Chapter 22. More variational inference

22.1.
For binary MRFs, the total energy is given by,

$$E(\mathbf{x}) = \sum_{i=1}^{n}[E_{x_i}(1)x_i + E_{x_i}(0)(1-x_i)] + \sum_{i=1}^{n}\sum_{j=1}^{n}[E_{x_i,x_j}(0,0)(1-x_i)(1-x_j)$$

$$+E_{x_i,x_j}(0,1)(1-x_i)x_j + E_{x_i,x_j}(1,0)x_i(1-x_j) + E_{x_i,x_j}(1,1)x_ix_j]$$

$$= \text{const}+\sum_{i=1}^{n}[(E_{x_i}(1)-E_{x_i}(0)+\sum_{j=1}^{n}(E_{x_i,x_j}(1,0)-E_{x_i,x_j}(0,0)+E_{x_j,x_i}(0,1)-E_{x_j,x_i}(0,0)))x_i]$$

$$+\sum_{i=1}^{n}\sum_{j=1}^{n}(E_{x_i,x_j}(0,0) + E_{x_i,x_j}(1,1) - E_{x_i,x_j}(1,0) - E_{x_i,x_j}(0,1))x_ix_j$$

$$= \text{const} + \sum_{i=1}^{n}(E'_{x_i}(1) - E'_{x_i}(0))x_i - \sum_{i=1}^{n}\sum_{j=1}^{n} E'_{x_i,x_j}(0,1)x_ix_j$$

$$= \text{const} + \sum_{i=1}^{n}(E'_{x_i}(1) - E'_{x_i}(0))x_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} E'_{x_i,x_j}(0,1)(x_i - x_j)^2.$$

The capacity of the cut of the graph is given by,

$$C(\mathbf{x}) = \sum_{p\in s\cup\{x_i:x_i=0\}}\ \sum_{q\in t\cup\{x_i:x_i=1\}} |\bar{pq}|$$

$$= \sum_{i=1}^{n} x_i \max(0, E'_{x_i}(1) - E'_{x_i}(0)) + \sum_{i=1}^{n}(1-x_i)\max(0, E'_{x_i}(0) - E'_{x_i}(1))$$

$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} E'_{x_i,x_j}(0,1)(x_i - x_j)^2 = E(\mathbf{x}) + \text{const}.$$

Therefore, the cost of the cut is equal to the energy of the corresponding assignment, up to constant.

22.2.
a. Let $\mathcal{P}_\alpha$ be the set of x_is such that $x_i = \alpha$, $\mathcal{P}_\beta$ be the set of x_is such that $x_i = \beta$. For $x_i \in \mathcal{P}_\alpha \cup \mathcal{P}_\beta$, define binary variable t_i such that $t_i = 0$ iff $x_i = \alpha$, and $t_i = 1$ iff $x_i = \beta$.
b. For each t_i, define energy function $E'(t_i, t_j) = E(x_i) + \sum_{x_j \notin \mathcal{P}_\alpha \cup \mathcal{P}_\beta} E_{x_i,x_j}(x_i, x_j)$. That is, the new energy function is the original energy function plus the sum of energies over all links to neighbors which are neither in the α nor in the β label. Then $E'(\mathbf{t}) = E(\mathbf{x}) + \text{const}$.
c. $E'_{uv}(1,0) + E'_{uv}(0,1) - E'_{uv}(0,0) - E'_{uv}(1,1) = E_{uv}(\beta,\alpha) + E_{uv}(\alpha,\beta) - E_{uv}(\alpha,\alpha) - E_{uv}(\beta,\beta) = 2E_{uv}(\alpha,\beta) \geq 0$, since E is semimetric.

22.3.
a. Since $\hat{x}$ is a local minimum in the α-expansion move space, $E(\hat{x}) \leq E(x')$. Since x^* is globally optimal, $E(x^*) \leq E(\hat{x})$.
b. Let I_α be the set of nodes and pairs of neighboring nodes contained inside V^α, B_α be the set of pairs of neighboring nodes on the boundary of V^α, and O_α be the set of nodes and pairs of neighboring nodes contained outside of V^α. We can decompose total energy into energies restricted to sets I_α, B_α, and O_α. We have:
1) $E_{O_\alpha}(x') = E_{O_\alpha}(\hat{x})$. This is obvious, by the defintion of x'.
2) $E_{I_\alpha}(x') = E_{I_\alpha}(x^*)$. This is obvious, by the defintion of x'.
3) $E_{B_\alpha}(x') \leq cE_{B_\alpha}(x^*)$. This holds because:

$$
\begin{aligned}
cE_{B_\alpha}(x^*) &= \sum_{(s,t)\in B_\alpha} c\epsilon_{st}(x_s^*, x_t^*) \geq \sum_{(s,t)\in B_\alpha} c \min_{s\neq t} \epsilon_{st}(\alpha,\beta) \\
&\geq \sum_{(s,t)\in B_\alpha} \frac{\max_{s\neq t}\epsilon_{st}(\alpha,\beta)}{\min_{s\neq t}\epsilon_{st}(\alpha,\beta)} \min_{s\neq t} \epsilon_{st}(\alpha,\beta) = \sum_{(s,t)\in B_\alpha} \max_{s\neq t} \epsilon_{st}(\alpha,\beta) \\
&\geq \max_x E_{B_\alpha}(x) \geq E_{B_\alpha}(x').
\end{aligned}
$$

Since $E(\hat{x}) \leq E(x')$, we have:

$$E_{O_\alpha}(\hat{x}) + E_{B_\alpha}(\hat{x}) + E_{I_\alpha}(\hat{x}) \leq E_{O_\alpha}(x') + E_{B_\alpha}(x') + E_{I_\alpha}(x')$$

Putting (1), (2) and (3) inside, we have:

$$E_{B_\alpha}(\hat{x}) + E_{I_\alpha}(\hat{x}) \leq cE_{B_\alpha}(x^*) + E_{I_\alpha}(x^*).$$

Now sum this over all α to get:

$$\sum_{\alpha}[E_{B_\alpha}(\hat{x}) + E_{I_\alpha}(\hat{x})] \leq \sum_{\alpha}[cE_{B_\alpha}(x^*) + E_{I_\alpha}(x^*)].$$

Let $B = \bigcup_\alpha B_\alpha$. For every $(s,t) \in B$, $\epsilon_{st}(\hat{x})$ appears twice on the left hand side, once in $E_{B_\alpha}(\hat{x})$ for $\alpha = x_s^*$ and once in $E_{B_\alpha}(\hat{x})$ for $\alpha = x_t^*$. Similarly, $\epsilon_{st}(x^*)$ appears 2c times on the right hand side. Combining these, we get:

$$E(\hat{x}) + E_B(\hat{x}) \leq E(x^*) + (2c-1)E_B(x^*) \leq 2cE(x^*).$$

Therefore, we get $E(\hat{x}) \leq 2cE(x^*)$.

22.4.

Minimization of $E_1 + E_2 + E_3$ is equivalent to the problem where each parts are copied and we add the constraint that the two copies of the same variable should take the same value. Formally:

$$\min_{\mathbf{x},\mathbf{y}_0,\mathbf{y}_1} [E_1(\mathbf{x}) + E_2(\mathbf{y_0}) + E_3(\mathbf{x}, \mathbf{y}_1)]$$

such that $\mathbf{y}_0 = \mathbf{y}_1$. The lagrange dual is formed by relaxing the coupling constraints on $\mathbf{y}_0$ and $\mathbf{y}_1$ by Lagrange multipliers,

$$g(\lambda) = \min_{\mathbf{x},\mathbf{y}_0,\mathbf{y}_1} [E_1(\mathbf{x}) + E_2(\mathbf{y_0}) + E_3(\mathbf{x}, \mathbf{y}_1) + \lambda(\mathbf{y}_0 - \mathbf{y}_1)]$$

For any value of λ, $g(\lambda)$ is a lower bound on the original energy function. Now we decompose the dual function to separate $\mathbf{y}_0$ and $\mathbf{y}_1$:

$$g(\lambda) \geq \tilde{g}(\lambda) = \min_{\mathbf{x},\mathbf{y}_1}(E_1(\mathbf{x}) + E_3(\mathbf{x}, \mathbf{y}_1) - \lambda \cdot \mathbf{y}_1) + \min_{\mathbf{y}_0}(E_2(\mathbf{y}_0) + \lambda \cdot \mathbf{y}_0)$$

Since the feasibility constraint is only imposed on $\mathbf{y}_0$, the first minimization problem, slave-0,

$$\min_{\mathbf{x},\mathbf{y}_1}(E_1(\mathbf{x}) + E_3(\mathbf{x}, \mathbf{y}_1) - \lambda \cdot \mathbf{y}_1)$$

becomes tractable. This only contains submodular potentials, so can be solved by graphcut.

The second minimization problem, slave-1,

$$\min_{\mathbf{y}_0}(E_2(\mathbf{y}_0) + \lambda \cdot \mathbf{y}_0)$$

is equivalent to K-label problem which has a simple tree structure and can be solved by max-product message passing algorithm.
Given the current value of λ, let $\bar{\mathbf{x}}(\lambda)$, $\bar{\mathbf{y}}_0(\lambda)$, and $\bar{\mathbf{y}}_1(\lambda)$ be the optimal solutions to two slave problems. The subgradient of the relaxed dual function at λ is given by:

$$\nabla \tilde{g}(\lambda) = \bar{\mathbf{y}}_0(\lambda) - \bar{\mathbf{y}}_1(\lambda).$$

which can be used to update λ by $\lambda \leftarrow \lambda + \gamma_t \nabla \tilde{g}(\lambda)$.

22.5.
Define the rectified truncated Gaussian as follows:

$$\mathcal{R}(x; \mu, \sigma^2, l, u) = \mathbf{1}_{[l,u]} \frac{\mathcal{N}(x; \mu, \sigma^2)}{\Phi(u; \mu, \sigma^2) - \Phi(l; \mu, \sigma^2)}$$

To compute its mean, we apply change of variables that transforms the distribution before truncation to a standard normal, as follows:

$$c = \frac{l - \mu}{\sigma}, d = \frac{u - \mu}{\sigma}$$

We note that:

$$\frac{\partial(\Phi(d) - \Phi(c))}{\partial \mu} = \int_l^u \frac{\partial \mathcal{N}(x; \mu, \sigma^2)}{\partial \mu} dx = \frac{\Phi(d) - \Phi(c)}{\sigma^2}(\mu_{\mathcal{R}} - \mu)$$

At the same time:

$$\frac{\partial(\Phi(d) - \Phi(c))}{\partial \mu} = \frac{\partial(\Phi(\mu; l, \sigma^2) - \Phi(\mu; u, \sigma^2))}{\partial \mu} = \mathcal{N}(\mu; l, \sigma^2) - \mathcal{N}(\mu; u, \sigma^2)$$

$$= \mathcal{N}(c; 0, 1) - \mathcal{N}(d; 0, 1).$$

Combining these, we get:

$$\mu_{\mathcal{R}} = \mu + \sigma \frac{\mathcal{N}(c) - \mathcal{N}(d)}{\Phi(d) - \Phi(c)}.$$

Now, return to the original problem.
If we have $y_g = +1$, the distribution is the rectified truncated Gaussian with $l = 0, u = \infty, c = -\frac{\mu^t_{h_g \to d_g}}{\sigma^t_{h_g \to d_g}}, d = \infty$:

$$\mu^t_g = \mu^t_{h_g \to d_g} + \sigma^t_{h_g \to d_g} \frac{\mathcal{N}(c)}{1 - \Phi(c)} = \mu^t_{h_g \to d_g} + \sigma^t_{h_g \to d_g} \frac{\mathcal{N}(-c)}{\Phi(-c)}$$

$$= \mu^t_{h_g \to d_g} + \sigma^t_{h_g \to d_g} \Psi(\frac{\mu^t_{h_g \to d_g}}{\sigma^t_{h_g \to d_g}}).$$

If we have $y_g = -1$, the distribution is the rectified truncated Gaussian with $l = -\infty, u = 0, c = -\infty, d = -\frac{\mu^t_{h_g \to d_g}}{\sigma^t_{h_g \to d_g}}$:

$$\mu^t_g = \mu^t_{h_g \to d_g} + \sigma^t_{h_g \to d_g} \frac{-\mathcal{N}(d)}{\Phi(d)} = \mu^t_{h_g \to d_g} - \sigma^t_{h_g \to d_g} \frac{\mathcal{N}(d)}{\Phi(d)}$$

$$= \mu^t_{h_g \to d_g} - \sigma^t_{h_g \to d_g} \Psi(-\frac{\mu^t_{h_g \to d_g}}{\sigma^t_{h_g \to d_g}}).$$

Combining these, we finally conclude:

$$\mu^t_g = \mu^t_{h_g \to d_g} + y_g \sigma^t_{h_g \to d_g} \Psi(\frac{y_g \mu^t_{h_g \to d_g}}{\sigma^t_{h_g \to d_g}}).$$