School of Computer Science
Carnegie Mellon

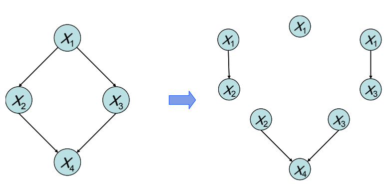
Probabilistic Graphical Models

Variational Inference II:

Mean Field Method and Variational Principle

Junming Yin

Lecture 15, March 7, 2012



Reading:

Recap

- Loopy belief propagation (sum-product) algorithm is a method to find the stationary point of Bethe free energy
  - based on direct approximation of Gibbs free energy
  - will revisit BP and Bethe approximation from another point of view later

- Today, we will look at another approximation inference method based on restricting the family of approximation distribution

Variational Methods

- "Variational": fancy name for optimization-based formulations
  - i.e., represent the quantity of interest as the solution to an optimization problem
  - *approximate* the desired solution by *relaxing/approximating* the *intractable* optimization problem

- Examples:
  - Courant-Fischer for eigenvalues: $\lambda_{\max}(A) = \max_{\|x\|_2=1} x^T A x$
    -
  - Linear system of equations: $Ax = b, A \succ 0, x^* = A^{-1}b$
    - variational formulation:

$$x^* = \arg\min_x \left\{ \frac{1}{2} x^T A x - b^T x \right\}$$

    - for large system, apply conjugate gradient method

Inference Problems in Graphical Models

- Undirected graphical model (MRF):

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

- The quantities of interest:
  - marginal distributions: $p(x_i) = \sum_{x_j, j \neq i} p(x)$
  - normalization constant (partition function): Z
- Question: how to represent these quantities in a variational form?

Variational Formulation

$$KL(Q \| P) = \underbrace{\boxed{-H_Q(X) - \sum_C E_Q \log \psi_C(x_C)}}_{F(P,Q)} + \log Z$$

Gibbs Free Energy

- $F(P,P) = -\log Z$ is a complicated function of true marginals and hard to compute

- Idea: construct a $F(P,Q)$ such that it has a nice functional form of beliefs (approximate marginals) and easy to optimize
  - approach 1: directly approximate with $\hat{F}(P,Q)$, e.g. Bethe approximation

$$F(P,Q) \approx \hat{F}(P,Q) = G_{\text{Bethe}}(\{q_i(x_i)\}, \{q_{ij}(x_i, x_j)\})$$

  - approach 2: restrict Q in a tractable class of distributions

Mean Field Methods

$$KL(Q \| P) = \underbrace{\boxed{-H_Q(X) - \sum_C E_Q \log \psi_C(x_C)}}_{F(P,Q)} + \log Z$$

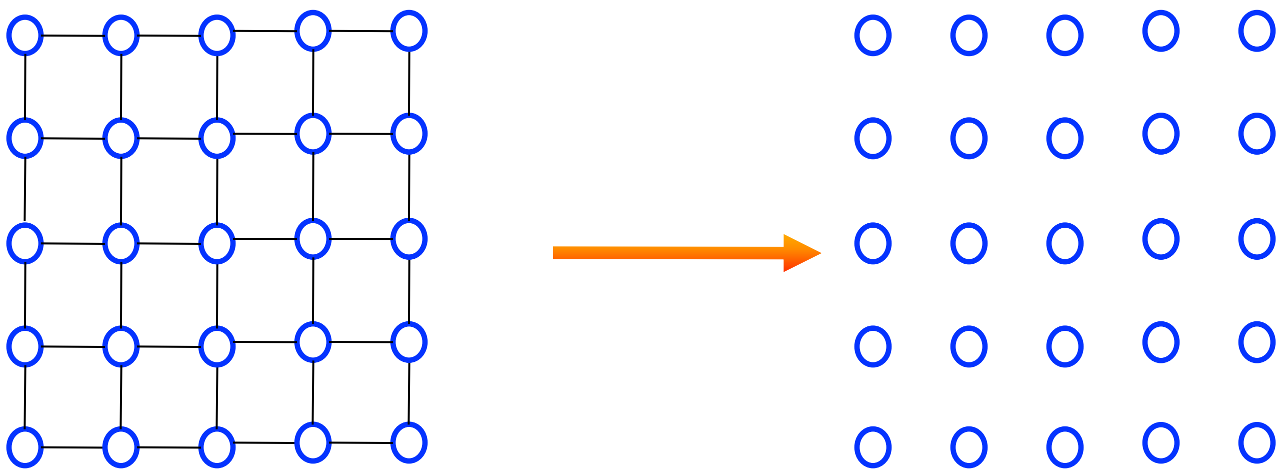
Gibbs Free Energy

- Restrict Q for which H_Q is feasible to compute
  - exact objective to minimize
  - tightened feasible set
  - yields a lower bound on the log partition function log Z
- Q is a "simple" *parameterized* approximating distribution
  - free parameters to tune are called variational parameters

Naïve Mean Field

- Completely factorized variational distribution

$$q(x) = \prod_{i \in V} q_i(x_i) \qquad H_Q = -\sum_i \sum_{x_i} q_i(x_i) \log q_i(x_i)$$



Naïve Mean Field Free Energy

- Consider a pairwise Markov random field

$$p(x) \propto \prod_{i \in V} \psi_i(x_i) \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j)$$

- Naïve mean field free energy

$$F(P,Q) = G_{MF}(q) = -\sum_{(i,j)\in E} \sum_{x_i, x_j} q_i(x_i) q_j(x_j) \log \psi_{ij}(x_i, x_j) - \sum_{i \in V} \sum_{x_i} q_i(x_i) \log \psi_i(x_i)$$

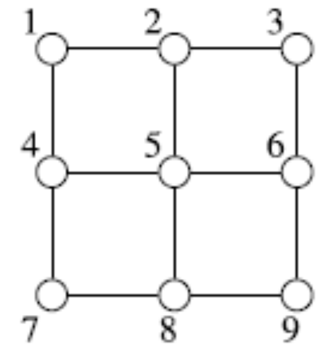
$$+ \sum_{i \in V} \sum_{x_i} q_i(x_i) \log q_i(x_i)$$

- Use coordinate descent to optimize with respect to q

Naïve Mean Field for Ising Model

- Ising model in {0,1} representation

$$p(x) \propto \exp\left\{ \sum_{i \in V} x_i \theta_i + \sum_{(i,j) \in E} x_i x_j \theta_{ij} \right\}$$

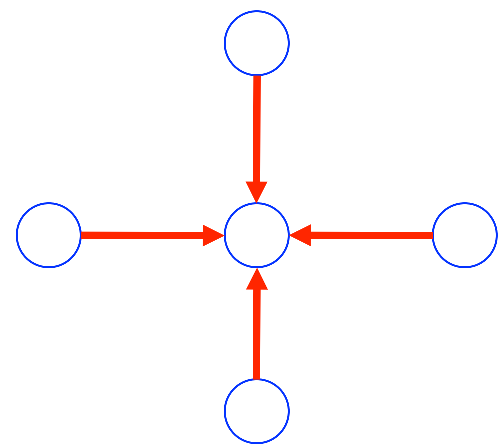


- The true marginals are the mean parameters

$$\mu_i = p(x_i = 1) = \mathbb{E}_p(x_i)$$

- The naïve mean field update equations

$$q_i \leftarrow \sigma\left(\theta_i + \sum_{j \in N(i)} \theta_{ij} q_j\right)$$



  - $q_i := q_i(x_i = 1) = \mathbb{E}_q[x_i]$ is the *variational mean parameter* at node i
  - the variational mean parameters are coupled among neighbors

Derivation

$$G_{\mathrm{MF}}(q) = -\sum_{(i,j)\in E} q_i q_j \theta_{ij} - \sum_{i\in V} q_i \theta_i$$
$$+ \sum_{i\in V} (q_i \log q_i + (1-q_i)\log(1-q_i))$$

$$\frac{dG_{\mathrm{MF}}(q)}{dq_i} = -\sum_{j\in N(i)} q_j \theta_{ij} - \theta_i + \log q_i - \log(1-q_i)$$

- Setting to zero gives us

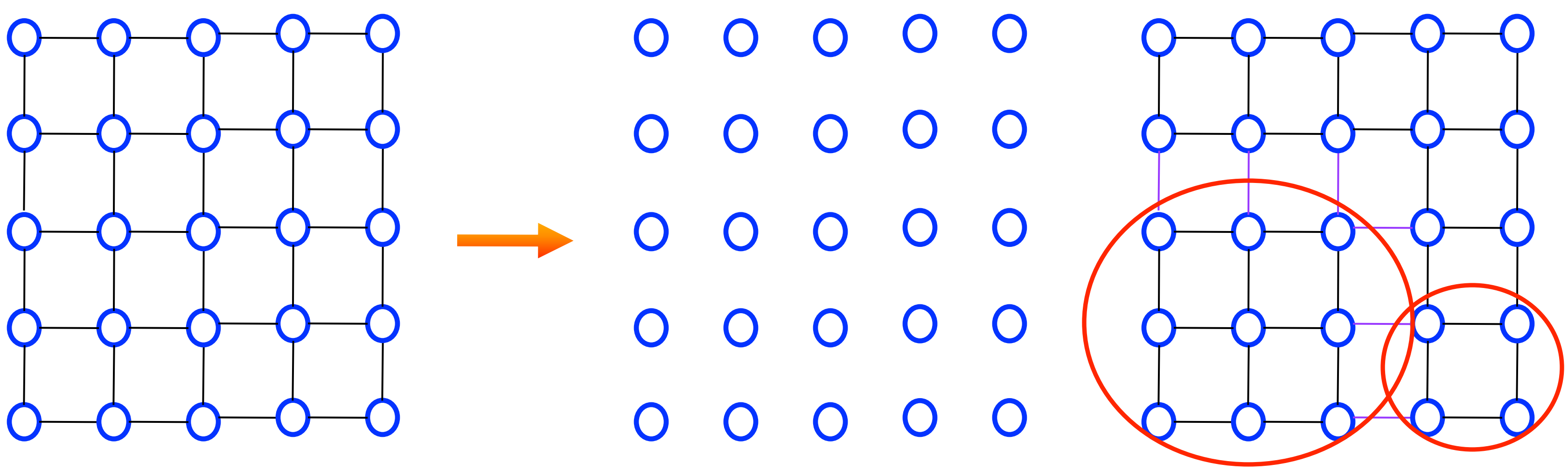
$$\theta_i + \sum_{j\in N(i)} q_j \theta_{ij} = \log \frac{q_i}{1-q_i}$$

- That is the mean field equation

$$q_i = \sigma\left(\theta_i + \sum_{j\in N(i)} q_j \theta_{ij}\right)$$

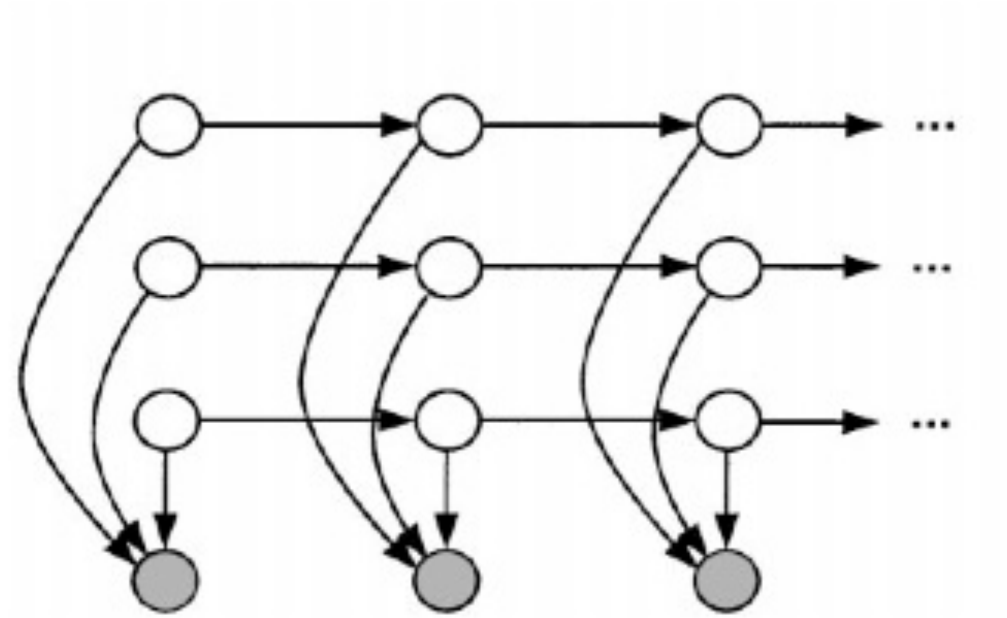
Structured Mean Field

- Mean field theory is general to any tractable sub-graphs
- Naïve mean field is based on the fully unconnected sub-graph
- Variants based on structured sub-graphs can be derived, such as trees, chains, and etc.



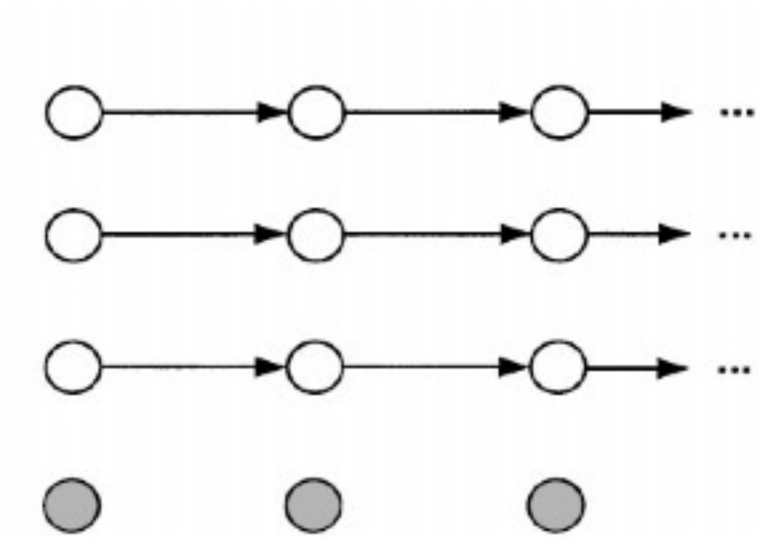
$$q(x) \propto \prod_{C \in \mathcal{C}} q_C(x_C)$$

Factorial HMM (Ghahramani & Jordan 97')



- Can be used to model multiple independent latent processes
- Exact inference is in general intractable (why?)
  - Complexity: $O(TMK^{M+1})$, which is exponential in the # of chains M

Structured Mean Field for Factorial HMM



- Structured mean field approximation (with variational parameter λ)

$$Q(\{S_t\} \mid \lambda) \propto \prod_{m=1}^{M} Q(S_1^{(m)} \mid \lambda) \prod_{t=2}^{T} Q(S_t^{(m)} \mid S_{t-1}^{(m)}, \lambda)$$

- The variational entropy term decouples into sum: one term for each chain
- In contrast to completely factorized Q, optimizing w.r.t. λ needs to run forward-backward algorithm as a subroutine

Summary so far

- Mean field methods minimizes KL divergence of variational distribution and target distribution by restricting the class of variational distributions
- It yields a lower bound of the log partition function, hence is a popular method to implement the approximate E-step of EM algorithm

Variational Principle

Inference Problems in Graphical Models

- Undirected graphical model (MRF):

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

- The quantities of interest:
  - marginal distributions: $p(x_i) = \sum_{x_j, j \neq i} p(x)$
  - normalization constant (partition function): Z
- Question: how to represent these quantities in a variational form?
  - Use tools from (1) exponential families; (2) convex analysis

Exponential Families

- Canonical parameterization (w.r.t measure ν)

$$p_\theta(x_1, \cdots, x_m) = \exp\left\{\theta^\top \phi(x) - A(\theta)\right\}$$

Canonical Parameters **Sufficient Statistics** **Log partition Function**

- Log normalization constant:

$$A(\theta) = \log \int \exp\{\theta^T \phi(x)\} dx$$

  - it is a convex function (Prop 3.1 in Wainwright & Jordan)
- Effective canonical parameters:

$$\Omega := \left\{\theta \in \mathbb{R}^d | A(\theta) < +\infty\right\}$$

- Regular family: Ω is an open set.

Examples:

Family	$\mathcal{X}$	$\boldsymbol{\nu}$	$\log p(\mathbf{x};\theta)$	$A(\theta)$
Bernoulli	$\{0,1\}$	Counting	$\theta x - A(\theta)$	$\log[1+\exp(\theta)]$
Gaussian	$\mathbb{R}$	Lebesgue	$\theta_1 x + \theta_2 x^2 - A(\theta)$	$\frac{1}{2}[\theta_1 + \log \frac{2\pi e}{-\theta_2}]$
Exponential	$(0,+\infty)$	Lebesgue	$\theta\,(-x) - A(\theta)$	$-\log\theta$
Poisson	$\{0,1,2\ldots\}$	Counting $h(x) = 1/x!$	$\theta x - A(\theta)$	$\exp(\theta)$

Graphical Models as Exponential Families

- Undirected graphical model (MRF):

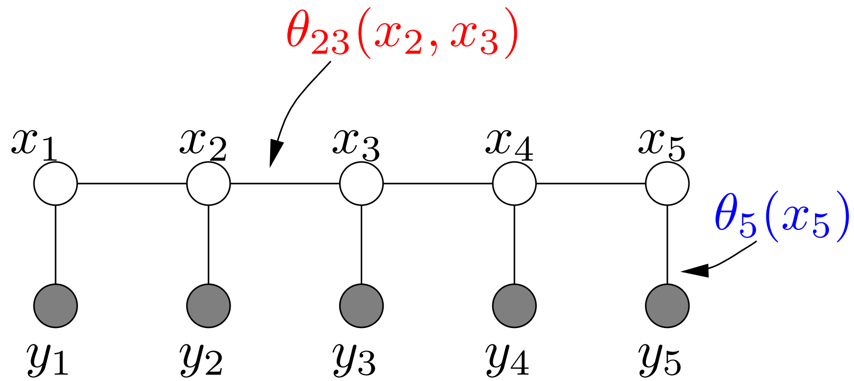
$$p(\mathbf{x};\theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi(\mathbf{x}_C;\theta_C)$$

- MRF in an exponential form:

$$p(\mathbf{x};\theta) = \exp\left\{ \sum_{C \in \mathcal{C}} \log \psi(\mathbf{x}_C;\theta_C) - \log Z(\theta) \right\}$$

  - $\log \psi(\mathbf{x}_C;\theta_C)$ can be written in a *linear* form after some reparameterization
  - Sufficient statistics must respect the structure of graph

Example: Hidden Markov Model



- What are the sufficient statistics?

$$\mathbb{I}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{otherwise,} \end{cases} \qquad \mathbb{I}_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \quad \text{and} \quad x_t = k, \\ 0 & \text{otherwise,} \end{cases}$$

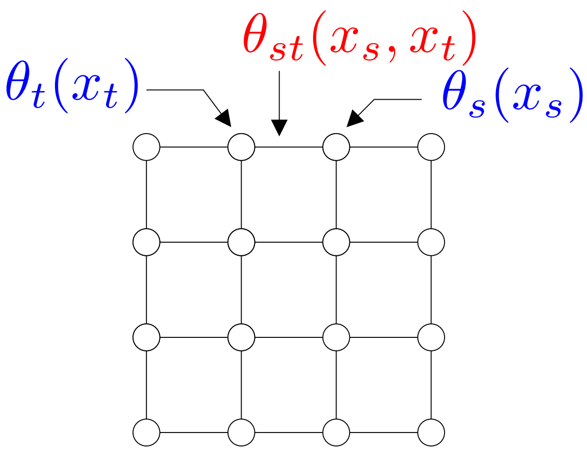
- What are the corresponding canonical parameters?

$$\theta_{st;jk} = \log P(x_t = k \mid x_s = j) \qquad \theta_{s;j} = \log P(y_s \mid x_s = j)$$

- A compact form

$$\theta_{st}(x_s, x_t) = \sum_{jk} \theta_{st;jk} \mathbb{I}_{st;jk}(x_s, x_t) = \log P(x_t \mid x_s)$$

Example: Discrete MRF



Indicators: $\mathbb{I}_j(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{otherwise} \end{cases}$

Parameters:
$$\theta_s = \{\theta_{s;j}, j \in \mathcal{X}_s\}$$
$$\theta_{st} = \{\theta_{st;jk}, (j,k) \in \mathcal{X}_s \times \mathcal{X}_t\}$$

Compact form:
$$\theta_s(x_s) := \sum_j \theta_{s;j} \mathbb{I}_j(x_s)$$
$$\theta_{st}(x_s, x_t) := \sum_{j,k} \theta_{st;jk} \mathbb{I}_j(x_s) \mathbb{I}_k(x_t)$$

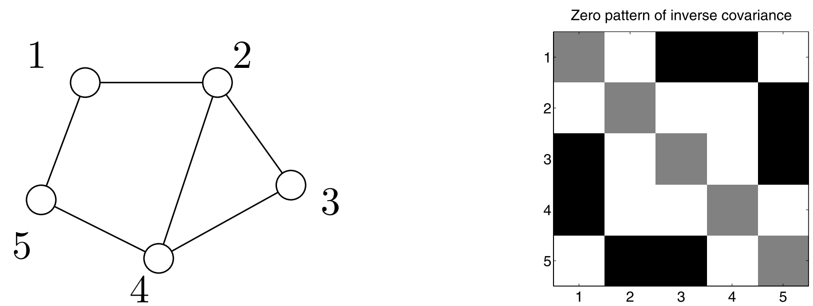
- In exponential form

$$p(\mathbf{x};\theta) \quad \propto \quad \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}$$

- Why is this representation is useful? How is it related to inference problem?
  - Computing the expectation of sufficient statistics (mean parameters) given the canonical parameters yields the marginals

Example: Gaussian MRF

- Consider a zero-mean multivariate Gaussian distribution that respects the Markov property of a graph
  - Hammersley-Clifford theorem states that the precision matrix $\Lambda = \Sigma^{-1}$ also respects the graph structure



(a) (b)

- Gaussian MRF in exponential form

$$p(\mathbf{x}) = \exp\left\{\frac{1}{2}\langle\Theta, \mathbf{x}\mathbf{x}^T\rangle - A(\Theta)\right\}, \text{where } \Theta = -\Lambda$$

  - Sufficient statistics are $\{x_s^2, s \in V; x_s x_t, (s,t) \in E\}$

Computing Mean Parameter: Bernoulli

- A single Bernoulli random variable

$$p(x;\theta) = \exp\{\theta x - A(\theta)\}, x \in \{0,1\}, A(\theta) = \log(1 + e^{\theta})$$

- Computing its mean parameter from canonical parameter:

$$\mu = p(x=1) = \mathbb{E}[x] = \frac{e^{\theta}}{1+e^{\theta}}$$

- Want to do it in a variational manner: cast the procedure of computing mean in an optimization-based formulation

Conjugate Dual Function

- Given any function $f(\theta)$, its conjugate dual function is:

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

- Conjugate dual is always a convex function: pointwise supremum of a class of linear functions

- Under some technical condition on f (convex and lower semi-continuous), the dual of dual is itself:

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

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

- See Convex Optimization book by Boyd for more details

Computing Mean Parameter: Bernoulli

- Compute the conjugate

$$A^*(\mu) \;:=\; \sup_{\theta \in \mathbb{R}} \left\{ \mu\theta - \log[1 + \exp(\theta)] \right\}$$

- Stationary condition $\mu = \dfrac{e^\theta}{1 + e^\theta}$

- We find $A^*(\mu) = \begin{cases} \mu \log \mu + (1-\mu)\log(1-\mu) & \text{if } \mu \in [0,1] \\ +\infty & \text{otherwise.} \end{cases}.$

- The variational form to compute mean:

$$A(\theta) = \max_{\mu \in [0,1]} \left\{ \mu \cdot \theta - A^*(\mu) \right\}.$$

- The optimum is achieved at $\mu = \dfrac{e^\theta}{1 + e^\theta}$

Next Step …

- The last identity is not a coincidence but a deep theorem in general exponential family
- However, for general graph models/exponential families, computing the conjugate dual (negative entropy) is intractable
- Moreover, the constrain set of mean parameter is hard to characterize
- Relaxing/Approximating them leads to different algorithms: loop belief propagation, naïve mean field, and etc.