An Introduction to Variational Methods for Graphical Models

By

Jordan, M., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.

Basics of Variational Methodology

- Exact inference in tree model can be done efficiently[1]
  - Message-passing algorithm
  - Junction-Tree algorithm
  - In general GM, exact inference is intractable
- We want to approximate the exact inference.
- Variational Approximation is a general method to approximate a complex function (e.g: ln(x)) by a family of simpler functions (e.g : linear).

See M.I Jordan, Graphical Models, in Statistical Science, 2004

Ideology of Variational Methods

- Application of variational methods converts a complex problem into a simpler problem
- The simpler problem is generally characterized by a decoupling of the degrees of freedom in the original problem.
- This decoupling is achieved via an expansion of the problem to include additional parameters, known as variational parameters, that must be fit to the problem at hand.
- This paradigm would be explained in detail with the help of 2 examples – QMR-DT and Boltzmann Machine.

A Simple Example

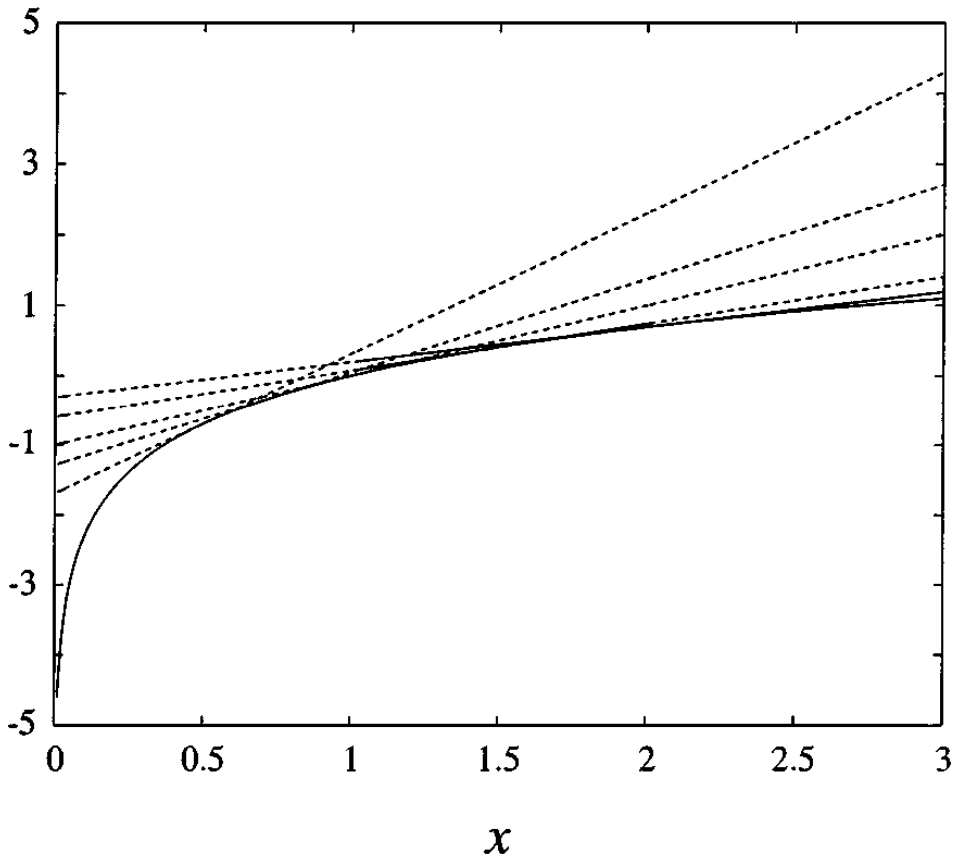
- Consider the logarithm function expressed variationally:

$$\ln(x) = \min_{\lambda}\{\lambda x - \ln\lambda - 1\}.$$

- Here λ is the variational parameter.
- Logarithm is a concave function.
- Each line above has slope λ and intercept (-ln λ-1)

Simple Example (Cont.)

- If we range across λ, the family of such lines forms an upper envelope of the logarithm function.
- Justification: We have converted a non-linear function into a linear function
- Cost: We have introduced a free parameter λ which must be set for each value of x.



Another Example

- Consider the logistic regression model:

$$f(x) = \frac{1}{1 + e^{-x}}$$

- This function is neither convex nor concave.
- So, a simple linear bound will not work.

Log Logistic function

- Consider the log logistic function:

$$g(x) = -\ln(1 + e^{-x})$$

- This function is concave. Thus, it can be bounded with linear functions.

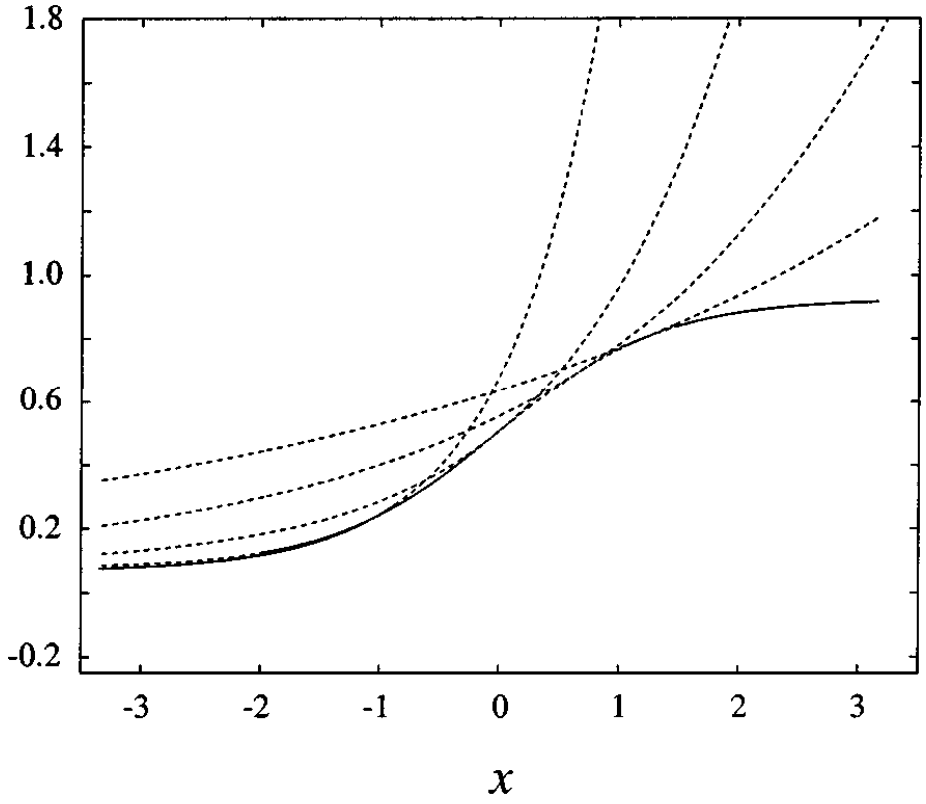
$$g(x) = \min_{\lambda}\{\lambda x - H(\lambda)\}$$

- Here $H(\lambda) = -\lambda \ln \lambda - (1-\lambda)\ln(1-\lambda)$
- Now taking the exponential on both sides, we get:

$$f(x) = \min_{\lambda}\left[e^{\lambda x - H(\lambda)}\right]$$

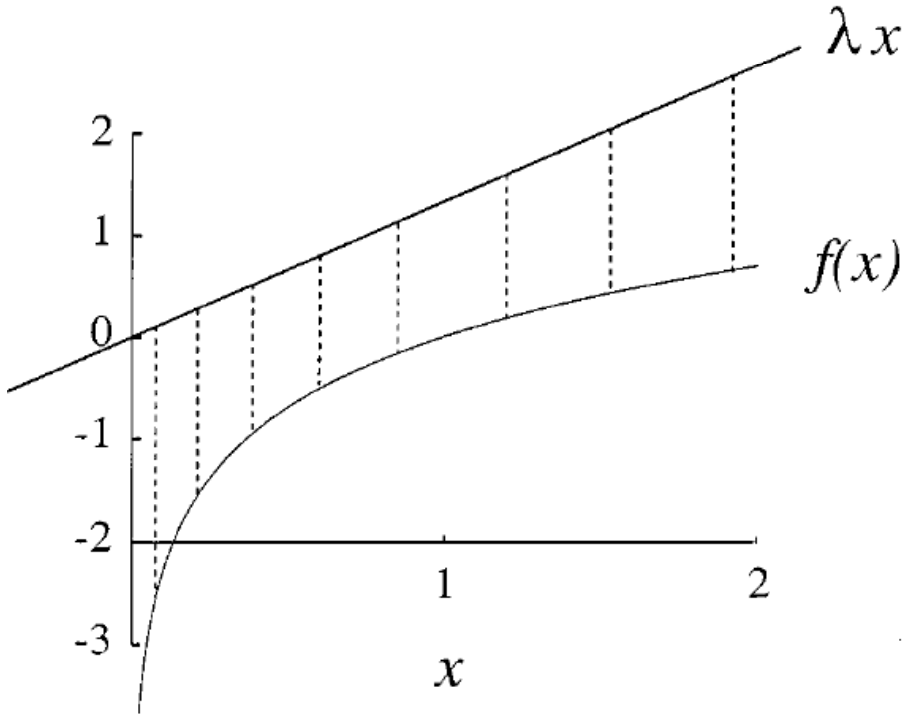
Upper bound of Logistic function

- For any value of λ, *we obtain an upper bound of the* logistic function for all values of *x.*
- Advantage: It is easier to compute joint probability when expressed variationally (Note that the exponentials are linear in x).



Convex Duality

- A principle way to estimate a convex/concave function by a family of linear functions.



Convex Duality

- A more general treatment of variational bounds.
- Any concave function f(x) can be represented via a conjugate or dual function as follows:

$$f(x) = \min_{\lambda} \{\lambda^T x - f^*(\lambda)\}$$

- Here x and λ are allowed to be vectors. The conjugate function can be obtained from the dual expression:

$$f^*(\lambda) = \min_{x} \{\lambda^T x - f(x)\}$$

Convex Duality

- For convex f(x), we get:

$$f(x) = \max_{\lambda}\{\lambda^T x - f^*(\lambda)\}$$

where

$$f^*(\lambda) = \max_{x}\{\lambda^T x - f(x)\}$$

Convex Duality - Non-linear case

- Convex Duality is not restricted to linear bounds.
- If f(x) is concave in x^2, we can write:

$$f(x) = \min_{\lambda}\{\lambda x^2 - \bar{f}^*(\lambda)\}$$

where $\bar{f}^*(\lambda)$ is the conjugate function of $\bar{f}(x) \equiv f(x^2)$

- Thus, the transformation yields a quadratic bound on f(x).

Summary

- The general methodology suggested by convex duality is the following.
- We wish to obtain upper or lower bounds on a function of interest.
- If the function is already convex or concave then we simply calculate the conjugate function.
- If the function is not convex or concave, then we look for an invertible transformation that renders the function convex or concave.
- We may also consider transformations of the argument of the function. We then calculate the conjugate function in the transformed space and transform back.

Joint and Conditional Probability

- So far, we discussed the local probability distributions at the nodes of a graphical model.
- How do these approximations translate into approximations for the global probabilities of interest:
  - Conditional distribution *P(H|E) that is our interest in the inference problem and*
  - *Marginal probability P(E)* that is our interest in learning problems?

Joint and Conditional Probabilities

- Suppose we have a lower bound and an upper bound for each of the local conditional pr $P(S_i \mid S_{\pi(i)})$
- Thus, we have:

$$P^U(S_i \mid S_{\pi(i)}, \lambda_i^U) \text{ and } P^L(S_i \mid S_{\pi(i)}, \lambda_i^L)$$

where λ_i^U and λ_i^L are variational parameterizations for the upper and lower bounds.

Joint and Conditional Probabilities

- Considering upper bounds, we get:

$$P(S) = \prod_i P\left(S_i \mid S_{\pi(i)}\right) \leq \prod_i P^U\left(S_i \mid S_{\pi(i)}, \lambda_i^U\right)$$

- For the marginal probability, we get:

$$P(E) = \sum_{\{H\}} P(H, E) \leq \sum_{\{H\}} \prod_i P^U\left(S_i \mid S_{\pi(i)}, \lambda_i^U\right)$$

- Key step - Variational forms should be chosen to carry out summation over H efficiently.
- To get the optimum value, the right hand side of above equation has to be minimized wrt λ_i^U

Important Distinction

- If we allow the variational parameters to be set optimally for each value of the argument *S,* then it is possible (in principle) to find optimizing settings of the variational parameters that recover the exact value of the joint probability.
- On the other hand, we are *not generally able to recover* exact values of the marginal by optimizing over variational parameters that depend only on the argument *E.*

Important Distinction(2)

- Consider, for example, the case of a node *that has parents in H.*
- As we range across {*H} there will be summands that will* involve evaluating the local probability $P(S_i \mid S_{\pi(i)})$ for different values of parents.

• If the variational parameter *depends only on E, we cannot in general expect to obtain* an exact representation for above probability *in each summand*

Loose and Tight bounds

- In particular, if $P(S_i \mid S_{\pi(i)})$ is nearly constant as we range across parents, then the bounds may be expected to be tight.
- Otherwise, one might expect that the bound would be loose.

Conditional Probability

$$P(H \mid E) = P(H, E)/P(E)$$

- To obtain upper and lower bounds on the conditional distribution, we must have upper and lower bounds on both the numerator and the denominator.
- Generally speaking, it is sufficient to obtain the lower and upper bounds on the denominator as the numerator involves fewer sums.
- If S = H U E, the numerator is simply a function evaluation.

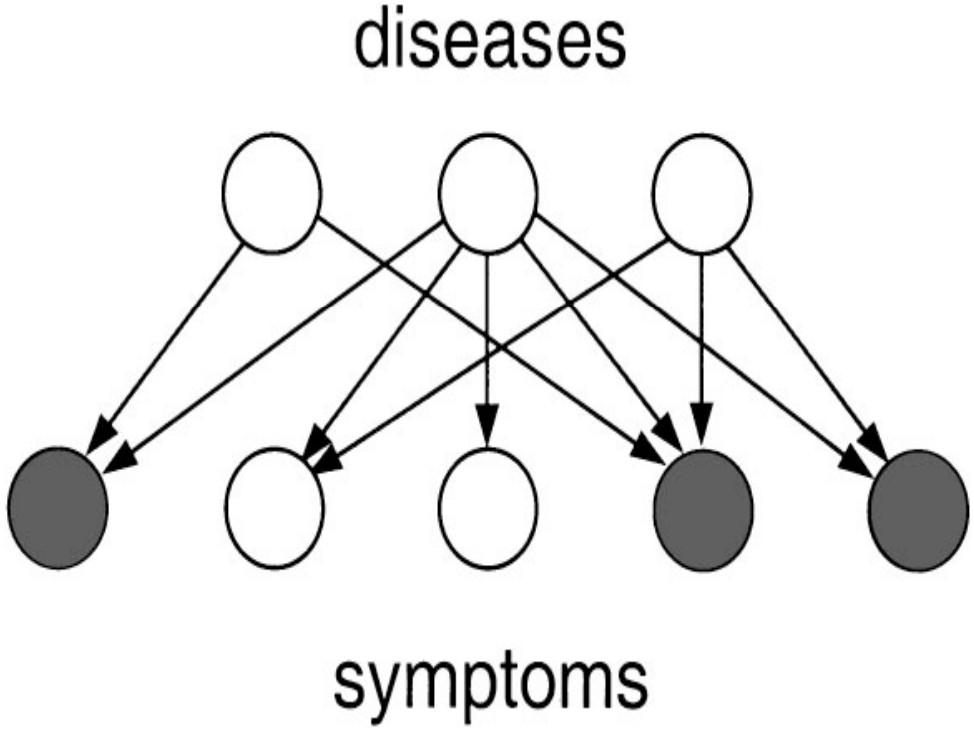
QMR-DT Database

QMR-DT database

- Example of graphical model – QMR-DT database
- Exact inference is infeasible
- QMR-DT database is a diagnostic system which uses fixed graphical model to answer queries.

QMR-DT database

- QMR-DT database is a bipartite graphical model
- Upper layer of nodes represent diseases and
- Lower layer of nodes represent symptoms
- Approximately 600 disease nodes and 4000 symptom nodes



Joint Probability in QMR-DT

- Evidence is a set of observed symptoms
- Represent the vector of findings (symptoms) with symbol f
- The symbol *d denotes* the vector of diseases
- All nodes are binary, thus the components *fi and di are binary random* variables
- The joint probability is given by:

$$P(f,d) = P(f \mid d)P(d)$$

$$= \left[\prod_i P(f_i \mid d)\right]\left[\prod_j P(d_j)\right].$$

Conditional Prob. in QMR-DT

- The conditional probabilities of the findings given the diseases, *P(fi|d), were obtained from* expert assessments under a "noisy-OR" model.

$$P(f_i = 0 \mid d) = (1 - q_{i0}) \prod_{j \in \pi(i)} (1 - q_{ij})^{d_j}$$

$$= \exp\left\{ - \sum_{j \in \pi(i)} \theta_{ij} d_j - \theta_{i0} \right\}$$

$$\text{where } \theta_{ij} \equiv -\ln(1 - q_{ij})$$

Conditional Prob. in QMR-DT

- The probability of a positive finding is given as follows:

$$P(f_i = 1 \mid d) = 1 - \exp\left\{ - \sum_{j \in \pi(i)} \theta_{ij} d_j - \theta_{i0} \right\}$$

- Products of the probabilities of positive findings yield cross products terms that are problematic for exact inference.
- Diagnostic calculation under the QMR-DT model is generally infeasible

Variational Approx. for QMR-DT

- "*findings* nodes" corresponding to symptoms that are not observed are omitted and have no impact on inference.
- Effects of negative findings on the disease probabilities can be handled in linear time because of the exponential form of the probability.

$$P(f_i = 0 \mid d) = (1 - q_{i0}) \prod_{j \in \pi(i)} (1 - q_{ij})^{d_j}$$

$$= \exp\left\{ - \sum_{j \in \pi(i)} \theta_{ij} d_j - \theta_{i0} \right\}$$

$$\text{where } \theta_{ij} \equiv -\ln(1 - q_{ij})$$

Variational Approx. for QMR-DT

- We focus on performing inference when there are positive findings.

$$P(f_i = 1 \mid d) = 1 - \exp\left\{ - \sum_{j \in \pi(i)} \theta_{ij} d_j - \theta_{i0} \right\}$$

- Function 1-exp(-x) is log concave. So, we can use variational approximation.

Calculating Upper bound

- The following variational approximation can be derived for the upper bound.

$$1 - e^{-x} \leq e^{\lambda x - f^*(\lambda)},$$

where the conjugate function is as follows:

$$f^*(\lambda) = -\lambda \ln \lambda + (\lambda + 1) \ln(\lambda + 1).$$

Node Decoupling using Variational Approx.

- Using the above variational approx., we get:

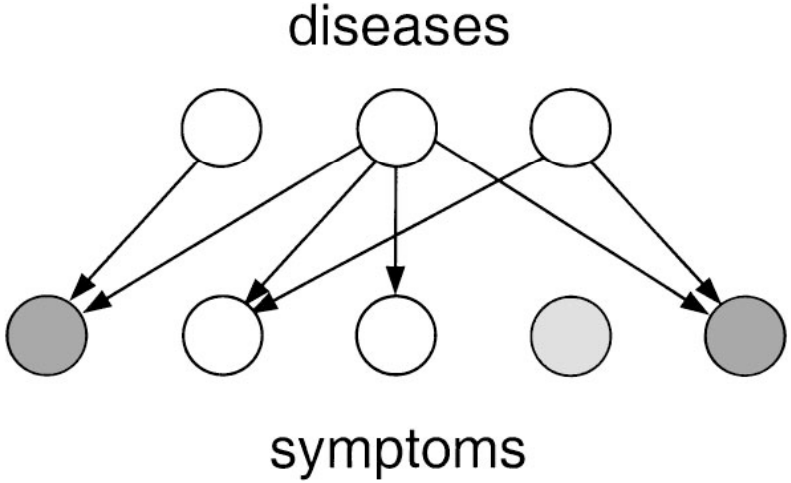
$$P(f_i = 1 \mid d) \leq \exp\left\{\lambda_i \left(\sum_{j \in \pi(i)} \theta_{ij} d_j + \theta_{i0}\right) - f^*(\lambda_i)\right\}$$

$$= e^{\lambda_i \theta_{i0} - f^*(\lambda_i)} \prod_{j \in \pi(i)} [e^{\lambda_i \theta_{ij}}]^{d_j}.$$

- In the original noisy-OR model, multiplication led to coupling of dj and dk nodes.
- But in the above expression, the contributions associated with the dj and dk nodes are uncoupled.

Node Decoupling shown graphically

- Thus the graphical effect of the variational transformation is to delink the *ith finding from the graph.*



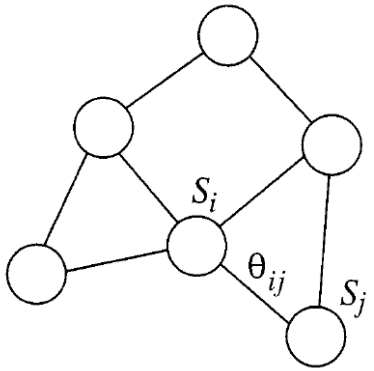
- This process of variational transformation is applied iteratively till the graph is simple enough that we can use exact inference on it.

Summary so far

- You've seen how QMR-DT, a graphical model, is "transformed" to another model so that it can be computed efficiently.
- The transformation relies on two insights:
  - Node coupling is the cause of intractability (e.g: complete independence is the easiest to deal with → no edge)
  - Convex duality theorem gives us a principled way to estimate a complex function f(x) by a family of simpler functions (linear, quadratic…), parameterized by λ.
- The transformation is carried one node at a time, until the graph is simple enough for exact inference

Boltzmann Machines

- Is a type of undirected graphical model, where we define potential function for every 2-node cliques.



- The joint distribution has the following form

$$P(S) = \frac{\exp\left\{\sum_{i<j}\theta_{ij}S_iS_j + \sum_i \theta_{i0}S_i\right\}}{Z}$$

Z is a normalizing faction

Decoupling a node S_i

- We want to “decouple” Si from the rest of the graph
- The marginal can be re-write as following

$$Z = \sum_{\{S\}} \exp\left\{ \sum_{j<k} \theta_{jk} S_j S_k + \sum_j \theta_{j0} S_j \right\}$$

$$= \sum_{\{S \setminus S_i\}} \sum_{S_i \in \{0,1\}} \exp\left\{ \sum_{j<k} \theta_{jk} S_j S_k + \sum_j \theta_{j0} S_j \right\}$$

Variational transformation

- The function is log-convex, thus, we bound it similarly to QMR-DT example as following

$$\ln\left[\sum_{S_i\in\{0,1\}}\exp\left\{\sum_{j<k}\theta_{jk}S_jS_k+\sum_j\theta_{j0}S_j\right\}\right]$$

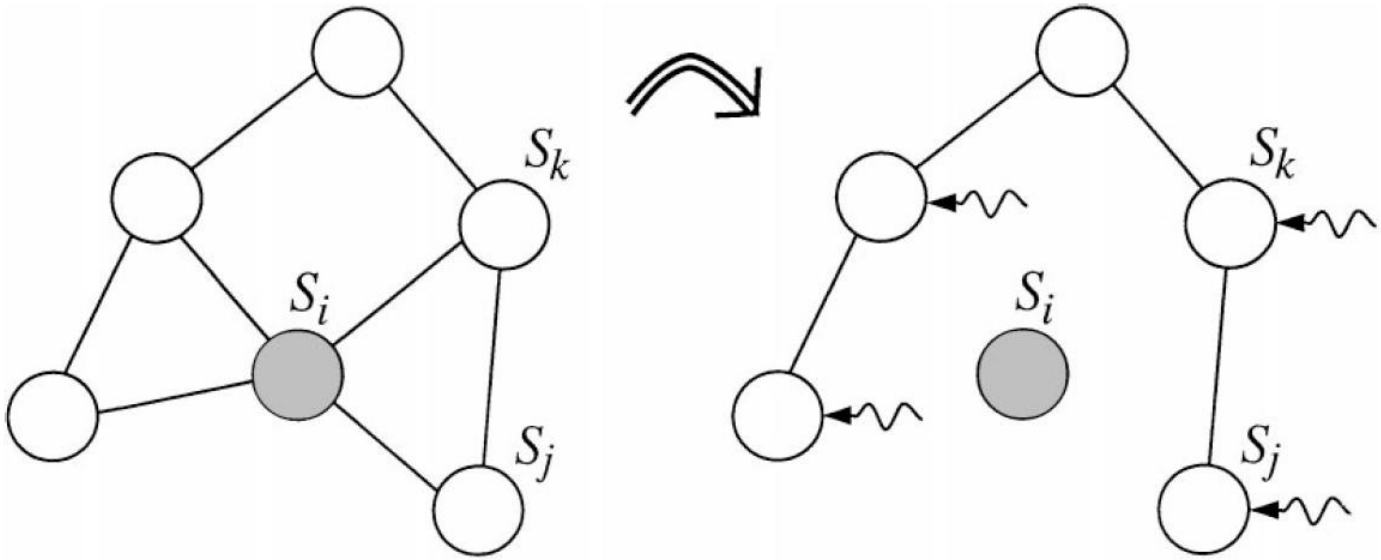
$$=\sum_{\{j<k\}\neq i}\theta_{jk}S_jS_k+\sum_{j\neq i}\theta_{j0}S_j+\ln\left[\sum_{S_i\in\{0,1\}}\exp\left\{\sum_{j\neq i}\theta_{ij}S_iS_j+\theta_{i0}S_i\right\}\right]$$

$$=\sum_{\{j<k\}\neq i}\theta_{jk}S_jS_k+\sum_{j\neq i}\theta_{j0}S_j+\ln\left[1+\exp\left\{\sum_{j\neq i}\theta_{ij}S_j+\theta_{i0}\right\}\right]$$

$$\geq\sum_{\{j<k\}\neq i}\theta_{jk}S_jS_k+\sum_{j\neq i}\theta_{j0}S_j+\lambda_i^L\left(\sum_{j\neq i}\theta_{ij}S_j+\theta_{i0}\right)+H(\lambda_i^L),$$

Graphical effect

- The effect of the approximation.
  - Si is now “independent”
  - Extra constants are introduced to its neighbors



Sequential VS Block approach

- In the above method, we “decouple” one node at a time, until the model behaves “nicely”. This is sequential approach.
- We can also transform a block of nodes at a time. This is block approach.

Block approach

- Suppose we need to approximate P(H|E), we introduce an approximating family function Q(H|E,λ) and choose the variational parameters λ such that

$$\lambda^* = \arg\min_{\lambda} D(Q(H \mid E, \lambda) \parallel P(H \mid E)),$$

- This yields the best lower bound for the log likelihood function P(E).
  - Proof on the next slide

Block approach

- Using Jensen's inequality.

$$\ln P(E) = \ln \sum_{\{H\}} P(H, E)$$

$$= \ln \sum_{\{H\}} Q(H \mid E) \cdot \frac{P(H, E)}{Q(H \mid E)}$$

$$\geq \sum_{\{H\}} Q(H \mid E) \ln \left[\frac{P(H, E)}{Q(H \mid E)} \right].$$

- Difference between the left and the right hand side is D(Q(H|E) || P(H|E)). This can also be justified by the convex duality theorem.

$$f^*(Q) = \min \left\{ \sum_{\{H\}} Q(H \mid E, \lambda) \ln P(H, E) - \ln P(E) \right\}$$

Conclusion

- Variational method offer an alternative to Sampling for approximate inference.
- The exact inference on graph is intractable in general, due to node coupling.
- Variational method transform a complex function to a family of simpler ones, giving the “best lower/upper bound” to the original function.
- The convex duality theorem gives a principle way to get the bound. There are other methods too.
- In general, effective use of variational method is kind of an art, requiring a lot of creativity (e.g: what node to transform, what order, what sub-structure….).