A DERIVATION OF $\beta(\mathbf{x})$

Equation 5 introduces $\beta(\mathbf{x})$ in order to allow the single optimization problem approximating $\log Z$,

$$\max_{i=1}^{k} \max_{\mathbf{x}\in S_i} \theta\phi(\mathbf{x}) - \log\gamma(\mathbf{x}, q_i)$$

to be written in the form

$$\theta\phi(\mathbf{x}) + \beta(\mathbf{x}) \; \forall \, \mathbf{x} \in C$$

This reformulation is needed because, more generally, the cutting-planes technique requires that the lower-bound of $\log Z$ be written in the form

$$\log Z(\theta) \geq \max_{\mathbf{x}\in C} f(\theta, g(\mathbf{x}))$$

i.e. as the maximization over a set C of variable configurations $\mathbf{x}$ of a term which is linear in the parameter vector θ and which contains some (possibly nonlinear) function of $\mathbf{x}$ (the term must be linear in θ in order for Equation 5 to be a linear program). If the lower bound is expressed in such a form, it can then be equivalently represented by linear constraints of the form

$$\alpha \geq f(\theta, g(\mathbf{x})) \; \forall \, \mathbf{x} \in C$$

This section completes the derivation of Equation 5 from Equation 4 by showing how Equation 4 can be written as a maximization over configurations $\mathbf{x}$.

Proposition 3. *There exists a set C and function $\beta(\mathbf{x})$ such that* $\log Z \geq \theta\phi(\mathbf{x}) + \beta(\mathbf{x}) \; \forall \, \mathbf{x} \in C$.

Proof. From Equation 4,

$$\log Z(\theta) \geq \max_{i=1}^{k} \max_{\mathbf{x}\in S_i} (\theta\phi(\mathbf{x}) - \log\gamma(\mathbf{x}, q_i))$$

$$= \max_{\mathbf{x}\in\bigcup_{i=1}^{k} S_i} (\theta\phi(\mathbf{x}) + \max_{i|\mathbf{x}\in S_i} (-\log\gamma(\mathbf{x}, q_i)))$$

$$= \max_{\mathbf{x}\in C} (\theta\phi(\mathbf{x}) + \beta(\mathbf{x}))$$

□

Where $\beta(\mathbf{x}) = \beta(\mathbf{x}, q_1, \ldots, q_k) = \max_{i|\mathbf{x}\in S_i} \log\gamma(\mathbf{x}, q_i)$ and C is the union of all $\mathbf{x}$ in each sampled set $S_i \sim q_i$. Intuitively, given any configuration of variables $\mathbf{x}$, $\beta(\mathbf{x})$ represents the maximum scale factor (importance weight) of $\mathbf{x}$ for all set-proposal distributions q_i. For multiple $S_i^t \sim q_i, t = 1, \ldots, T$, it is necessary once again that

$$\log Z(\theta) \geq \max_{i=1}^{k} \operatorname*{median}_{t=1,\ldots,T} \max_{\mathbf{x}\in S_i^t} (\theta\phi(\mathbf{x}) - \log\gamma(\mathbf{x}, q_i))$$

be written in the form

$$\theta\phi(\mathbf{x}) + \beta(\mathbf{x}) \; \forall \, \mathbf{x} \in C$$

Taking the same approach,

$$\log Z(\theta) \geq \max_{i=1}^{k} \operatorname*{median}_{t=1,\ldots,T} \max_{\mathbf{x}\in S_i^t} (\theta\phi(\mathbf{x}) - \log\gamma(\mathbf{x}, q_i))$$

$$= \max_{\mathbf{x}\in\bigcup_{i=1}^{k}\bigcup_{t=1}^{T} S_i^t} (\theta\phi(\mathbf{x}) + \max_{i|\mathbf{x}\in\bigcup_{t=1}^{T} S_i^t} \operatorname*{median}_{t=1,\ldots,T} (-\log\gamma(\mathbf{x}, q_i))$$

In practice it also works well to replace the median with the max, as Corollary 3 proves an approximate lower bound and the bound is made tighter by taking the max over T samples. Making this substitution,

$$\log Z(\theta) \geq \max_{i=1}^{k} \max_{t=1}^{T} \max_{\mathbf{x}\in S_i^t} (\theta\phi(\mathbf{x}) - \log\gamma(\mathbf{x}, q_i))$$

$$= \max_{\mathbf{x}\in\bigcup_{i=1}^{k}\bigcup_{t=1}^{T} S_i^t} (\theta\phi(\mathbf{x}) + \max_{i|\mathbf{x}\in\bigcup_{t=1}^{T} S_i^t} (-\log\gamma(\mathbf{x}, q_i)))$$

$$= \max_{\mathbf{x}\in C} (\theta\phi(\mathbf{x}) + \beta(\mathbf{x}))$$