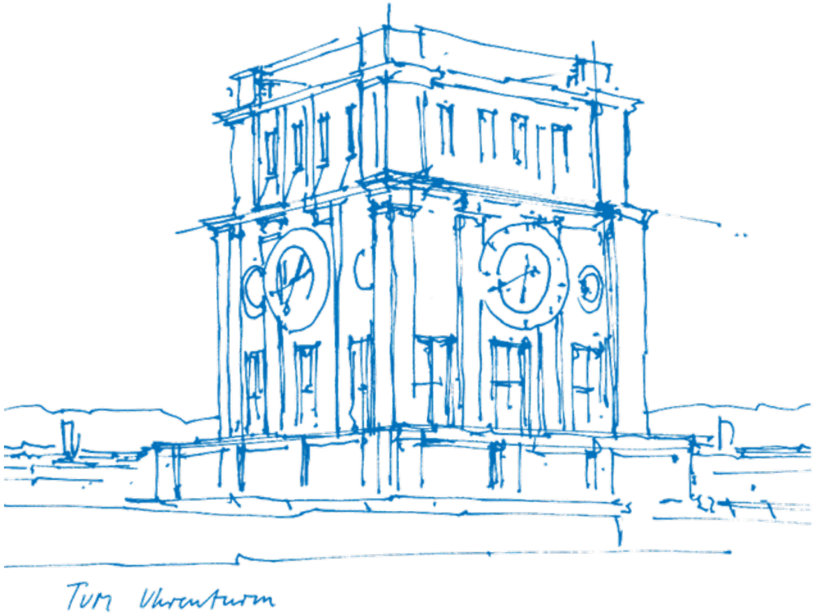
Computer Vision & Artificial Intelligence
Department of Informatics
Technical University of Munich

III : Inference on Graphical Models

Tao Wu, Yuesong Shen, Zhenzhang Ye

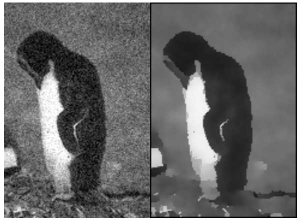
Computer Vision & Artificial Intelligence
Technical University of Munich



Motivation

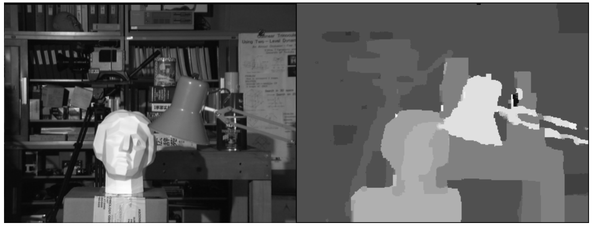
– Many computer vision tasks boil down to inference on graphical models.

Denoising



Optical flow

Stereo matching



Inpainting

Super-resolution

1. **Probabilistic inference**: compute marginal distribution
$$p(y) = \sum_x p(y, x).$$

2. **MAP inference**: compute maximum of conditional distribution
$$\arg\max_y p(y|x).$$

Exact Inference

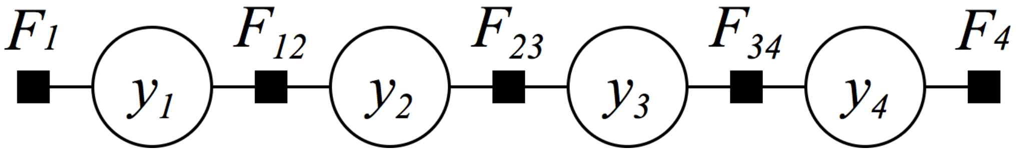
Outline of the Section

- Basic idea: Variable elimination.
- Junction tree algorithm on arbitrary MRFs.
- Belief propagation on tree factor graphs.

Example: Marginal Query on a "Chain" MRF

Joint distribution represented by MRF:

$$p(y_1, y_2, y_3, y_4) = \frac{1}{Z}\phi_1(y_1) \cdot \phi_{12}(y_1, y_2) \cdot \phi_{23}(y_2, y_3) \cdot \phi_{34}(y_3, y_4) \cdot \phi_4(y_4).$$



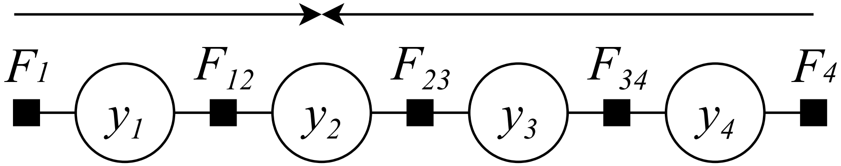
Query about marginal distribution $p(y_2) = ?$

Variable Elimination

Apply **variable elimination** (VE) to the marginal query:

$$
\begin{aligned}
p(y_2) &= \sum_{y_1,y_3,y_4} p(y_1, y_2, y_3, y_4) \\
&= \sum_{y_1,y_3,y_4} \frac{1}{Z}\phi_1(y_1)\phi_{12}(y_1, y_2)\phi_{23}(y_2, y_3)\phi_{34}(y_3, y_4)\phi_4(y_4) \\
&= \frac{1}{Z}\underbrace{\sum_{y_1}\Big(\phi_1(y_1)\phi_{12}(y_1, y_2)\Big)}_{=:\ m_{1\to2}(y_2)}\sum_{y_3}\Big(\phi_{23}(y_2, y_3)\underbrace{\sum_{y_4}\big(\phi_{34}(y_3, y_4)\phi_4(y_4)\big)}_{=:\ m_{4\to3}(y_3)}\Big) \\
&= \frac{1}{Z}m_{1\to2}(y_2)\underbrace{\sum_{y_3}\Big(\phi_{23}(y_2, y_3)m_{4\to3}(y_3)\Big)}_{=:\ m_{3\to2}(y_2)} \\
&= \frac{1}{Z}m_{1\to2}(y_2)m_{3\to2}(y_2), \\
Z &= \sum_{y_2} m_{1\to2}(y_2)m_{3\to2}(y_2).
\end{aligned}
$$

Variable Elimination and Beyond



- This algorithm is called **sum-product** VE.
- Sum-product VE yields *exact* inference (of one node marginal) on any *tree-structured factor graph*.
- Observed nodes (a.k.a. *evidence*) can be introduced as reduced factors.
- A similar algorithm can be derived for MAP inference – simply switch all "sum" to "max". The resulting algorithm is called **max-product** VE.
- We shall consider two different extensions beyond VE:

1. Inference on arbitrary MRFs? ⇝ **Junction tree algorithm**.
2. Compute all node/factor marginals at one shot? ⇝ **Belief propagation**.

Junction Tree

- For an undirected graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, the **junction tree** of $\mathcal{H}$ is a tree $\mathcal{T}$ s.t.
  1. The nodes of $\mathcal{T}$ consist of the *maximal cliques* of $\mathcal{H}$.
  2. The edge S_{ij} between two nodes C_i, C_j of $\mathcal{T}$ (i.e. two maximal cliques of $\mathcal{H}$) is given by $S_{ij} = C_i \cap C_j$ (known as the *running intersection property*).
- $\mathcal{H}$ is **triangulated** if every cycle of length ≥ 4 has a *chord*. (A chord is an edge that is not part of the cycle but connects two vertices of the cycle.)
- Theorem [Lauritzen '96]: A graph has a junction tree iff it is triangulated.

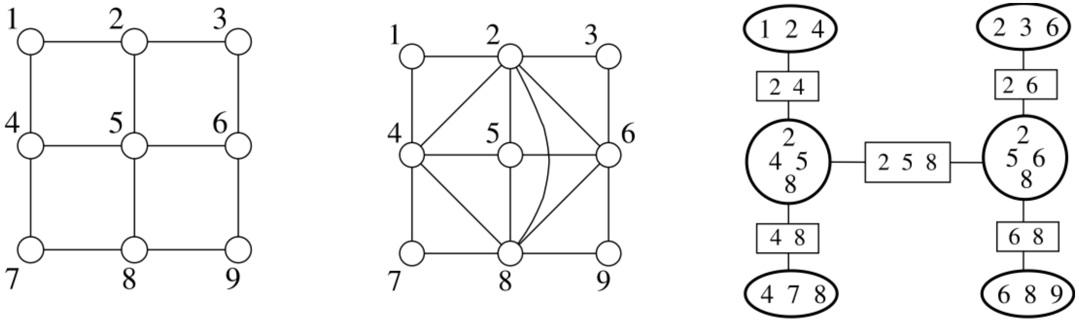
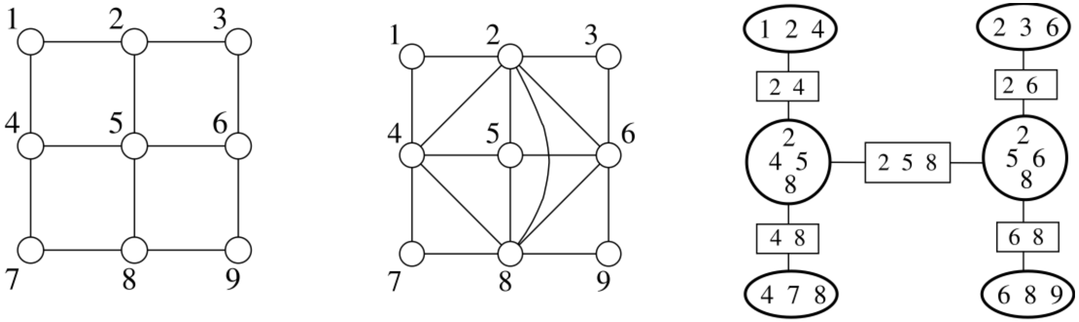


Figure:[1] (a) Original graph; (b) Triangulation of (a); (c) Junction tree for (b).

[1] Wainwright and Jordan, "Graphical Models, Exponential Families, and Variational Inference".

Junction Tree Algorithm (Sketch)



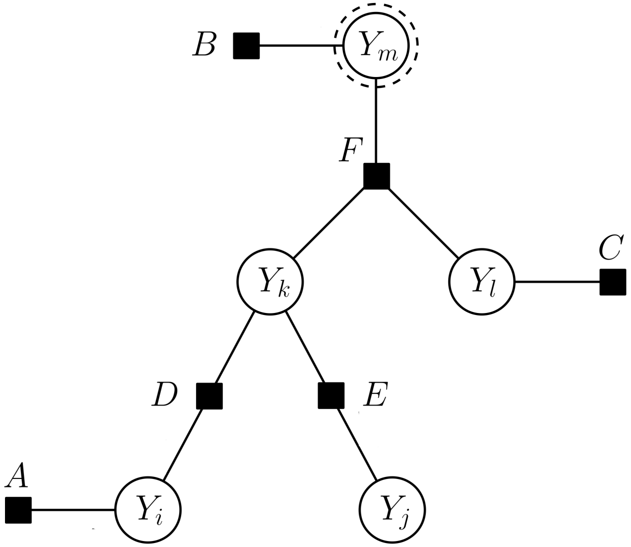
Sum-product message passing on a junction tree $\mathcal{T}$ appears like:

$$m_{C_i \to C_j}(y_{C_j \cap C_i}) = \sum_{y_{C_i \setminus C_j}} \phi_{C_i}(y_{C_i}) \prod_{C_k \in \mathrm{nbr}_{\mathcal{T}}(C_i) \setminus \{C_j\}} m_{C_k \to C_i}(y_{C_i \cap C_k}).$$

Overall **junction tree algorithm** for exact inference on an arbitrary MRF:

1. Given a MRF with cycles, triangulate it by adding edges as necessary.
2. Form a junction tree $\mathcal{T}$ for the triangulated MRF.
3. Run VE on the junction tree $\mathcal{T}$.

Belief Propagation on Tree Factor Graphs[2]



- Factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$: assumed to be a tree.
- Neighbors of a variable or factor node:

$$\begin{aligned} \mathrm{nbr}_{\mathcal{G}}(i) &= \{F \in \mathcal{F} : (i, F) \in \mathcal{E}\}, \\ \mathrm{nbr}_{\mathcal{G}}(F) &= \{i \in \mathcal{V} : (i, F) \in \mathcal{E}\}. \end{aligned}$$

- (Log-domain) energies: $E_F(y_F) = -\log \phi_F(y_F)$.

[2] Illustrations for BP are extracted from Nowozin & Lampert, 2011.

BP: Leaf-to-Root Stage

0. Pick $Y_r \in \mathcal{V}$ as the tree root (e.g. Y_m in the figure).

1a. Schedule the leaf-to-root messages.

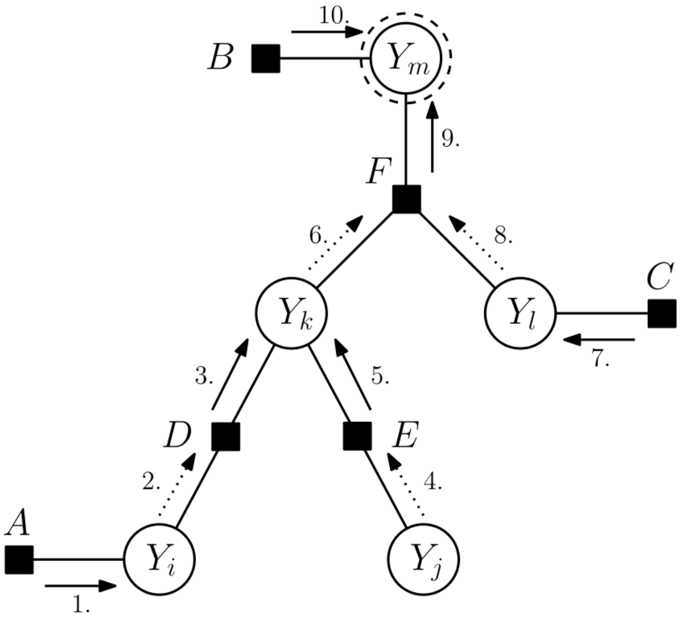


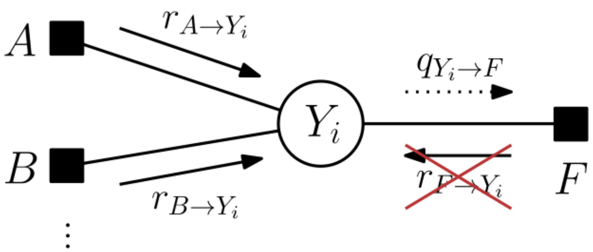
Figure: Belief propagation: leaf-to-root stage.

1b. Compute all leaf-to-root messages (detailed in the next slide).

BP: Compute Messages

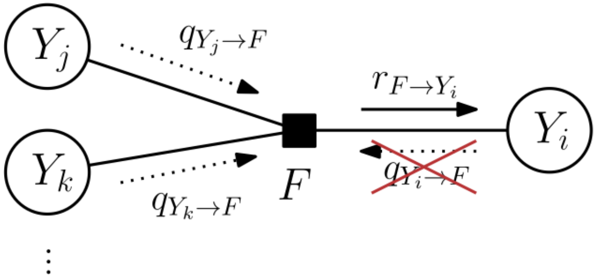
- Compute variable-to-factor message:

$$q_{i\to F}(y_i) = \sum_{F'\in \mathrm{nbr}_{\mathcal{G}}(i)\setminus\{F\}} r_{F'\to i}(y_i).$$



- Compute factor-to-variable message:

$$r_{F\to i}(y_i) = \log \sum_{y_{F\setminus\{i\}}} \exp\Big(-E_F(y_F) + \sum_{i'\in \mathrm{nbr}_{\mathcal{G}}(F)\setminus\{i\}} q_{i'\to F}(y_{i'})\Big).$$



BP: Compute the Partition Function

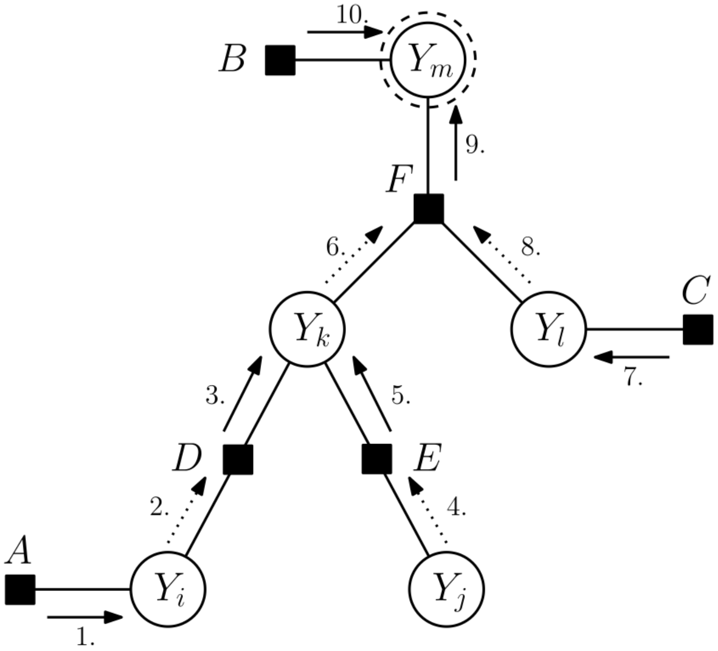


Figure: Belief propagation: leaf-to-root stage.

1c. Compute the log partition function:

$$\log Z = \log \sum_{y_r} \exp\left(\sum_{F \in \mathrm{nbr}_{\mathcal{G}}(r)} r_{F \to r}(y_r) \right).$$

BP: Root-to-Leaf Stage

2a. Schedule the root-to-leaf messages.

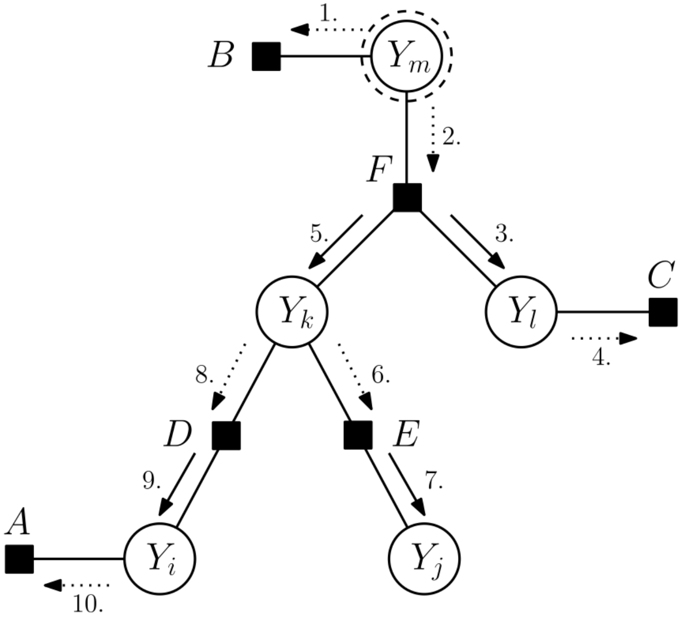


Figure: Belief propagation: root-to-leaf stage.

2b. Compute the root-to-leaf messages using the same formulas on page 12.

BP: Compute Factor / Variable Marginals

2c. Alongside Step 2b, combine messages and compute factor marginals:

$$\mu_F(y_F) := p(y_F) = \exp\Big(-E_F(y_F) + \sum_{i \in \mathrm{nbr}_{\mathcal{G}}(F)} q_{i\to F}(y_i) - \log Z \Big),$$

as well as variable marginals:

$$\mu_i(y_i) := p(y_i) = \exp\Big(\sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} r_{F\to i}(y_i) - \log Z \Big).$$

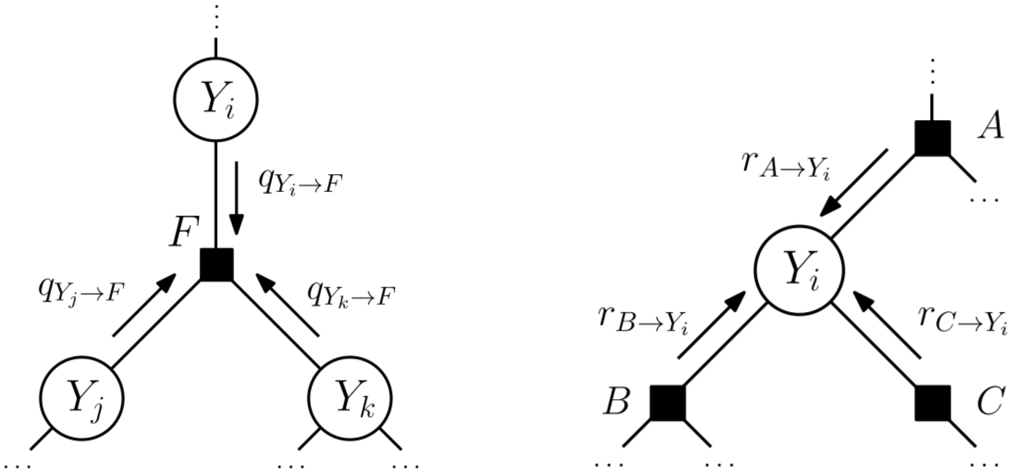


Figure: (left) Factor marginal; (right) Variable marginal.

BP on Pairwise MRFs (as exercise)

For a pairwise MRF $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, the joint distribution is factorized by

$$p(y) = \exp\Big(- \sum_{i \in \mathcal{V}} E_i(y_i) - \sum_{(i,j) \in \mathcal{E}} E_{ij}(y_i, y_j) - \log Z \Big).$$

BP on such pairwise MRF can be simplified:

- Variable-to-variable message is computed by

$$m_{i \to j}(y_j) = \log \sum_{y_i} \exp\Big(- E_i(y_i) - E_{ij}(y_i, y_j) + \sum_{k \in \mathrm{nbr}_{\mathcal{H}}(i) \setminus \{j\}} m_{k \to i}(y_i) \Big).$$

- Variable marginal is computed by

$$\mu_i(y_i) = \exp\Big(- E_i(y_i) + \sum_{k \in \mathrm{nbr}_{\mathcal{H}}(i)} m_{k \to i}(y_i) - \log Z \Big).$$

Further Reading

- Koller & Friedman, Chapters 9, 10.
- Murphy, Chapter 20.
- Nowozin & Lampert, Section 3.1.

Variational Inference

Outline of this Section

- Basic idea: Variational inference.
- Mean field (MF) method.
- Loopy belief propagation (LBP).

Approximation by Tractable Distributions

- Goal: probabilistic inference on joint distribution $p(y)$ represented by *general* MRF (i.e. possibly with loops).
- Instead of tackling the inference on p directly, we first seek for an approximation q within a family $\mathcal{Q}$ consisting of "tractable" distributions:

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q \,|\, p\right).$$

- The **Kullback-Leibler (KL) divergence** (a.k.a. *relative entropy*) between two distributions q, p (assuming the "absolute continuity" $q \ll p$) is defined by

$$\mathrm{KL}\left(q \,|\, p\right) = \sum_y q(y) \log \frac{q(y)}{p(y)}.$$

- Basic properties of KL:

1. $\mathrm{KL}\left(q \,|\, p\right) = 0$ iff $p = q$.
2. $\mathrm{KL}\left(q \,|\, p\right) \geq 0 \ \forall q, p$.
3. $\mathrm{KL}\left(\cdot \,|\, \cdot\right)$ is not symmetric. Nor does it satisfy the triangle inequality.

Preliminaries to Variational Inference

- Represented by a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$, p takes the form

$$p(y) = \exp\Big(- \sum_{F \in \mathcal{F}} E_F(y_F) - \log Z \Big).$$

- Plug p into KL divergence $\rightsquigarrow$

$$\begin{aligned} \mathrm{KL}\,(q \,|\, p) &= \sum_y q(y) \log \frac{q(y)}{p(y)} = \sum_y q(y) \log q(y) - \sum_y q(y) \log p(y) \\ &= -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F} \mu_F[q](y_F) E_F(y_F) + \log Z. \end{aligned}$$

- $H(q)$ is the **entropy** of distribution q.
- $\mu_F[q]$ is the marginal distribution of q over variables Y_F.
- $F_{\text{Gibbs}}(q; p) := \mathrm{KL}\,(q \,|\, p) - \log Z = -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F} \mu_F[q](y_F) E_F(y_F)$ is called the **Gibbs free energy**.
- $\mathrm{KL}\,(q \,|\, p) \geq 0 \;\Rightarrow\;$ $\log Z$ is lower bounded by $-F_{\text{Gibbs}}(q; p)$.

Mean Field Approximation

In (naive) **mean field** method, $\mathcal{Q}$ consists of q factorized by only unaries:

$$q(y) = \prod_{i \in \mathcal{V}} q_i(y_i).$$

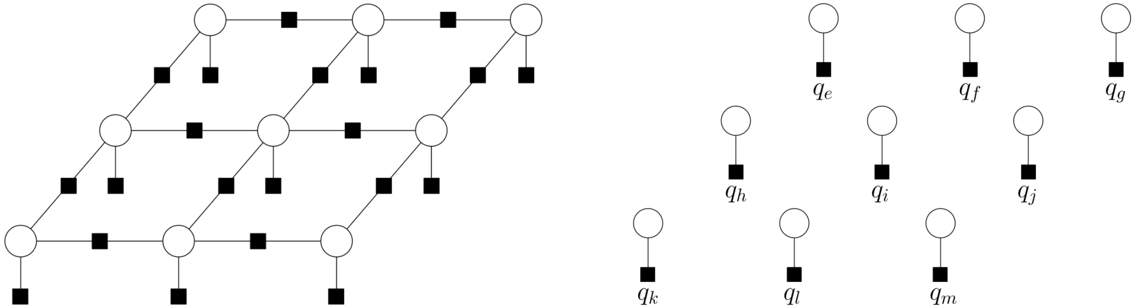


Figure: (left) Original factor graph; (right) (Naive) mean field approximation.

- Such q is "tractable" because $\{q_i(y_i)\}$ provide variable marginals.
- Quick facts: $$H(q) = \sum_{i \in \mathcal{V}} H(q_i) = -\sum_{i \in \mathcal{V}} \sum_{y_i} q_i(y_i) \log q_i(y_i),$$ $$\mu_F[q](y_F) = \prod_{i \in \text{nbr}_{\mathcal{G}}(F)} q_i(y_i).$$

Mean Field (MF) Approximation

Derivation of MF approximation:

$$\begin{aligned} q^* &= \arg\min_{q\in\mathcal{Q}} \mathrm{KL}\,(q \,|\, p) = \arg\min_{q\in\mathcal{Q}} F(q; p) \\ &= \arg\min_{q\in\mathcal{Q}} -H(q) + \sum_{F\in\mathcal{F}} \sum_{y_F} \mu_F[q](y_F) E_F(y_F) \\ &= \arg\min_{\{q_i\}_{i\in\mathcal{V}}} \sum_{i\in\mathcal{V}} \sum_{y_i} q_i(y_i) \log q_i(y_i) + \sum_{F\in\mathcal{F}} \sum_{y_F} \Bigg(\prod_{i\in \mathrm{nbr}_{\mathcal{G}}(F)} q_i(y_i) \Bigg) E_F(y_F). \end{aligned}$$

Each q_i lies in the probability simplex Δ_i, i.e.

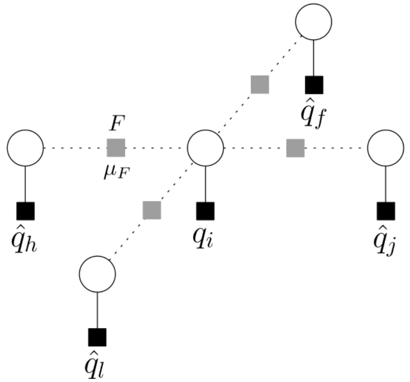
$$q_i(y_i) \geq 0 \;\; \forall y_i,$$
$$\sum_{y_i} q_i(y_i) = 1.$$

The optimization can be resolved by *coordinate descent* (next slide).

MF Update Formula

For each block q_i, fix $\widehat{q}_{i'}(y_{i'}) = q_{i'}(y_{i'})\ \forall i' \neq i$ and solve:

$$q_i^* = \arg\min_{q_i \in \Delta_i} \sum_{y_i} q_i(y_i) \log q_i(y_i) + \sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} \sum_{y_F} \Bigg(\prod_{i' \in \mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\}} \widehat{q}_{i'}(y_{i'}) \Bigg) q_i(y_i) E_F(y_F).$$



We obtain an analytical solution via Lagrange multiplier λ for $\sum_{y_i} q_i^*(y_i) = 1$:

$$q_i^*(y_i) = \exp\Bigg(-1 - \sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} \sum_{y_{F\setminus\{i\}}} \Bigg(\prod_{i' \in \mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\}} \widehat{q}_{i'}(y_{i'}) \Bigg) E_F(y_F) + \lambda \Bigg)$$

$$\propto \exp\Bigg(- \sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} \sum_{y_{F\setminus\{i\}}} \Bigg(\prod_{i' \in \mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\}} \widehat{q}_{i'}(y_{i'}) \Bigg) E_F(y_F) \Bigg).$$

Some Remarks on MF

- The term $\prod_{i' \in \text{nbr}_{\mathcal{G}}(F) \setminus \{i\}} \widehat{q}_{i'}(y_{i'})$ is taken to be 1 if $\text{nbr}_{\mathcal{G}}(F) \setminus \{i\} = \emptyset$.
- For a pairwise MRF $\mathcal{H}$, the MF update rule can be simplified as

$$q_i^*(y_i) \propto \exp\Big(- E_i(y_i) - \sum_{j \in \text{nbr}_{\mathcal{H}}(i)} \sum_{y_j} \widehat{q}_j(y_j) E_{ij}(y_i, y_j) \Big).$$

- MF is an iterative procedure which converges to a *locally optimal* solution q^*.
- Upon convergence, $\{q_i^*\}$ directly provide (approximate) variable marginals.
- The tractable family $\mathcal{Q}$ can be more sophisticated than factorizations of unaries in naive mean field. $\rightsquigarrow$ *Structured mean field* approximation.

From Belief Propagation to Loopy Belief Propagation

- Previously we have seen how belief propagation works on tree factor graphs.
- We can use similar update rules to derive **loopy belief propagation** (LBP).
- Although LBP does not guarantee the convergence (if at all) to the true marginal, it often performs well and is widely used in practice[3].
- In the following, we first present the LBP algorithm and then interpret it from perspective of variational inference.

[3] Murphy et al., “Loopy Belief Propagation for Approximate Inference: An Empirical Study”.

Loopy Belief Propagation

On a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$, LBP proceeds as follows.

0. Initialize all variable-to-factor messages: $q_{i\to F}(y_i) = 0$. Then iterate:
1. Compute all factor-to-variable messages:

$$r_{F\to i}(y_i) = \log \sum_{y_{F\setminus\{i\}}} \exp\Big(- E_F(y_F) + \sum_{i'\in \mathrm{nbr}_{\mathcal{G}}(F)\setminus\{i\}} q_{i'\to F}(y_{i'})\Big).$$

2. Compute all (normalized) variable-to-factor messages:

$$\bar{q}_{i\to F}(y_i) = \sum_{F'\in \mathrm{nbr}_{\mathcal{G}}(i)\setminus\{F\}} r_{F'\to i}(y_i),$$

$$\delta_{i\to F} = \log \sum_{y_i} \exp\Big(\bar{q}_{i\to F}(y_i)\Big),$$

$$q_{i\to F}(y_i) = \bar{q}_{i\to F}(y_i) - \delta_{i\to F}.$$

Loopy Belief Propagation (cont'd)

3. Compute all factor marginals:

$$\mu_F(y_F) \propto \exp\Big(- E_F(y_F) + \sum_{i \in \mathrm{nbr}_{\mathcal{G}}(F)} q_{i \to F}(y_i) \Big).$$

4. Compute all variable marginals:

$$\mu_i(y_i) \propto \exp\Big(\sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} r_{F \to i}(y_i) \Big).$$

Differences compared to BP:

- The normalization constants in the computation of marginals differ at each factor/variable.
- The log partition function is not directly available, but it can be approximated by the **Bethe free energy**:

$$- \log Z \approx F_{\mathrm{Bethe}}(\mu; p) := \sum_{i \in \mathcal{V}} (1 - |\,\mathrm{nbr}_{\mathcal{G}}(i)|) \sum_{y_i} \mu_i(y_i) \log \mu_i(y_i)$$
$$+ \sum_{F \in \mathcal{F}} \sum_{y_F} \mu_F(y_F) \Big(E_F(y_F) + \log \mu_F(y_F) \Big).$$

Interpretation of LBP

On a pairwise MRF $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, LBP can be interpreted as an attempt to solve:

$$\underset{\{\mu_i\}_{i\in\mathcal{V}},\{\mu_{ij}\}_{(i,j)\in\mathcal{E}}}{\text{minimize}} \sum_{i\in\mathcal{V}}(1-|\operatorname{nbr}_{\mathcal{H}}(i)|)\sum_{y_i}\mu_i(y_i)\log\mu_i(y_i)$$

$$+\sum_{(i,j)\in\mathcal{E}}\sum_{y_i,y_j}\mu_{ij}(y_i,y_j)\Big(E_{ij}(y_i,y_j)+\log\mu_{ij}(y_i,y_j)\Big)$$

$$\text{subject to } \mu_i(y_i)\geq 0,\ \mu_{ij}(y_i,y_j)\geq 0,\ \sum_{y_i}\mu_i(y_i)=1,\ \sum_{y_i}\mu_{ij}(y_i,y_j)=\mu_j(y_j).$$

- The constraints impose *local consistency* between node marginals $\{\mu_i\}$ and edge marginals $\{\mu_{ij}\}$.
- However, $\{\mu_i\}, \{\mu_{ij}\}$ under these constraints are may not be marginals of any joint distribution on $\mathcal{H}$ (i.e. outer approximation of *marginal polytope*).
- LBP updates can be derived from an iterative algorithm for the above constrained optimization.
- An amazing theory on variational inference arise in this context — we point those interested to the "monster" paper [Jordan & Wainwright, 2008].

LBP vs. MF

$(+)$ (Naive) MF optimizes over only variable marginals; LBP optimizes over variable and factor marginals under local consistency constraints.

$(+)$ LBP does exact inference on factor graphs without loops; MF is exact on a strict subclass of factor graphs, on which all true factor marginals are factorized by $\mu_F(y_F) = \prod_{i \in \mathrm{nbr}_{\mathcal{G}}(F)} \mu_i(y_i)$ (hence an inner approximation of marginal polytope).

$(+)$ While both being approximate inference techniques, LBP tends to be more accurate than MF in practice.

$(-)$ MF provides a lower bound of the log partition function (given by negative Gibbs free energy), while LBP does not.

$(-)$ Compared to LBP, it is easier to extend MF to distributions other than discrete and Gaussian, due to the simplicity of working with only variable marginals.

Further Reading

- Murphy, Chapters 21, 22.
- Nowozin & Lampert, Sections 3.2, 3.3.
- Koller & Friedman, Chapter 11.
- Jordan & Wainwright, Chapters 4, 5.

Sampling-based Inference

Outline of the Section

- Monte Carlo (MC) method.
- Markov chain Monte Carlo (MCMC) method.
- Sampling of Bayesian network and Markov random field.

Basic Principle of Sampling

Given a distribution p, we can approximate p using a finite sequence of **samples** $\{x_n\}_{n=1}^N$ in the sense that:

$$\mathbb{E}_{x\sim p}[f(x)] = \sum_x f(x)p(x) \approx \frac{1}{N}\sum_{n=1}^N f(x_n) \quad \text{for any function } f.$$

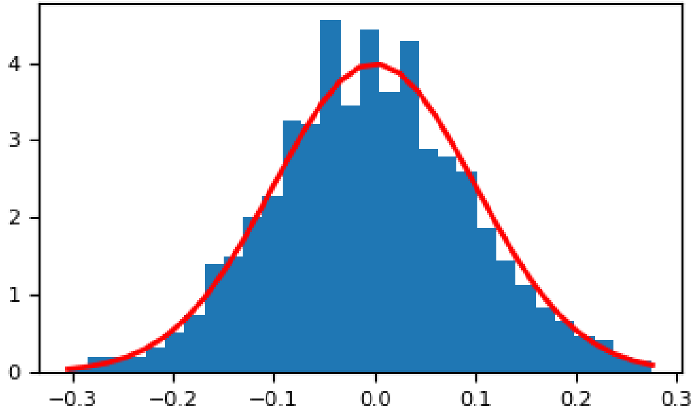


Figure: Sampling of a Gaussian[4].

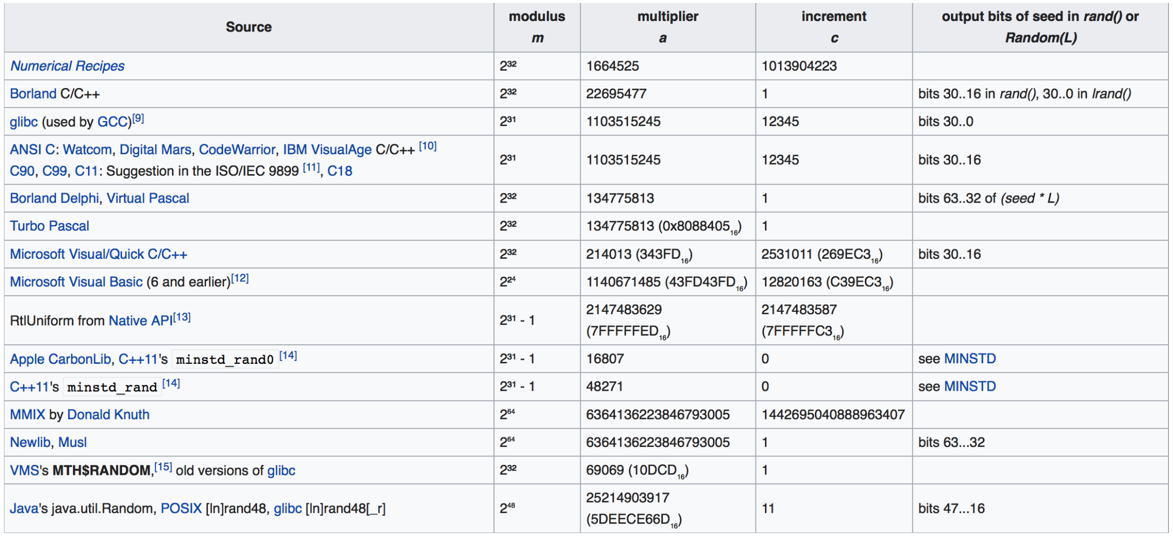
[4] https://docs.scipy.org/doc/numpy/reference/generated/numpy.random.normal.html

Pseudo-Random Number Generator

Linear congruential generator for sampling Unif(0, 1):

$$x_{n+1} = (a \cdot x_n + c) \mod m.$$

- Most fundamental sampler above all.
- The generated samples are *pseudo-random* — $\{x_n\}$ are "deterministic" if the generator (i.e. parameters a, c, m) and the *seed* x_0 are fixed.



Source	modulus *m*	multiplier *a*	increment *c*	output bits of seed in *rand()* or *Random(L)*
Numerical Recipes	2^{32}	1664525	1013904223	
Borland C/C++	2^{32}	22695477	1	bits 30..16 in *rand()*, 30..0 in *lrand()*
glibc (used by GCC)[9]	2^{31}	1103515245	12345	bits 30..0
ANSI C: Watcom, Digital Mars, CodeWarrior, IBM VisualAge C/C++ [10] C90, C99, C11: Suggestion in the ISO/IEC 9899 [11], C18	2^{31}	1103515245	12345	bits 30..16
Borland Delphi, Virtual Pascal	2^{32}	134775813	1	bits 63..32 of *(seed * L)*
Turbo Pascal	2^{32}	134775813 (0x8088405$_{16}$)	1	
Microsoft Visual/Quick C/C++	2^{32}	214013 (343FD$_{16}$)	2531011 (269EC3$_{16}$)	bits 30..16
Microsoft Visual Basic (6 and earlier)[12]	2^{24}	1140671485 (43FD43FD$_{16}$)	12820163 (C39EC3$_{16}$)	
RtlUniform from Native API[13]	$2^{31} - 1$	2147483629 (7FFFFFED$_{16}$)	2147483587 (7FFFFFC3$_{16}$)	
Apple CarbonLib, C++11's `minstd_rand0` [14]	$2^{31} - 1$	16807	0	see MINSTD
C++11's `minstd_rand` [14]	$2^{31} - 1$	48271	0	see MINSTD
MMIX by Donald Knuth	2^{64}	6364136223846793005	1442695040888963407	
Newlib, Musl	2^{64}	6364136223846793005	1	bits 63...32
VMS's **MTH$RANDOM**,[15] old versions of glibc	2^{32}	69069 (10DCD$_{16}$)	1	
Java's java.util.Random, POSIX [ln]rand48, glibc [ln]rand48[_r]	2^{48}	25214903917 (5DEECE66D$_{16}$)	11	bits 47...16

Figure: Common used linear congruential generators[5].

[5] https://en.wikipedia.org/wiki/Linear_congruential_generator

Sampling Gaussians

- Sample univariate Gaussian distribution by **Box-Muller method**:

1. Sample $(z_1, z_2) \sim p_z(z_1, z_2) = \frac{1}{\pi}\mathbf{1}\{z_1^2 + z_2^2 \leq 1\}$ (i.e. uniform distribution supported on the unit 2D circle).
2. Perform the Box-Muller transformation and output x_1, x_2:

$$x_i = z_i\sqrt{\frac{-2\log(z_1^2+z_2^2)}{z_1^2+z_2^2}}, \quad i \in \{1,2\}.$$

Fact: x_1, x_2 are two independent samples of Normal$(0, 1)$:

$$p_x(x_1, x_2) = p_z(z_1, z_2)\left|\frac{\partial(z_1, z_2)}{\partial(x_1, x_2)}\right| = \frac{1}{\sqrt{2\pi}}\exp(-x_1^2/2) \cdot \frac{1}{\sqrt{2\pi}}\exp(-x_2^2/2).$$

- Sample multivariate Gaussian distribution, $y \sim$ Normal(μ, Σ), by:

1. Perform Cholesky decomposition $\Sigma = LL^\top$.
2. Sample $x \sim$ Normal$(0, I)$, and output $y := Lx + \mu$.

Fact: $\mathbb{E}[y] = \mu$, and $\text{Var}[y] = L\,\text{Var}[x]L^\top = LIL^\top = \Sigma$.

Sampling by Inverse CDF

Sample a distribution via **inverse Cumulative Distribution Function**:

- Let $U \sim \text{Unif}(0,1)$ and F_p be the CDF for (univariate) distribution p, i.e.

$$F_p(y) := \int_{-\infty}^{y} p(x)dx = \int_{-\infty}^{\infty} \mathbf{1}\{x \leq y\}p(x)dx.$$

- Note that $X \sim p \;\Leftrightarrow P(X \leq y) = F_p(y)$.

- We assert $F_p^{-1}(U) \sim p$, since

$$\begin{aligned} P(F_p^{-1}(U) \leq y) &= P(U \leq F_p(y)) && \text{(since } F_p \text{ is monotone)} \\ &= F_p(y). && \text{(since } P(U \leq u) = u \;\; \forall u \in [0,1]) \end{aligned}$$

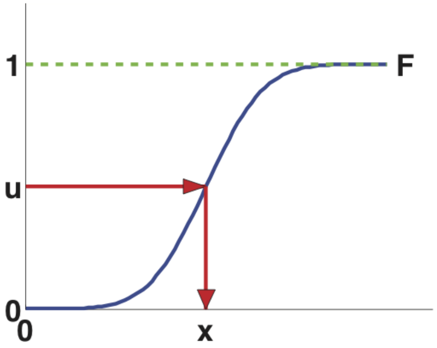


Figure: Sampling using inverse CDF [Murphy, Figure 23.1].

Rejection Sampling

- Inverse CDF sampling requires explicit knowledge of F_p^{-1}.
- **Rejection Sampling**:

  Require: *unnormalized* target distribution $\widetilde{p}$ (i.e. $\widetilde{p}(x)/Z_p = p(x)$ for target distribution p), *proposal distribution* q and constant $M > 0$ s.t. $Mq(x) \geq \widetilde{p}(x)\ \forall x$ ($\Rightarrow p \ll q$).

1. Sample $x \sim q$, and $u \sim \text{Unif}(0, 1)$.
2. If $u > \frac{\widetilde{p}(x)}{Mq(x)}$, reject the proposed sample x; otherwise, accept x.

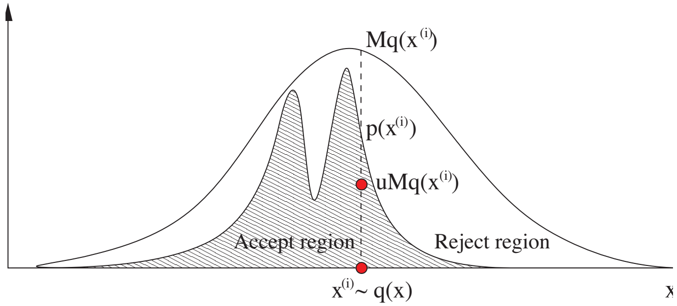


Figure: Rejection sampling [Murphy, Figure 23.2].

- Proof: (univariate case) $P(x \leq y | x \text{ accepted}) = \frac{P(x \leq y,\ x \text{ accepted})}{P(x \text{ accepted})} =$

$$\frac{\iint \mathbf{1}\{u \leq \widetilde{p}(x)/(Mq(x)),\ x \leq y\} q(x) du\, dx}{\iint \mathbf{1}\{u \leq \widetilde{p}(x)/(Mq(x))\} q(x) du\, dx} = \frac{\frac{1}{M}\int_{-\infty}^{y} \widetilde{p}(x) dx}{\frac{1}{M}\int_{-\infty}^{\infty} \widetilde{p}(x) dx} = F_p(y).$$

Importance Sampling

- In rejection sampling, $P(x \text{ accepted}) = \frac{1}{M}\int_{-\infty}^{\infty}\widetilde{p}(x)dx$, i.e., many proposed samples are potentially wasted.
- In contrast, **importance sampling** uses all samples by weighting them:

$$\mathbb{E}_{x\sim p}[f(x)] = \int f(x)\frac{p(x)}{q(x)}q(x)dx \approx \frac{1}{N}\sum_{n=1}^{N} w_n f(x_n),$$

with $x_n \sim q$ i.i.d. and $w_n = \frac{p(x_n)}{q(x_n)}$.

- Extend importance sampling to *unnormalized* distributions $\widetilde{p}$, $\widetilde{q}$:

$$\mathbb{E}_{x\sim p}[f(x)] = \frac{Z_q}{Z_p}\int f(x)\frac{\widetilde{p}(x)}{\widetilde{q}(x)}q(x)dx \approx \frac{Z_q}{Z_p}\frac{1}{N}\sum_{n=1}^{N}\frac{\widetilde{p}(x_n)}{\widetilde{q}(x_n)}f(x_n), \quad x_n \sim q \text{ i.i.d.}$$

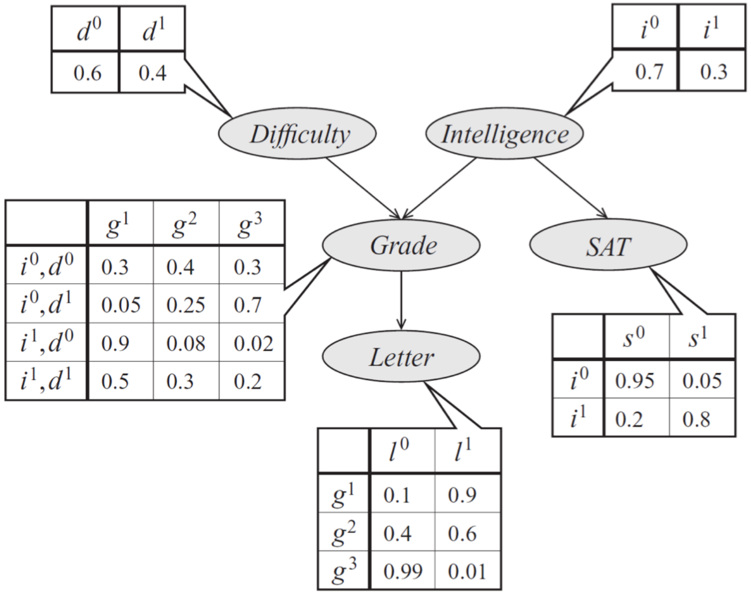
$$\frac{Z_p}{Z_q} = \int \frac{1}{Z_q}\widetilde{p}(x)dx = \int \frac{\widetilde{p}(x)}{\widetilde{q}(x)}q(x)dx \approx \frac{1}{N}\sum_{n=1}^{N}\frac{\widetilde{p}(x'_n)}{\widetilde{q}(x'_n)}, \quad x'_n \sim q \text{ i.i.d.}$$

We often take $x'_n = x_n$. For finite N, this yields a *biased estimator* of p.

Sampling of Bayesian Network

Recall that the distribution represented by BN is given by

$$p(x) = \prod_{i \in \mathcal{V}} p\big(x_i \big| (x_j)_{j \in \mathrm{Pa}_{\mathcal{G}}(i)}\big).$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Ancestral sampling: Given that no variables are observed, we can follow the topological order of the BN and sample each individual conditional distribution.

Sampling of BN with Evidence

In case the BN $\mathcal{G}$ contains observed nodes (called **evidence**), we can modify ancestral sampling (AS) as follows:

- **Logic sampling**: Perform AS. Whenever a sampled node takes different value from the evidence, reject the whole sample and start again.
- LS is closely related to rejection sampling. Unsurprisingly, it is inefficient for wasting samples.
- **Likelihood weighting**: Perform AS. Whenever node i is observed (written $i \in \mathcal{O}$), we *clamp* the observed value $\bar{x}_i$ and *weight* the whole sample by the probability of the clamped node $p(\bar{x}_i | x_{\mathrm{Pa}_{\mathcal{G}}(i)})$.
- LW can be interpreted as importance sampling with weights given by:

$$w(x) = \frac{p(x)}{q(x)} = \frac{\prod_{i \in \mathcal{V}} p(x_i | x_{\mathrm{Pa}_{\mathcal{G}}(i)})}{\prod_{i \in \mathcal{V} \setminus \mathcal{O}} p(x_i | x_{\mathrm{Pa}_{\mathcal{G}}(i)}) \prod_{i \in \mathcal{O}} \delta_{\bar{x}_i}(x_i)} = \prod_{i \in \mathcal{O}} p(\bar{x}_i | x_{\mathrm{Pa}_{\mathcal{G}}(i)}).$$

$\delta_{\bar{x}}$ denotes the **Dirac distribution** defined by $\delta_{\bar{x}}(x) = \begin{cases} 1 & \text{if } x = \bar{x}, \\ 0 & \text{otherwise.} \end{cases}$

Towards Markov Chain Monte Carlo

- Monte Carlo sampling requires exact or rough knowledge of the partition function (of a MRF), hence impractical for high dimensional distributions.
- Instead of generating i.i.d. samples, **Monte Carlo Markov Chain** (MCMC) constructs a *Markov chain* using *"adaptive" proposal distributions*, in a way that the Markov chain converges to a *stationary distribution* identical to the target distribution.

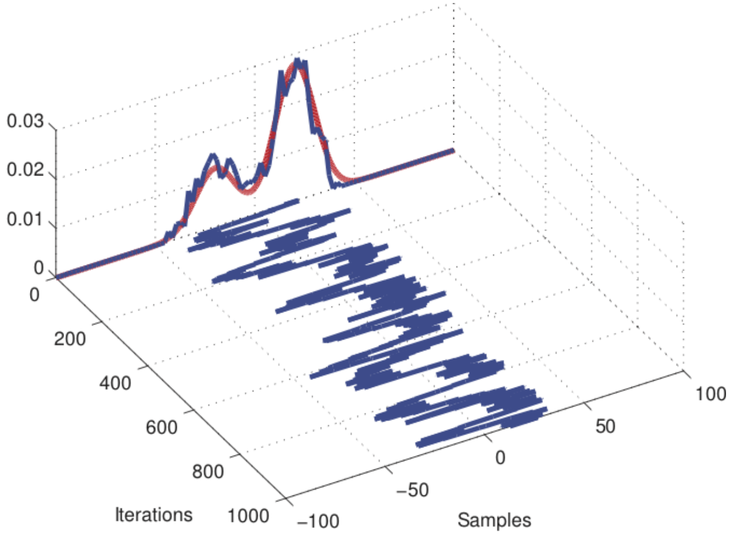


Figure: Sampling by MCMC [Murphy, Figure 24.7].

Markov Chain

- The (discrete-time) **Markov chain** (MC) is a sequence of RVs $(X_n)_{n=1}^{\infty}$ satisfying the Markov property:

$$P(X_{n+1} = x | X_1, ..., X_n \text{ given}) = P(X_{n+1} = x | X_n \text{ given}).$$

"The future depends on the past only through the present."

- Further assume:

1. All X_n has a *finite state space* $\mathcal{X}$.
2. The MC is *time-homogeneous*, i.e., the transition probability is time-independent

$$P(X_{n+1} = x' | X_n = x) =: \pi(x'|x) \quad \forall n,$$

with $\pi(x'|x) \geq 0$, $\sum_{x'} \pi(x'|x) = 1$. π is the **transition kernel** of the MC.

- Denote by p_n the distribution at time step n:

$$p_n(x) = P(X_n = x) \quad \Rightarrow \quad p_{n+1}(x') = \sum_x p_n(x)\pi(x'|x).$$

Relevant Notions on Markov Chain

- p_* is a **stationary distribution** for the MC if

$$p_*(x') = \sum_x p_*(x)\pi(x'|x) \;\; \forall x' \in \mathcal{X}.$$

- The MC is **irreducible** if

$$\forall x, x' \in \mathcal{X} \;\; \exists n(x, x') \;\; \text{s.t. } P(X_n = x'|X_0 = x) > 0,$$

  i.e., it is possible to get to any state from any state in finite steps.

- A state $x \in \mathcal{X}$ has *period* T_x if

$$T_x = \text{gcd}\{n > 0 : P(X_n = x|X_0 = x) > 0\},$$

  i.e., any loop over state x must occur in a multiple of T_x steps.
  We say the MC is **aperiodic** if $T_x = 1 \; \forall x \in \mathcal{X}$.

- The MC is **regular** if

$$\exists n \;\; \text{s.t. } P(X_n = x'|X_0 = x) > 0 \;\; \forall x, x' \in \mathcal{X}.$$

  Fact: MC is regular $\Rightarrow$ MC is irreducible and aperiodic.

Convergence to Stationary Distribution

Theorem 1: If the transition kernel π of a Markov chain satisfies the *detailed balance condition* for some distribution p_*:

$$p_*(x)\pi(x'|x) = p_*(x')\pi(x|x') \quad \forall x, x' \in \mathcal{X},$$

then p_* is a stationary distribution for the Markov chain.

Proof: $\sum_x p_*(x)\pi(x'|x) = \sum_x p_*(x')\pi(x|x') = p_*(x')\sum_x \pi(x|x') = p_*(x')$.

Theorem 2[6]: Every irreducible, aperiodic, finite-state Markov chain has a limiting distribution

$$p_*(x') = \lim_{n\to\infty} \sum_x P(X_n = x'|X_0 = x)p_0(x),$$

regardless of the initial distribution p_0. Indeed, p_* is equal to the unique stationary distribution of the MC.

[6] [Murphy, Theorem 17.2.1]

Metropolis-Hastings Algorithm

Metropolis-Hastings (MH) algorithm:

Input: unnormalized target distribution $\widetilde{p}$ (i.e. $p_*(x) = \widetilde{p}(x)/Z_p$), proposal distribution $q(\cdot|\cdot)$, initial sample x_0. Loop $n = 0, 1, 2, \ldots$ as follows:

1. Set $x = x_n$. Sample $x' \sim q(x'|x)$.
2. Compute acceptance probability $\alpha = \dfrac{\widetilde{p}(x')q(x|x')}{\widetilde{p}(x)q(x'|x)}$.
3. Compute $r = \min(1, \alpha)$. Sample $u \sim \text{Unif}(0, 1)$.
4. Set new sample to: $x_{n+1} = \begin{cases} x' & \text{if } u < r, \\ x_n & \text{if } u \geq r. \end{cases}$

Some remarks:

- For a given target distribution p_*, a proposal distribution q is valid if $\text{supp}(p_*) \subset \cup_x \text{supp}(q(\cdot|x))$, i.e. $\forall x'$ with $p_*(x') > 0 \; \exists x$ s.t. $q(x'|x) > 0$.
- If q is symmetric, i.e. $q(x'|x) = q(x|x')$, then MH simplifies to the Metropolis algorithm with $\alpha = \frac{\widetilde{p}(x')}{\widetilde{p}(x)}$. Hastings made the correction for asymmetric q.

Analysis of MH Algorithm

We analyze with convergence of the MH algorithm:

1. MH generates a Markov chain with the transition kernel:

$$\pi(x'|x) = \begin{cases} q(x'|x)r(x'|x) & \text{if } x' \neq x, \\ q(x|x) + \sum_{x' \neq x} q(x'|x)(1 - r(x'|x)) & \text{if } x' = x. \end{cases}$$

   $r(x'|x)$ is the conditional probability that x' is accepted after being proposed. We will show that the Markov chain satisfies the detailed balance condition:

$$p_*(x)\pi(x'|x) = p_*(x')\pi(x|x').$$

2. Let two states x and x' ($x \neq x'$) be arbitrarily fixed. Either

$$p_*(x)\pi(x'|x) \leq p_*(x')\pi(x|x'), \tag{$\dagger$}$$

   or the reversed inequality holds. Without loss of generality, we proceed with inequality (†).

Analysis of MH Algorithm (cont'd)

$$p_*(x)\pi(x'|x) \leq p_*(x')\pi(x|x'). \tag{$\dagger$}$$

3. $(\dagger) \;\Rightarrow\; \alpha(x'|x) = \dfrac{p_*(x')q(x|x')}{p_*(x)q(x'|x)} \leq 1 \;\Rightarrow\; r(x'|x) = \alpha(x'|x)$

   $\Rightarrow\; \pi(x'|x) = q(x'|x)r(x'|x) = q(x'|x)\dfrac{p_*(x')q(x|x')}{p_*(x)q(x'|x)} = \dfrac{p_*(x')}{p_*(x)}q(x|x').$

4. $(\dagger) \;\Rightarrow\; \alpha(x|x') = \dfrac{p_*(x)q(x'|x)}{p_*(x')q(x|x')} \geq 1 \;\Rightarrow\; r(x|x') = 1$

   $\Rightarrow\; \pi(x|x') = q(x|x')r(x|x') = q(x|x').$

5. Combining (3) and (4), we conclude that $p_*(x)\pi(x'|x) = p_*(x')\pi(x|x')$. Hence, by Theorem 1, p_* is a stationary distribution for the Markov chain.

6. If in addition the Markov chain generated by the MH algorithm is irreducible and aperiodic, then by Theorem 2 the Markov chain converges to the unique stationary distribution p_*.

Gibbs Sampling

Gibbs sampling:

Input: unnormalized target distribution $\widetilde{p}((x_i)_{i\in\mathcal{V}})$, initial sample x^0.

Loop $n \in \{0, 1, 2, ...\},\ i \in \mathcal{V}$:

Sample $x_i^{n+1} \sim p(x_i | x^{n+1}_{\{0,...,i-1\}}, x^n_{\{i+1,...,|\mathcal{V}|\}})$.

Some remarks:

- If p is represented by a graphical model (i.e. BN or MRF), then sampling of x_i^{n+1} only involves the Markov blanket of i.
- Gibbs sampling can be interpreted as the MH algorithm with the proposal:

$$q(x'|x) = \delta_{x_{\mathcal{V}\setminus\{i\}}}(x'_{\mathcal{V}\setminus\{i\}})p(x'_i|x_{\mathcal{V}\setminus\{i\}}),$$

and 100% acceptance rate:

$$\alpha = \frac{p(x')q(x|x')}{p(x)q(x'|x)} = \frac{p(x'_i|x'_{\mathcal{V}\setminus\{i\}})p(x'_{\mathcal{V}\setminus\{i\}})\delta_{x_{\mathcal{V}\setminus\{i\}}}(x'_{\mathcal{V}\setminus\{i\}})p(x_i|x'_{\mathcal{V}\setminus\{i\}})}{p(x_i|x_{\mathcal{V}\setminus\{i\}})p(x_{\mathcal{V}\setminus\{i\}})p(x'_i|x_{\mathcal{V}\setminus\{i\}})} = 1.$$

Example: Gibbs Sampling for Pairwise CRF

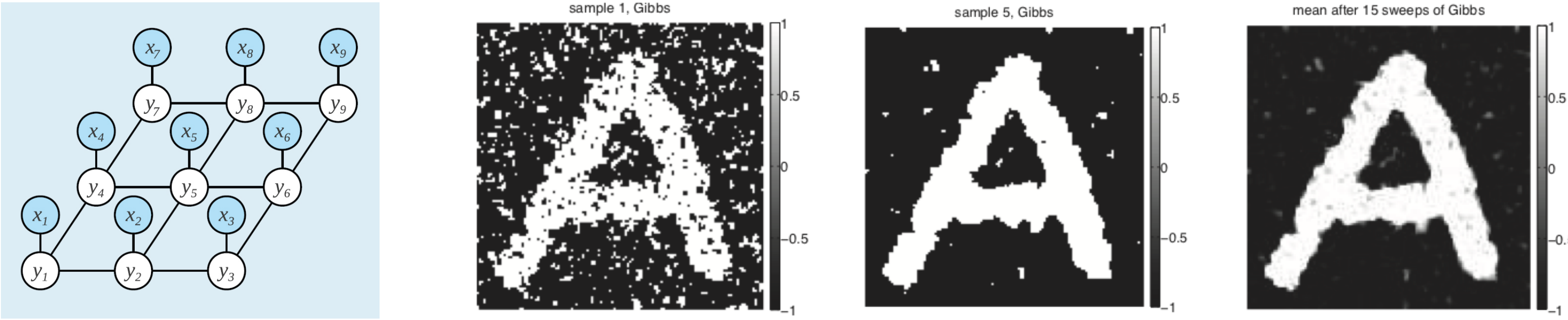


Figure: Gibbs Sampling for Pairwise CRF[7].

We can apply Gibbs sampling to find

$$y \sim p(y|x) \propto \exp\left(-\sum_{i \in \mathcal{V}} E_i(y_i; x_i) - \sum_{(i,j) \in \mathcal{E}} E_{ij}(y_i, y_j)\right).$$

For each $i \in \mathcal{V}$, sample (e.g. by inverse CDF method):

$$y_i^{n+1} \sim p(y_i | x_i, y^n_{\text{nbr}(i)}) \propto \exp\left(-E_i(y_i) - \sum_{j \in \text{nbr}(i)} E_{ij}(y_i, y_j^n)\right).$$

[7] Sampled images taken from [Murphy, Figure 24.1].

Further Reading

- Murphy, Chapters 23, 24.
- Nowozin & Lampert, Sections 3.4.
- Koller & Friedman, Chapter 12.