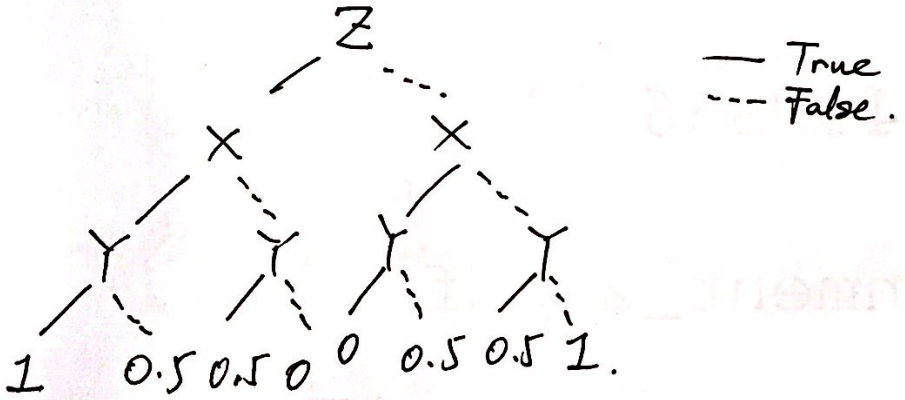
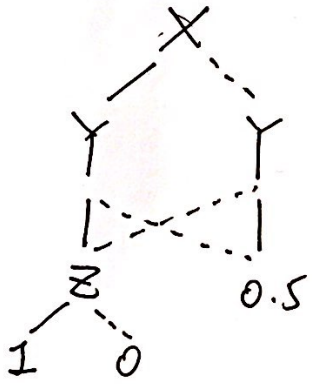
Q1. Sol: (a). Tabular

X	Y	Z	Pr
T	T	T	1
T	T	F	0
F	T	T	0.5
F	T	F	0.5
T	F	T	0.5
T	F	F	0.5
F	F	T	0
F	F	F	0

(b). Tree.



(c).



1. The ADD is compact, therefore, it is effecient in space using.
2. ADD operations can avoid state enumeration.

Q2. Sol: For detailed balance, if the finite-state Markov chain has a unique distribute s.t.

$$\pi(x)T(x \to x') = \pi(x')T(x' \to x)$$

Then we could say the markov chain is detailed balance.

Since $\pi(x)T(x \to x') = \pi(x)q(x'|x)\min\left[1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right]$

where $\min\left[1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right] = a(x'|x^{(t-1)})$ is the acceptance rate

$$\Rightarrow \pi(x)T(x \to x') = \min\left(\pi(x)q(x'|x), \pi(x')q(x|x')\right)$$

$$= \pi(x')q(x|x')\min\left[1, \frac{\pi(x)q(x'|x)}{\pi(x')q(x|x')}\right]$$

$$= \pi(x')q(x|x')a(x|x'^{(t-1)})$$

$$= \pi(x')T(x' \to x)$$

$\Rightarrow$ The Metropolic-Hastings holds the detailed-balance property

Q3. Sol: a. Since $P(y|x) = \frac{\pi_i \Psi_i(y_i, x_i)}{Z(x)}$, $D = \{(x^d, y^d)\}_d$.

where $\Psi_i = \exp[\Sigma_j \lambda_{i,j} f_{i,j}(y_i, x_i)]$.

$\Rightarrow \log P(y|x) = \Sigma_i \Sigma_j \lambda_{i,j} f_{i,j}(y_i, x_i) - \log Z(x)$.

Since all data samples are about to be used to train the same stacked graph, CRF.

$\Rightarrow \log P(y|x) = \Sigma_i \Sigma_j \lambda_j f_j(y_i, x_i) - \log Z(x)$

$\Rightarrow \lambda_j^* = \arg\max_{\lambda_j} (\Sigma_i \Sigma_j \lambda_j f_j(y_i, x_i) - \log Z(x))$

since $\frac{\partial \log P}{\partial \lambda_j} = \Sigma_i f_j(y_i, x_i) - \frac{\partial}{\partial \lambda_j} \log Z(x)$.

$= \Sigma_i [f_j(y_i, x_i) - \frac{\partial}{\partial \lambda_j} \log Z(x_i)]$.

since $Z(x_i) = \Sigma_y \pi_j \Psi_j(x_i, y)$

$\Rightarrow \frac{\partial \log P}{\partial \lambda_j} = \Sigma_i [f_j(y_i, x_i) - \Sigma_y f_j(x_i, y)]$.

$\frac{1}{I} \cdot \frac{\partial \log P}{\partial \lambda_j} = \frac{1}{I} \Sigma_i [f_j(y_i, x_i) - \Sigma_y f_j(x_i, y)]$

$= E_{(y_i|x_i)} f_j(x_i, y_i) - I \cdot E_{(y|x_i)} f_j(x_i, y)$.

Observed mean (from empirical data) — Expected mean.

b). For MRF, the denominator would be Z, which $= \Sigma_y \Sigma_x \pi_j \Psi_j(x, y) =$ constant for each λ_j possible value.

$\Rightarrow$ In MRF, the second term would be computed once in a single iteration of SGD. However, in CRF, the second term would be computed I times in a single iteration of SGD.