg}{dp} = \frac{dx}{dp}$; but $\frac{dp}{dx} = \frac{d^2 f}{dx^2}$

$$\frac{d^2 f^*}{} = \frac{1}{d^2 f/ \ }$$ > 0

all this works in ND as well
(ex: prove it!)

Thm (Fenchel-Moreau).

$$f = (f^*)^*$$

iff

f is proper, lower semi-continuous and convex

OR

$$f \equiv \infty$$

OR

$$f \equiv -\infty$$

Now consider $A(\theta) = \log Z(\theta)$.
for an exp. dist.

1) $$A^*(\mu) = \sup_\theta \left(\langle \theta, \mu \rangle - A(\theta) \right)$$

is defined, convex.

2) for $\mu \in \mathcal{M}$ ← marginal polytope.

write $$\theta(\mu) = \Lambda^{-1}(\mu)$$

then $$A^*(\mu) = -H(p(x; \theta(\mu))).$$

Proof of 2 (sketch):

$$\Lambda^{-1}(\mu) = \theta \quad \text{such that}$$

$$E_{p(x;\theta)}\left[\varphi(x)\right] = \mu$$

(by defn of Λ)

But if $\mu = E_{p(x;\theta)}[\phi] = \nabla_\theta A(\theta)$

then θ is sup

so

$$-H(p(x;\theta(\mu))) = E_{p(x;\theta(\mu))}\left[\langle \theta, \phi(x) \rangle - A(\theta)\right]$$

$$= \langle \theta, \mu \rangle - A(\theta)$$

$$= A^*(\mu) \qquad \text{(cause } \theta \text{ is sup)}$$

3

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\}$$

(A is lower semicontinuous – see notes; then Fenchel-Moreau means

$$(A^*)^* = A .$$

and $$(A^*)^*(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\}$$

<u>Compare</u>: with

$$E_Q = -E_Q \log p + E_Q \log q$$

which we <u>minimized</u> to build var model

marginals of Q

$$E_Q \log p \longrightarrow \langle \theta, \mu \rangle$$

params of

Now we have

$$\sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} . \leftarrow \text{C}$$

is attained at

$$\mu = E_{p(x;\theta)}[\phi]$$

So solving C gives

- log partition function
- set of mean pars
  (for our purposes,
  arg max.)

<u>BUT</u>

$\mathcal{M}$ is hard.

$A^*(\mu)$ is hard.

Now we can unify algs.

<u>Mean field, single tree, etc</u>

for any $\mu \in \mathcal{M}$,

$$A(\theta) \geq \langle \mu, \theta \rangle - A^*(\mu).$$

<u>Now</u> consider $T \subset \mathcal{M}$

corresponding to models that are tractable
$\equiv$ can compute $A^*(\mu)$

then solve

$$\sup_{\mu \in T} \{ \langle \theta, \mu \rangle - A^*(\mu) \} = A_{MF}(\theta)$$

same as our exp but w - sign of max

must have $A_{MF}(\theta) \leq A(\theta)$

loopy BP

1) recall for a tree structured model

$$H(q) = -\sum_{\substack{\cancel{\text{values}} \\ i \in \text{verts}}} \left[\sum_{x \in \text{values}} q_i(x_i) \log q_i(x_i) \right] - \sum_{i,j \in \text{edges}} \left[\sum_{x_i, x_j \in \text{values}} q_{ij}(x_i, x_j) \cdot \log \left[\frac{q_{ij}(x_i, x_j)}{q_i(x_i)\, q_j(x_j)} \right] \right]$$

2) approximate

$$A^*(\mu) \approx -H(\mu)$$

computed using tree expression

Bethe approx

3) $\mathbb{L} \supset \mathcal{M}$

local polytope, consistency constraints for pairwise marginals

4) Solve

$$\sup_{\mu \in \mathcal{L}} \left\{ \langle \theta, \mu \rangle + H_B(\mu) \right\} = A_{LBP}(\theta)$$

we must have

$$A_{LBP}(\theta) \geq A(\theta)$$

But we can now explore other approximations:

- eg. insert constraints so that

$$\mathcal{L} \supseteq \mathcal{L}^c \supseteq \mathcal{M}$$

defined by marginal constraints

marginal + extra

true marginal polytope

heres one construction.

Assume some vector μ, which *might* be in $\mathcal{M}$

Construct the matrix

marg for first var

$$M = \begin{bmatrix} 1 & \mu_1 & \mu_2 & \cdots \\ \mu_1 & & \mu_{12} & \\ \mu_2 & \mu_{21} & & \\ \vdots & & & \end{bmatrix}$$

M is a covariance matrix

so $M \succeq 0$