

Variational Inference

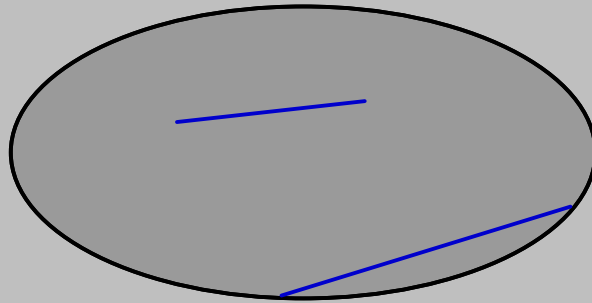
(basic principles)

Ross A. Lippert
MIT Department of Mathematics

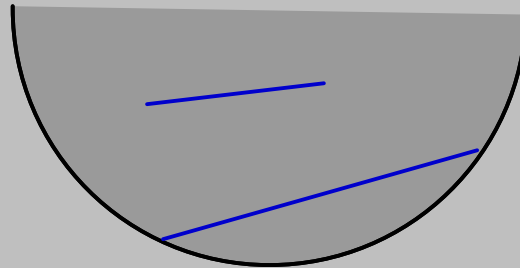
November 3, 2005

Concave/convex preliminaries

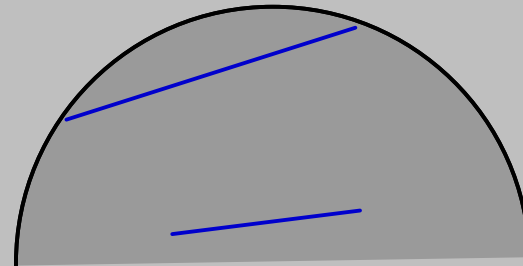
Convex sets



Convex functions



Concave functions



Concave/convex preliminaries (cont.)

Due to my inability to cope with signs, we will deal with **concave**, rather than **convex** functions.

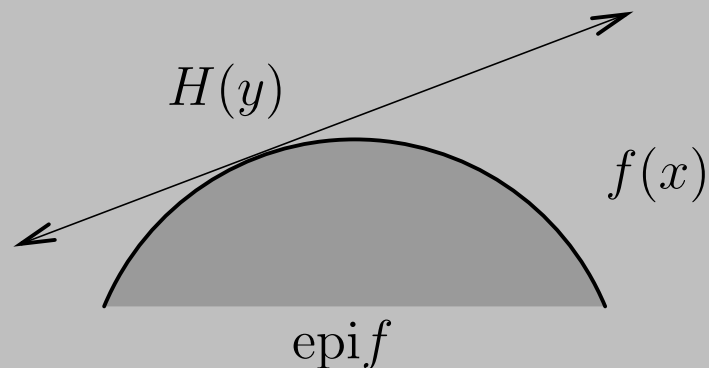
A general **concave** function $f : \mathbb{R}^n \rightarrow [-\infty, \infty)$,

$$f\left(\sum_i \rho_i x_i\right) \geq \sum_i \rho_i f(x_i), \quad \text{where } \sum_i \rho_i = 1, \rho_i \geq 0.$$

definition	notes
$\text{dom} f = \{x : f(x) > -\infty\}$	is a convex set
$\text{epi} f = \{(x, e) : e \leq f(x)\} \subset \mathbb{R}^{n+1}$	a technically useful convex set
f is <i>closed</i> \leftrightarrow $\text{epi} f$ is closed	a technical condition
f is <i>proper</i> \leftrightarrow $\text{dom} f \neq \emptyset$	a technical condition

Concave gradients and “subgradients”

If $\nabla f(x)$ exists, $f(y) \leq H(y) = f(x) + (y - x)^t \nabla f(x)$:



Concave functions **minorize** their tangent hyperplanes.

Generalize *gradient* with *subgradient* at x is $\partial f(x)$:

$$s \in \partial f(x) : f(y) \leq f(x) + (y - x)^t s,$$

$\partial f(x) = \emptyset$ when $x \notin \text{dom } f$ (conversely when f is **closed**).

For just about anything we do: f is differentiable

$$\partial f(x) = \{\nabla f(x)\}$$

The Fenchel conjugate

a.k.a. **Fenchel/Fenchel-Legendre/Legendre conjugate/transform/dual** ^a

The Fenchel conjugate f^* is

$$f^*(z) = \inf_y \{y^t z - f(y)\}.$$

Easy checks: f^* is **concave** and $g \leq f \Leftrightarrow g^* \geq f^*$.

(Not so easy check: f^* is **closed**.)

When **closed** and **concave**, $f^{**} = f$:

$$f(y) = \inf_z \{y^t z - f^*(z)\}.$$

Analogy with **Laplace transform**.

functions/operations in $t \leftrightarrow$ functions/operations in s

^aThis use of the word “dual” here is unfortunate

Fenchel conjugate, differentiable case

$$f^*(z) = \inf_y \{y^t z - f(y)\}$$

Minimum (in y , fixed z) when $z - \nabla f(y) = 0$, defines $y_*(z)$,

$$\begin{aligned} f^*(z) &= (y_*(z))^t z - f(y_*(z)) \quad (\text{when min is attained}) \\ f^*(z) &= \begin{cases} y^t z - f(y) & \text{where } \nabla f(y) = z \\ -\infty & \text{no such } y \end{cases} \end{aligned}$$

General way to express the f, f^* relation is, (**write this down**)

$$\left. \begin{aligned} & \left\{ \begin{array}{ll} f^*(z) + f(y) = y^t z & \text{if } z \in \partial f(y) \\ f^*(z) + f(y) < y^t z & \text{else} \end{array} \right\} \end{aligned} \right\}$$

called *the Fenchel-Young theorem*.

A simple example

Let $f(y) = -y^2$, $y \in \mathbb{R}$.

Finding the conjugate, $f^*(z) = \inf_y \{yz - f(y)\}$,

$$z - \nabla f(y) = 0$$

$$\Rightarrow z(y) + 2y = 0 \quad (\text{thus } y(z) = -\frac{1}{2}z)$$

$$\begin{aligned} f^*(z) &= y(z)z - f(y(z)) \\ &= \left(-\frac{1}{2}z\right)z + \left(\frac{1}{2}z\right)^2 \\ &= -\frac{1}{4}z^2 \end{aligned}$$

The Fenchel-Young thm

$$\begin{aligned} f(y) + f^*(z) &= -y^2 - \frac{1}{4}z^2 \\ &= yz - \left(y + \frac{1}{2}z\right)^2 \end{aligned}$$

For more convex analysis:
Borwein and Lewis's *Convex analysis and non-linear optimization*

The intro to convex analysis is over!
Now let's talk about inference...

Cummulants and concavity

The goal:

$$\text{compute } Z = \sum_{x \in S} \Phi(x), \text{ where } |S| \text{ is very big.}$$

(think of S as a sample space, with sample point x)

The goal (let $\phi = -\log(\Phi)$)

$$-\log(Z) = A = -\log \left(\sum_{x \in S} \exp(-\phi(x)) \right).$$

A is a **concave** function of ϕ . (divergence of sum $\Rightarrow A = -\infty$).

Note, $A(\phi + C1) = A(\phi) + C$ (C is any constant).

Let's play the conjugation game on $A(\phi)$.

Conjugation of $A(\phi)$

With $\phi, \rho : S \rightarrow \mathbb{R}$, $A(\phi) = -\log(\sum_{x \in S} \exp(-\phi(x)))$

$$A^*(\rho) = \inf_{\phi} \left\{ \sum_{x \in S} \phi(x) \rho(x) - A(\phi) \right\},$$

Optimality when

$$\begin{aligned} \rho(x) &= \frac{\partial A(\phi)}{\partial \phi(x)} = \exp(A(\phi) - \phi(x)) \\ &\left(\text{only possible if } \sum_{x \in S} \rho(x) = 1 \right) \\ \Rightarrow \phi(x) &= A(\phi) - \log(\rho(x)) \\ \Rightarrow A^*(\rho) &= \begin{cases} \sum_x \rho(x) \log(1/\rho(x)) & \rho \in \mathcal{P}(S) \\ -\infty & \text{else} \end{cases} \end{aligned}$$

where $\mathcal{P}(S)$ are *distributions* on S (i.e. $\rho(x) \geq 0$, $\sum_x \rho(x) = 1$).

$A^*(\rho)$ conclusions

$\text{dom}A^* = \mathcal{P}(S)$ and $A^*(\rho) = \text{entropy}$.

Optimum attained when $\rho(x) = \exp(A(\phi) - \phi(x))$.

Same game with an exponential family: $\phi(x) = \theta^t T(x)$,

$$A(\theta) = A(\theta^t T) \equiv -\log \left(\sum_{x \in S} \exp(-\theta^t T(x)) \right)$$

$A(\theta)$ is concave function in θ (b.c. $\frac{\partial^2 A}{\partial \theta^2} = -\text{covariance of } T$).

With $\theta, \mu \in \mathbb{R}^n$,

$$A^*(\mu) = \inf_{\theta} \{ \theta^t \mu - A(\theta) \}$$

Optimum when $(\exists \theta : \mu \in \partial A(\theta))$

$$\mu = \frac{\partial A(\theta)}{\partial \theta} = \sum_x T(x) \exp(A(\theta) - \theta^t T(x)) \equiv E_{\theta}[T]$$

maybe θ exists, maybe θ does not.

Conjugation game on expfams

$\theta : \mu = E_\theta[T]$ does not exist then $A^*(\mu) = -\infty$ (i.e. $\mu \notin \text{dom}A^*$).

$\theta : \mu = E_\theta[T]$ exists, then $A^*(\mu) > -\infty$ (i.e. $\mu \in \text{dom}A^*$).

Let $\rho_\theta(x) \equiv \exp(A(\theta) - \theta^t T(x))$

and plug in $\mu = \sum_x T(x) \exp(A(\theta) - \theta^t T(x)) = \sum_x T(x) \rho_\theta(x)$

$$\begin{aligned} A^*(\mu) &= \theta^t \mu - A(\theta) \\ &= \theta^t \left(\sum_x T(x) \rho_\theta(x) \right) - A(\theta) \\ &= \sum_x [\theta^t T(x)] \rho_\theta(x) - A(\theta) \\ &= \sum_x [A(\theta) - \log(\rho_\theta(x))] \rho_\theta(x) - A(\theta) \\ &= \sum_x \rho_\theta(x) \log(1/\rho_\theta(x)) \\ &= \text{entropy of } \rho_\theta. \end{aligned}$$

$A^*(\mu)$ conclusions

$\text{dom}A^* =$ realizable means of T

For $\mu \in \text{dom}A^*$, $A^*(\mu) =$ entropy.

$\text{dom}A^*(\mu)$ is convex but otherwise can be **hard** to determine.

What have we learned?

- Cummulants conjugate to entropies.
- Approximations to cummulants will conjugate to approximations to entropy (and vice versa).
- Insights into entropy should be useful.

Jordan and Wainwright: focuses on approximations in terms of μ -space and entropy.

Bi-conjugation

Remember how $A^{**} = A$? Suppose $B^*(\mu) \sim A^*(\mu)$, then **we hope** that

$$\begin{aligned} B(\theta) &\sim A(\theta) \\ \frac{\partial}{\partial \theta} B(\theta) &\sim \frac{\partial}{\partial \theta} A(\theta) \end{aligned}$$

Suppose $B(\theta)$ is a cumulant for some: ρ'_{θ}

$$\begin{aligned} B(\theta) &= \inf_{\mu} \{ \theta^t \mu - B^*(\mu) \} \\ &= \inf_{\theta'} \{ \theta^t E_{\theta'}[T] - H(\rho'_{\theta'}) \} \\ &= \inf_{\theta'} \left\{ \sum_x \theta^t T(x) \rho'_{\theta'}(x) - H(\rho'_{\theta'}) \right\} = \inf_{\theta'} \text{free energy} \\ &= \inf_{\theta'} \left\{ \sum_x [\theta^t T(x) - \log(1/\rho'_{\theta'}(x))] \rho'_{\theta'}(x) \right\} \\ B(\theta) - A(\theta) &= \inf_{\theta'} \left\{ \sum_x \log(\rho'_{\theta'}(x)/\rho_{\theta}(x)) \rho'_{\theta'}(x) \right\} = \inf_{\theta'} D(\rho'_{\theta'} || \rho_{\theta}). \end{aligned}$$

Bi-conjugation recovers the usual **idols of worship**.

Approximations

Approximations to $A(\theta)$ can be obtained either directly, **or** by approximating $A^*(\mu)$ and conjugating back.

Conjugation reinterprets approximations.

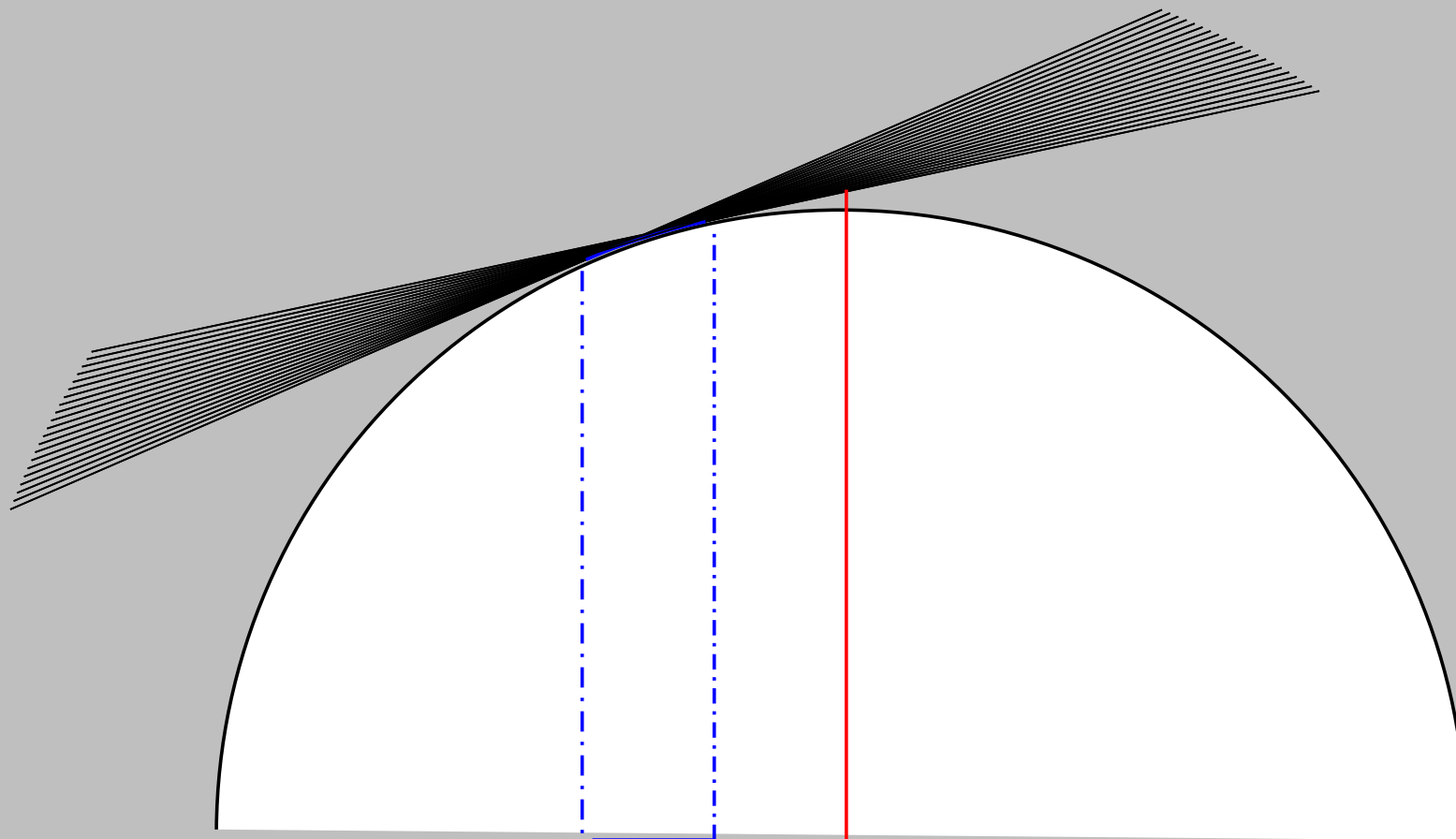
minorants \rightarrow^* majorants

majorants \rightarrow^* minorants

The basic flavors of approximation:

- **upper bound:** Replace $A(\theta)$ with a majorant. Common majorizations come from *mean field*.
- **lower bound:** Replace $A(\theta)$ with a minorant. The literature is pretty sketchy. Convexified Bethe (i.e. tree weighted BP) would be an example.
- **heuristic:** Inspiration gives a similar $\tilde{A}(\theta) = A(\theta)$. We'd like $\text{dom}\tilde{A}^* \sim \text{dom}A^*$. Examples include the regular Bethe and loopy BP.

Mean field



Mean field (upper bound)

Partition variables: $A(\theta_1, \theta_2)$. Concavity says (for any ψ, θ_1, θ_2)

$$\begin{aligned} A(\theta_1, \theta_2) &\leq A(\psi, 0) + \mu^t \begin{pmatrix} \theta_1 - \psi \\ \theta_2 \end{pmatrix} \quad \text{where } \mu(\psi) = \nabla A(\psi, 0) \\ &= A(\psi, 0) - \mu^t \begin{pmatrix} \psi \\ 0 \end{pmatrix} + \mu^t \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \equiv B(\mu(\psi), \theta_1, \theta_2) \end{aligned}$$

Say $A(\psi, 0)$ is tractable:

- Can compute $A(\psi, 0)$
- Can compute $\mu(\psi) = \nabla A(\psi, 0) = (E_\psi[T_1], E_\psi[T_2])$
- Can compute $E_\psi[T_i T_j^t]$ for $i, j \in \{1, 2\}$

$$\left(E_\psi[g] = \sum_{x \in S} g(x) \exp(A(\psi, 0) - \psi^t T_1(x)) \right)$$

Best bound is

$$A(\theta_1, \theta_2) \leq \inf_{\psi} \{B(\nabla A(\psi, 0), \theta_1, \theta_2)\} \equiv \tilde{A}(\theta_1, \theta_2)$$

Mean field equations

$$\tilde{A}(\theta_1, \theta_2) = \inf_{\psi} \left\{ A(\psi, 0) - \left(\frac{\partial A(\psi, 0)}{\partial \theta_1} \right)^t \psi + \left(\frac{\partial A(\psi, 0)}{\partial \theta_1} \right)^t \theta_1 + \left(\frac{\partial A(\psi, 0)}{\partial \theta_2} \right)^t \theta_2 \right\}$$

not convex, but we can still try,

$$0 = \nabla_{\psi} B = \left(\frac{\partial^2 A(\psi, 0)}{\partial \theta_1^2} \right) (\theta_1 - \psi) + \left(\frac{\partial^2 A(\psi, 0)}{\partial \theta_1 \partial \theta_2} \right) \theta_2$$

i.e.

$$\begin{aligned} \frac{\partial^2 A(\psi, 0)}{\partial \theta_1^2} &= \left. \frac{\partial^2 A(\theta_1, \theta_2)}{\partial \theta_1^2} \right|_{\psi, 0} &= -E_{\psi}[T_1 T_1^t] + E_{\psi}[T_1] E_{\psi}[T_1^t] \equiv -C_{\psi} \\ \frac{\partial^2 A(\psi, 0)}{\partial \theta_1 \partial \theta_2} &= \left. \frac{\partial^2 A(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2} \right|_{\psi, 0} &= -E_{\psi}[T_1 T_2^t] + E_{\psi}[T_1] E_{\psi}[T_2^t] \equiv -D_{\psi} \\ 0 &= -C_{\psi}(\theta_1 - \psi) - D_{\psi} \theta_2 \\ \psi &= \theta_1 + C_{\psi}^{-1} D_{\psi} \theta_2 \end{aligned}$$

Fixed point iteration might converge, might not. We have the **gradient** regardless.

Mean field with mean parameters

Let $A_1(\psi) = A(\psi, 0)$. $\mu_1(\psi) = \frac{\partial A(\psi, 0)}{\partial \theta_1}$ is invertible.

The fixed point equations, c.o.v. to $\nu = \mu_1(\psi)$,

$$\begin{aligned} \nu &= \mu_1 \left(\theta_1 + C_{\mu_1^{-1}(\nu)}^{-1} D_{\mu_1^{-1}(\nu)} \theta_2 \right) \\ &= \mu_1 \left(\theta_1 + \frac{\partial}{\partial \nu} E_{\mu_1^{-1}(\nu)} [T_2^t \theta_2] \right) \end{aligned}$$

Last part from,

$$\begin{aligned} \frac{\partial}{\partial \nu} E_{\mu_1^{-1}(\nu)} [T_2^t] &= \left[\frac{\partial \mu_1(\psi)}{\partial \psi} \right]^{-1} \frac{\partial}{\partial \psi} E_\psi [T_2^t] \\ \frac{\partial \mu_1(\psi)}{\partial \psi} &= \frac{\partial^2 A(\psi, 0)}{\partial \theta_1^2} \\ \frac{\partial}{\partial \psi} E_\psi [T_2^t] &= \frac{\partial^2 A(\psi, 0)}{\partial \theta_1 \partial \theta_2} \end{aligned}$$

Sometimes $E_{\mu_1^{-1}(\psi)} [T_2^t \theta_2]$ is more convenient to calculate (like in the next example).

Simple Ising model example

From the Wainwright and Jordan paper.

$$\exp(-A(\theta)) = \sum_{\bar{x} \in \{0,1\}^n} \exp \left(- \sum_i \theta_i x_i - \sum_{ij} \theta_{ij} x_i x_j \right).$$

T_1 's are x_i and T_2 's are $x_i x_j$.

$$A(\psi, 0) = - \log \prod_{i=1}^n (1 + \exp(-\psi_i)) = - \sum_{i=1}^n \log(1 + \exp(-\psi_i))$$

$$\mu_1(\psi) = \frac{\exp(-\psi)}{1 + \exp(-\psi)} \quad (\text{componentwise})$$

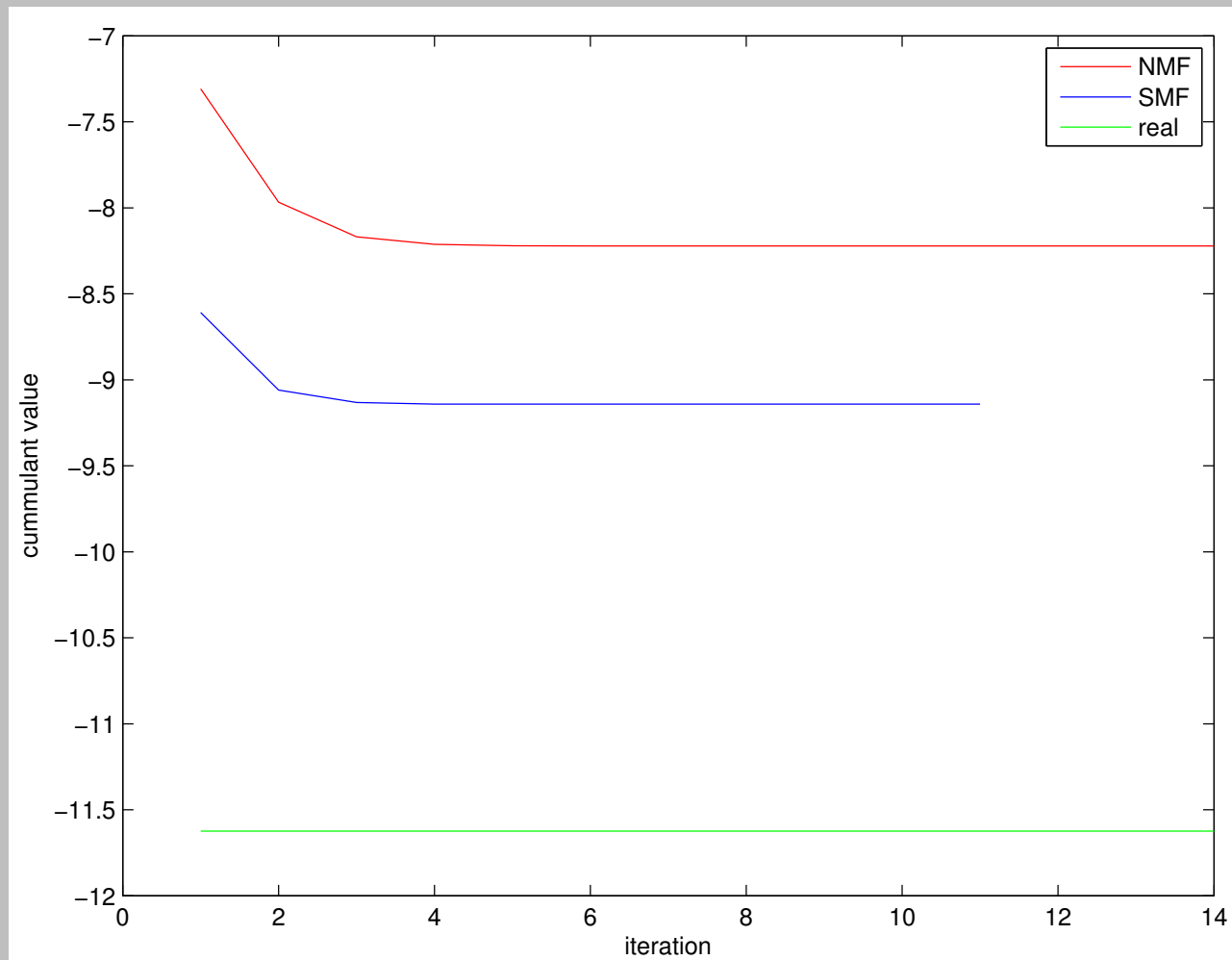
$$E_{\mu_1^{-1}(\nu)}[x_i x_j] = \nu_i \nu_j$$

$$E_{\mu_1^{-1}(\nu)} \left[\sum_{ij} \theta_{ij} x_i x_j \right] = \sum_{ij} \theta_{ij} \nu_i \nu_j$$

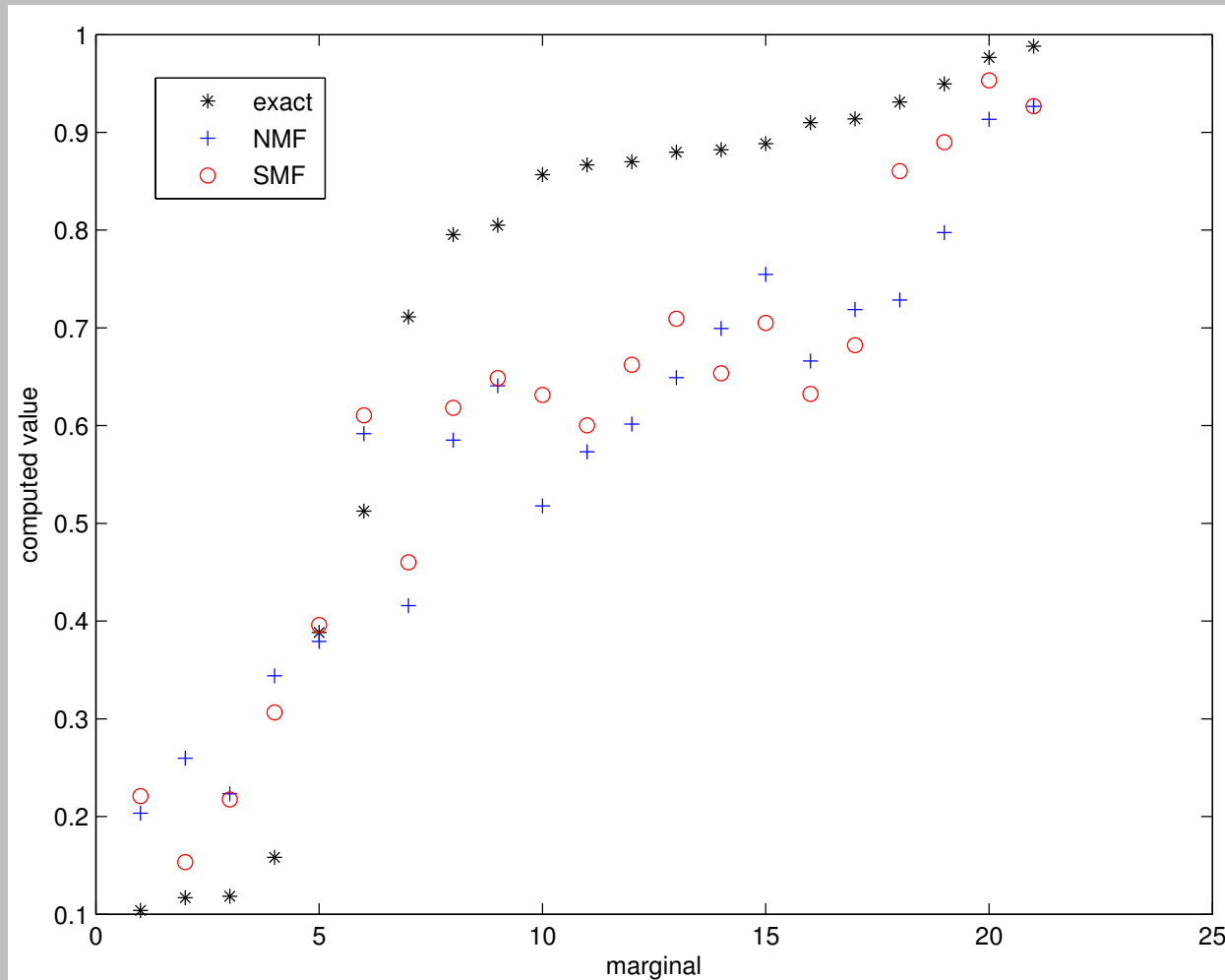
and the update is

$$\nu \leftarrow \mu_1(\theta + [\theta_{ij}]\nu).$$

A comparison of MF Ising approximations



A comparison of MF Ising approximations



A common mean field trick

The meanfield approach can also be used to approximate one distribution with another.

$$\begin{aligned}p_{\psi}(y) &= \exp(A_1(\psi) - \psi^t T_1(y)) \\p_{\theta_2}(y) &= \exp(A_2(\theta_2) - \theta_2^t T_2(y))\end{aligned}$$

with A_1 tractable, A_2 intractable.

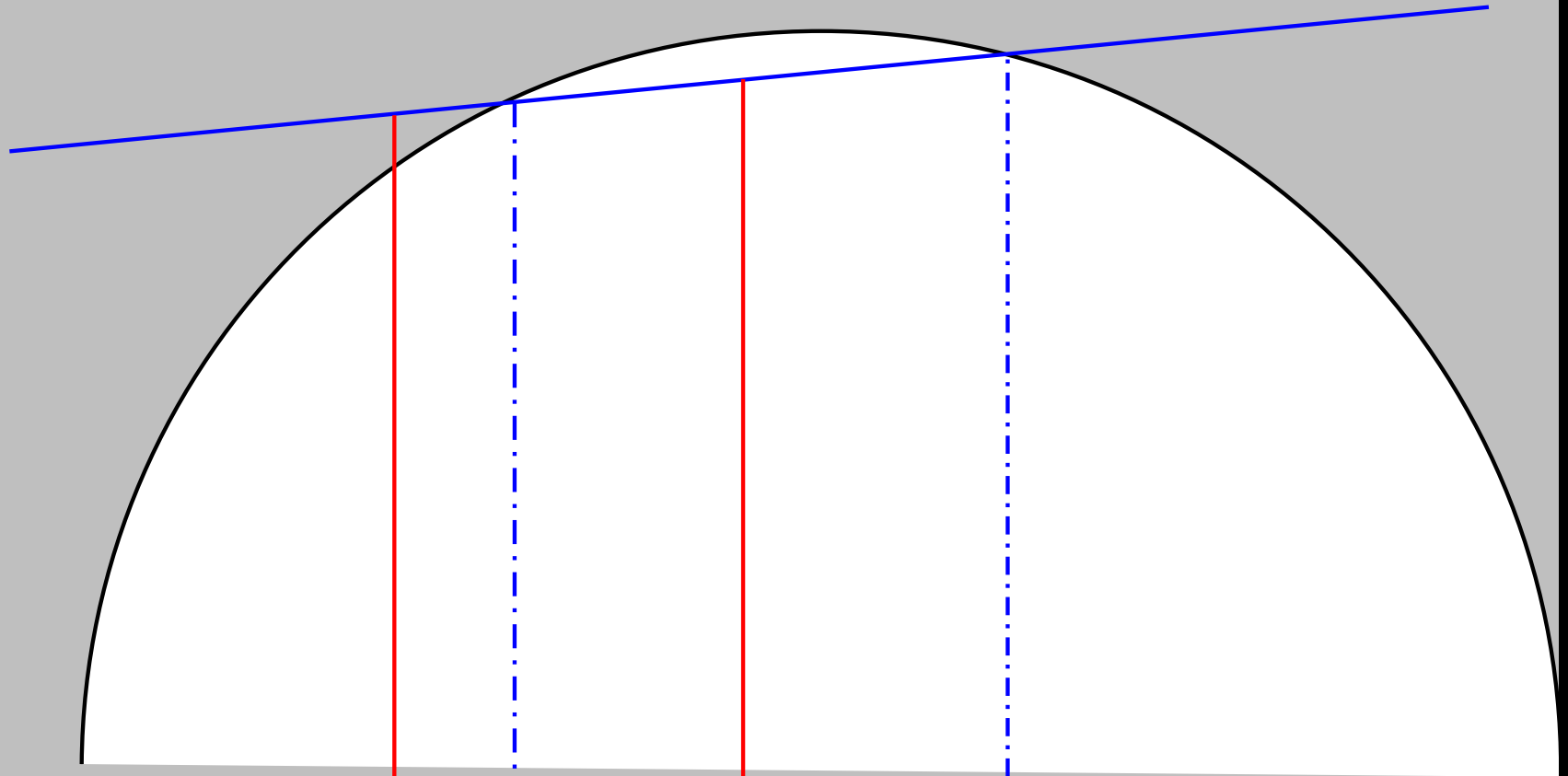
$$A(\theta_1, \theta_2) \equiv -\log \left(\sum_y \exp(-\theta_1^t T_1(y) - \theta_2^t T_2(y)) \right)$$

thus $A_2(\theta_2) = A(0, \theta_2)$. Upper bound by MF: $\psi \leftarrow C_{\psi}^{-1} D_{\psi} \theta_2$ has a cute form

$$\begin{aligned}\psi &\leftarrow \frac{\partial}{\partial \nu} E_{\mu_1^{-1}(\nu)} [T_2^t \theta_2] \\ \nu &\leftarrow E_{\psi} [T_1]\end{aligned}$$

Mixture model approximations are sort of like this.

Loopy BP



Bethe/Kikuchi/Yedidia approximations

Recall **concavity** means

$$\sum_{i=1}^k \lambda_i A(\theta^i) \leq A\left(\sum_{i=1}^k \lambda_i \theta^i\right) \quad \lambda_i \geq 0, \sum_i \lambda_i = 1.$$

Make a **constraint**: $\sum_i \lambda_i \theta^i = \theta$ (the λ_i fixed with θ^i varying).

- $A(\theta^i)$ are “tractable” (parts of θ^i are 0)
- we can get a lower bound on $A(\theta)$.

Simple case: $n = 3, k = 2, \lambda_i > 0$, with $A(\theta_1, \theta_2, 0)$ and $A(0, \theta_2, \theta_3)$ tractable,

$$\lambda_1 A(\theta_1/\lambda_1, (\theta_2 + \alpha)/\lambda_1, 0) + \lambda_2 A(0, (\theta_2 - \alpha)/\lambda_2, \theta_3/\lambda_2) \leq A(\theta_1, \theta_2, \theta_3).$$

for any $\alpha \in \mathbb{R}$,

(Generally) let $Z_j^i \in \{0, 1\}$ with $1 \leq i \leq k$ and $1 \leq j \leq n$, **constraints** are

$$\begin{aligned} Z_j^i \theta_j^i &= 0 \\ \sum_{i=1}^k \lambda_i \theta^i &= \theta. \end{aligned}$$

Continuing simple example

Best lower bound (in α),

$$\sup_{\alpha} \{ \lambda_1 A(\theta_1/\lambda_1, (\theta_2 + \alpha)/\lambda_1, 0) + \lambda_2 A(0, (\theta_2 - \alpha)/\lambda_2, \theta_3/\lambda_2) \}$$

If $\lambda_i > 0$, $\sum_i \lambda_i = 1$, then [above] convex in α

$$\left. \frac{\partial A(\theta_1/\lambda_1, \psi, 0)}{\partial \psi} \right|_{\psi=(\theta_2+\alpha)/\lambda_1} - \left. \frac{\partial A(0, \psi, \theta_3/\lambda_2)}{\partial \psi} \right|_{\psi=(\theta_2-\alpha)/\lambda_2} = 0$$

for optimality.

$$E_{(\theta_1/\lambda_1, (\theta_2+\alpha)/\lambda_1, 0)}[T_2] - E_{(0, (\theta_2-\alpha)/\lambda_2, \theta_3/\lambda_2)}[T_2] = 0$$

Generally

$$\lambda_i > 0, \sum_i \lambda_i = 1$$

$$\tilde{A}(\theta) = \sup_{\sum \lambda_i \theta^i = \theta, Z_j^i \theta_j^i = 0} \left\{ \sum_i \lambda_i A(\theta^i) \right\} \leq A(\theta).$$

Enforce constraints with Lagrange multipliers, like

$$\tau^t(\theta - \sum_i \lambda_i \theta^i) - \sum_{ij} z_j^i Z_j^i \theta_j^i$$

giving optimality conditions

$$\frac{\partial}{\partial \theta_j^i} \left(\sum_{i=1}^k \lambda_i A(\theta^i) + \tau^t(\theta - \sum_i \lambda_i \theta^i) - \sum_{ij} z_j^i Z_j^i \theta_j^i \right) = 0$$

$$\forall i, j : \lambda_i \frac{\partial A(\theta^i)}{\partial \theta_j} - \lambda_i \tau_j - z_j^i Z_j^i = 0$$

$$\Rightarrow \forall i, j : \text{if } Z_j^i = 0, \frac{\partial A(\theta^i)}{\partial \theta_j} = \tau_j$$

Beliefs, τ , consistent across overlaps.

Big picture

Maximality (in θ^i with constraints) implies “intersecting statistics” have consistent expectations across the pieces.

The τ Lagrange multipliers are *pseudo-marginals* or *beliefs*.

$$\begin{aligned}\sum_i \lambda_i \frac{\partial A(\theta^i)}{\partial \theta} &= \tau \\ \sum_i \lambda_i \theta^i &= \theta \\ Z_j^i \theta_j^i &= 0.\end{aligned}$$

The Fenchel Conjugate view

$$\begin{aligned}
 \tilde{A}^*(\mu) &\geq A^*(\mu) \\
 \tilde{A}^*(\mu) &= \inf_{\theta} \left\{ \theta^t \mu - \sup_{\sum \lambda_i \theta^i = \theta, \theta_j^i Z_j^i = 0} \left\{ \sum_i \lambda_i A(\theta^i) \right\} \right\} \\
 &= \inf_{\theta, \sum \lambda_i \theta^i = \theta, \theta_j^i Z_j^i = 0} \left\{ \theta^t \mu - \sum_i \lambda_i A(\theta^i) \right\} \\
 &= \inf_{\theta_j^i Z_j^i = 0} \left\{ \sum_i \lambda_i (\theta^i)^t \mu - \sum_i \lambda_i A(\theta^i) \right\} \\
 &= \sum_i \lambda_i \left(\inf_{\theta_j^i Z_j^i = 0} \{ \theta^t \mu - A(\theta) \} \right) \\
 &= \sum_i \lambda_i A_i^*(\mu) \quad \text{where } A_i^*(\mu) = \inf_{\theta_j^i Z_j^i = 0} \{ \theta^t \mu - A(\theta) \}.
 \end{aligned}$$

If $A_i^*(\mu)$ are pleasant, this is a good approach.

The Fenchel Conjugate view (cont.)

$$\forall i : A_i^*(\mu) \geq A^*(\mu)$$
$$\text{dom}\tilde{A}^*(\mu) = \bigcap_{i=1}^k \text{dom}A_i^*$$

In *Bethe, Yedidia and Kikuchi*, you can “**cheat**” by *relaxing the constraints* on λ_i .

- Loses the guarantees of global optimality
- Loses lower bounding properties
- Seems to work pretty well on multinomial models

Multinomial example: tree weighted BP

Node values $\in S_0 = \{1, \dots, N\}$, $S = S_0^n$, with parameters $\theta_i(a), \theta_{ij}(a, b)$ at nodes i and $(i, j) \in E$ with $a, b \in S_0$.

For **trees**,

$$A^*(\mu) = \sum_i \sum_{a \in S_0} \mu_i(a) \log \frac{1}{\mu_i(a)} + \sum_{(i,j) \in E} \sum_{a,b \in S_0} \mu_{ij}(a,b) \log \frac{\mu_i(a)\mu_j(b)}{\mu_{ij}(a,b)}$$

with constraints: $\sum_a \mu_i(a) = 1$, and $\sum_b \mu_{ij}(a,b) = \sum_c \mu_{ki}(c,a) = \mu_i(a)$.

Superposition of spanning tree subgraphs, t_1, t_2, \dots, t_k

$$\tilde{A}^*(\mu) = \sum_i \sum_{a \in S_0} \mu_i(a) \log \frac{1}{\mu_i(a)} + \sum_{i < j} \rho_{ij} \sum_{a,b \in S_0} \mu_{ij}(a,b) \log \frac{\mu_i(a)\mu_j(b)}{\mu_{ij}(a,b)}$$

where $\rho_{ij} = \sum_{t_k: (i,j) \in E_{t_k}} \lambda_k$.

In the Bethe approximation, $\rho_{ij} = 1$ (relaxed constraints).

tree weighted BP (cont.)

Bi-conjugation to get $\tilde{A}(\theta)$ and μ_{opt} .

$$\tilde{A}(\theta) = \inf_{\mu} \left\{ \sum_{a,i} \theta_i(a) \mu_j(a) + \sum_{a,b,i,j} \theta_{ij}(a,b) \mu_{ij}(a,b) - \tilde{A}^*(\mu) \right\}$$

with constraints $\sum_a \mu_i(a) = 1$ and $\sum_b \mu_{ij}(a,b) = \sum_c \mu_{ki}(c,a) = \mu_i(a)$.

Lagrange multiplier terms of the form

$$\begin{aligned} & \sum_i \alpha_i [1 - \sum_a \mu_i(a)] \\ & + \sum_{(i,j) \in E} \beta_{ij}(a) [\mu_i(a) - \sum_b \mu_{ij}(a,b)] + \sum_{(j,i) \in E} \beta_{ij}(a) [\mu_i(a) - \sum_c \mu_{ij}(c,a)] \end{aligned}$$

The local optimality conditions are then

$$\begin{aligned} \theta_i(a) - \alpha_i + \sum_{j:(i,j) \in E} \beta_{ij}(a) + \beta_{ji}(a) &= \frac{\partial \tilde{A}^*(\mu)}{\partial \mu_i(a)} \\ \theta_{ij}(a,b) - \beta_{ij}(a) - \beta_{ji}(a) &= \frac{\partial \tilde{A}^*(\mu)}{\partial \mu_{ij}(a,b)} \end{aligned}$$

tree weighted BP (cont.)

$$\tilde{A}^*(\mu) = \sum_i \sum_{a \in S_0} \mu_i(a) \log \frac{1}{\mu_i(a)} + \sum_{i < j} \rho_{ij} \sum_{a, b \in S_0} \mu_{ij}(a, b) \log \frac{\mu_i(a) \mu_j(b)}{\mu_{ij}(a, b)}$$

Absorbing annoying constants into the Lagrange multipliers,

$$\theta_i(a) - \alpha_i + \sum_{j: (i,j) \in E} \beta_{ij}(a) + \beta_{ji}(a) = -\log(\mu_i(a))$$

$$\theta_{ij}(a, b) - \beta_{ij}(a) - \beta_{ji}(a) = -\rho_{ij} \log \frac{\mu_{ij}(a, b)}{\mu_i(a) \mu_j(b)}$$

i.e.

$$\begin{aligned} \mu_i(a) &\propto \exp \left(- \left(\theta_i(a) + \sum_{j \in \mathcal{N}(i)} \beta_{ij}(a) \right) \right) \\ \mu_{ij}(a, b) &\propto \mu_i(a) \mu_j(b) \exp \left([-\theta_{ij}(a, b) + \beta_{ij}(a) + \beta_{ji}(b)] / \rho_{ij} \right) \end{aligned}$$

tree weighted BP (end)

Let $M_{ij}(a) = \exp(-\beta_{ij}(a)/\rho_{ij})$

$$\mu_i(a) \propto \exp(-\theta_i(a)) \prod_{j \in \mathcal{N}(i)} M_{ij}(a)^{\rho_{ij}}$$

$$\mu_{ij}(a, b) \propto \frac{\mu_i(a) \mu_j(b) \exp(-\theta_{ij}(a, b))}{M_{ij}(a) M_{ji}(b)}$$

enforcing the constraints,

$$\begin{aligned} M_{ji}(b) &\propto \sum_a \frac{\mu_i(a) \exp(-\theta_{ij}(a, b))}{M_{ij}(a)} \\ &= \sum_a \exp(-\theta_i(a) - \theta_{ij}(a, b)) \frac{\prod_{k \in \mathcal{N}(i)} M_{ik}(a)^{\rho_{ik}}}{M_{ij}(a)} \\ M_{ij}(a) &\propto \sum_b \exp(-\theta_j(b) - \theta_{ij}(a, b)) \frac{\prod_{k \in \mathcal{N}(j)} M_{jk}(b)^{\rho_{jk}}}{M_{ji}(b)} \end{aligned}$$

which are general update equations for message passing.

Conclusions

- The cummulant is **concave**
- Its Fenchel conjugate is the *mean parametrized entropy*
- Convex analysis has ideas of how to bound concave functions
- Resulting *best bound* problems may or may not be concave/convex
- Insights into the form of the entropy can also motivate approximations