Proofs for "Information Geometry and Minimum Description Length Networks"

An Approximation of $\ln N(\mathcal{B}, \boldsymbol{\alpha})$

As the value of $\ln N(\mathcal{B}, \boldsymbol{\alpha})$ does not depend on the choice of the coordinate system, we abuse notation and vary $\boldsymbol{\eta} = \left(\eta^{(1)}, \ldots, \eta^{(\dim \mathcal{S})}\right)$ to an *ideal coordinate system*, where $g(\boldsymbol{\eta})$ is everywhere identity. In theory, this is possible locally. However, for convenience, we assume that such a coordinate system exists globally. By definition,

$$\begin{aligned}
\ln N(\mathcal{B}, \boldsymbol{\alpha}) &= \ln \left(\int_{\boldsymbol{\eta} \in \mathcal{S}} \sum_{i=1}^{m} \alpha_i \exp\left(-D(\boldsymbol{\eta} \,\|\, \boldsymbol{\eta}_i)\right) d\boldsymbol{\eta} \right) \\
&= \ln \left(\sum_{i=1}^{m} \alpha_i \int_{\boldsymbol{\eta} \in \mathcal{S}} \exp\left(-D(\boldsymbol{\eta} \,\|\, \boldsymbol{\eta}_i)\right) d\boldsymbol{\eta} \right) \\
&\approx \ln \Bigg(\sum_{i=1}^{m} \alpha_i \int_{\eta^{(1)}} \cdots \int_{\eta^{(\dim \mathcal{S})}} \exp\left(-\frac{1}{2} (\boldsymbol{\eta} - \boldsymbol{\eta}_i)^T g(\boldsymbol{\eta}_i) (\boldsymbol{\eta} - \boldsymbol{\eta}_i) \right) \\
&\qquad\qquad \times \sqrt{|g(\boldsymbol{\eta})|} d\eta^{(1)} \cdots d\eta^{(\dim \mathcal{S})} \Bigg) \\
&= \ln \left(\sum_{i=1}^{m} \alpha_i \int_{\eta^{(1)}} \cdots \int_{\eta^{(\dim \mathcal{S})}} \exp\left(-\frac{1}{2} \|\boldsymbol{\eta} - \boldsymbol{\eta}_i\|_2^2 \right) d\eta^{(1)} \cdots d\eta^{(\dim \mathcal{S})} \right) \\
&\approx \ln \left(\sum_{i=1}^{m} \alpha_i \exp\left(\frac{\dim \mathcal{S}}{2} \ln(2\pi) \right) \right) = \frac{\dim \mathcal{S}}{2} \ln(2\pi).
\end{aligned} \tag{S.1}$$

The first "$\approx$" is by approximating D to a square distance, which is only accurate when $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_i$ are close enough. The second "$\approx$" is by relaxing the domain of the integration from $\mathcal{S}$ to $\Re^{\dim \mathcal{S}}$. This is a rough approximation for a general $\mathcal{S}$, to show the order of the term $\ln N(\mathcal{B}, \boldsymbol{\alpha})$, and to show its weak dependence to $\mathcal{B}$ and $\boldsymbol{\alpha}$. More accurate approximations based on specific choices of $\mathcal{S}$ can lead to better implementations of MDL networks and better criteria in accordance to MDL.

Proof of $E(\mathcal{N}, A) \leq \hat{E}(\mathcal{N}, A)$ (HARDN)

Proof. $\forall l$, $\forall i$,

$$\sum_{j=1}^{n_{l+1}} \alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right) \geq \max_j \left[\alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right)\right]. \quad \text{(S.2)}$$

As $-\ln(x)$ is monotonically decreasing,

$$\begin{aligned} E(\mathcal{N}, A) &= -\sum_{l=0}^{L-1} \sum_{i=1}^{n_l} \ln\left(\sum_{j=1}^{n_{l+1}} \alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right)\right) \\ &\leq -\sum_{l=0}^{L-1} \sum_{i=1}^{n_l} \ln \max_j \left[\alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right)\right] \\ &= -\sum_{l=0}^{L-1} \sum_{i=1}^{n_l} \max_j \left[\ln \alpha_{l+1,j} - D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right] \\ &= \sum_{l=0}^{L-1} \sum_{i=1}^{n_l} \min_j \left[-\ln \alpha_{l+1,j} + D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right] = \hat{E}(\mathcal{N}, A). \end{aligned} \quad \text{(S.3)}$$

□

Proof of $E(\mathcal{N}, A) \leq \bar{E}(\mathcal{N}, A, B)$ (SOFTN)

Proof. Because of the convexity of $-\ln(x)$,

$$\begin{aligned} E(\mathcal{N}, A) &= -\sum_{l=0}^{L-1} \sum_{i=1}^{n_l} \ln\left(\sum_{j=1}^{n_{l+1}} \alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right)\right) \\ &= -\sum_{l=0}^{L-1} \sum_{i=1}^{n_l} \ln\left(\sum_{j=1}^{n_{l+1}} \beta_{li}^j \cdot \frac{\alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right)}{\beta_{li}^j}\right) \\ &\leq \sum_{l=0}^{L-1} \sum_{i=1}^{n_l} \sum_{j=1}^{n_{l+1}} \beta_{li}^j \left[-\ln\left(\frac{\alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right)}{\beta_{li}^j}\right)\right] \\ &= \sum_{l=0}^{L-1} \sum_{i=1}^{n_l} \sum_{j=1}^{n_{l+1}} \beta_{li}^j \left(\ln \frac{\beta_{li}^j}{\alpha_{l+1,j}} + D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right) = \bar{E}(\mathcal{N}, A, B). \end{aligned} \quad \text{(S.4)}$$

□

Proof of Theorem 3

Proof. Denote the true distribution with the components $\{\boldsymbol{\eta}_i^t\}$ and the weights $\{\alpha_i^t\}$ by $True(\boldsymbol{x})$. By eq. (7), $\forall \mathcal{N}$, $\forall A$, when $n \to \infty$,

$$\begin{aligned}
E(\mathcal{N}, A) = & -n \int True(\boldsymbol{x}) \ln \left(\sum_{j=1}^{n_1} \alpha_{1j} \exp\left(-D(\boldsymbol{\eta}(\boldsymbol{x}) \,\|\, \boldsymbol{\eta}_{1j})\right) \right) d\boldsymbol{x} \\
& - \sum_{l=1}^{L-1} \sum_{i=1}^{n_l} \ln \left(\sum_{j=1}^{n_{l+1}} \alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right) \right) \\
= & -n \int True(\boldsymbol{x}) \ln \left(\sum_{j=1}^{n_1} \alpha_{1j} p\left(\boldsymbol{x} \,|\, \boldsymbol{\eta}_{1j}\right) \right) d\boldsymbol{x} + constant \\
& - \sum_{l=1}^{L-1} \sum_{i=1}^{n_l} \ln \left(\sum_{j=1}^{n_{l+1}} \alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right) \right).
\end{aligned} \tag{S.5}$$

We construct an MDL network $\mathcal{N}^t$, where $\mathcal{L}_1^t$ is given by $\{\boldsymbol{\eta}_{1i}^t = \boldsymbol{\eta}_i^t\}$ with the weights $\{\alpha_{1i}^t = \alpha_i^t\}$. The rest of the cells $\{\boldsymbol{\eta}_{li}^t\}$ in higher levels, including their weights $\{\alpha_{li}^t\}$ are given by the sub-optimal solution which minimizes the above eq. (S.5) with $\mathcal{L}_1^t$ and its weights fixed. Given that $\mathcal{L}_0$ is fixed by infinite samples corresponding to the truth, $\forall \mathcal{N}$, $\forall A$,

$$\begin{aligned}
E(\mathcal{N}, A) - E(\mathcal{N}^t, A^t) = & n \int True(\boldsymbol{x}) \ln \frac{True(\boldsymbol{x})}{\sum_{j=1}^{n_1} \alpha_{1j} p\left(\boldsymbol{x}_i \,|\, \boldsymbol{\eta}_{1j}\right)} d\boldsymbol{x} \\
& - \sum_{l=1}^{L-1} \sum_{i=1}^{n_l} \ln \left(\sum_{j=1}^{n_{l+1}} \alpha_{l+1,j} \exp\left(-D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j})\right) \right) \\
& + \sum_{l=1}^{L-1} \sum_{i=1}^{n_l} \ln \left(\sum_{j=1}^{n_{l+1}} \alpha_{l+1,j}^t \exp\left(-D(\boldsymbol{\eta}_{li}^t \,\|\, \boldsymbol{\eta}_{l+1,j}^t)\right) \right).
\end{aligned} \tag{S.6}$$

If $\{\boldsymbol{\eta}_{1j}\}$ in $\mathcal{N}$ or $\{\alpha_{1j}\}$ in A does not correspond to $True(\boldsymbol{x})$, the first term on the right-hand-side of eq. (S.6) will go to $+\infty$ as $n \to \infty$. The second term is always non-negative, because of the non-negativity of D. Because of the sub-optimality discussed earlier, the third term is lower-bounded, as in

$$\sum_{l=1}^{L-1} \sum_{i=1}^{n_l} \ln \left(\sum_{j=1}^{n_{l+1}} \alpha_{l+1,j}^t \exp\left(-D(\boldsymbol{\eta}_{li}^t \,\|\, \boldsymbol{\eta}_{l+1,j}^t)\right) \right) \geq -\sum_{i=1}^{n_1} D(\boldsymbol{\eta}_i^t \,\|\, \tilde{\boldsymbol{\eta}}), \tag{S.7}$$

where $\tilde{\boldsymbol{\eta}}$ can be any distribution, e.g., the right-handed Bregman centroid $\{\boldsymbol{\eta}_i^t\}$. The right-hand-side of eq. (S.7) is the negative cost of a simple structure (one cell in $\mathcal{L}_2$) to represent $\mathcal{L}_1^t$, which is upper-bounded by the sub-optimal negative cost on the left-hand-side. Integrating all the three terms on the right-hand-of eq. (S.6), $E(\mathcal{N}, A) > E(\mathcal{N}^t, A^t)$. Hence, in the optimal solution, $\mathcal{L}_1$ must be exactly $\{\boldsymbol{\eta}_i^t\}$ and the weights must be exactly $\{\alpha_i^t\}$. □

Proof of Theorem 4

Proof. By the definition of $D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2)$ in section 2.3 as a Bregman divergence, $\forall \boldsymbol{\theta}(\boldsymbol{\eta})$, we have

$$\begin{aligned} gain(\boldsymbol{\eta}) = \quad & D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2) - D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}) - D(\boldsymbol{\eta} \,\|\, \boldsymbol{\eta}_2) \\ = + & \Big(\psi^\star(\boldsymbol{\eta}_1) - \psi^\star(\boldsymbol{\eta}_2) - \boldsymbol{\theta}_2^T (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) \Big) \\ - & \Big(\psi^\star(\boldsymbol{\eta}_1) - \psi^\star(\boldsymbol{\eta}) - \boldsymbol{\theta}^T (\boldsymbol{\eta}_1 - \boldsymbol{\eta}) \Big) \\ - & \Big(\psi^\star(\boldsymbol{\eta}) - \psi^\star(\boldsymbol{\eta}_2) - \boldsymbol{\theta}_2^T (\boldsymbol{\eta} - \boldsymbol{\eta}_2) \Big) \\ = & (\boldsymbol{\theta}_2 - \boldsymbol{\theta})^T (\boldsymbol{\eta} - \boldsymbol{\eta}_1). \end{aligned} \tag{S.8}$$

Let $\boldsymbol{\theta}_{lc} = (\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)/2$ be the left-handed Bregman centroid of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, then $\boldsymbol{\theta}_2 - \boldsymbol{\theta}_{lc} = \boldsymbol{\theta}_{lc} - \boldsymbol{\theta}_1$. Therefore,

$$gain(\boldsymbol{\eta}_{lc}) = (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_{lc})^T (\boldsymbol{\eta}_{lc} - \boldsymbol{\eta}_1) = (\boldsymbol{\theta}_{lc} - \boldsymbol{\theta}_1)^T (\boldsymbol{\eta}_{lc} - \boldsymbol{\eta}_1). \tag{S.9}$$

On the other hand, $\forall \boldsymbol{\eta}_a, \boldsymbol{\eta}_b \in \mathcal{S}$, $\boldsymbol{\eta}_a \neq \boldsymbol{\eta}_b$,

$$\begin{aligned} D(\boldsymbol{\eta}_a \,\|\, \boldsymbol{\eta}_b) + D(\boldsymbol{\eta}_b \,\|\, \boldsymbol{\eta}_a) = + & \Big(\psi^\star(\boldsymbol{\eta}_a) - \psi^\star(\boldsymbol{\eta}_b) - \boldsymbol{\theta}_b^T (\boldsymbol{\eta}_a - \boldsymbol{\eta}_b) \Big) \\ + & \Big(\psi^\star(\boldsymbol{\eta}_b) - \psi^\star(\boldsymbol{\eta}_a) - \boldsymbol{\theta}_a^T (\boldsymbol{\eta}_b - \boldsymbol{\eta}_a) \Big) \\ = & (\boldsymbol{\theta}_a - \boldsymbol{\theta}_b)^T (\boldsymbol{\eta}_a - \boldsymbol{\eta}_b) > 0. \end{aligned} \tag{S.10}$$

By eqs. (S.9) and (S.10),

$$gain(\boldsymbol{\eta}_{lc}) = D(\boldsymbol{\eta}_{lc} \,\|\, \boldsymbol{\eta}_1) + D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_{lc}) > 0 \quad \text{(which proves ①)}. \tag{S.11}$$

Similarly, we let $\boldsymbol{\eta}_{rc} = (\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)/2$ be the righted-handed Bregman centroid, then

$$\begin{aligned} gain(\boldsymbol{\eta}_{rc}) &= (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_{rc})^T (\boldsymbol{\eta}_{rc} - \boldsymbol{\eta}_1) = (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_{rc})^T (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_{rc}) \\ &= D(\boldsymbol{\eta}_2 \,\|\, \boldsymbol{\eta}_{rc}) + D(\boldsymbol{\eta}_{rc} \,\|\, \boldsymbol{\eta}_2). \end{aligned} \tag{S.12}$$

By eqs. (S.11) and (S.12), $\exists \boldsymbol{\eta} \in \mathcal{S}$ satisfying

$$gain(\boldsymbol{\eta}) \geq \max\{D(\boldsymbol{\eta}_{lc} \,\|\, \boldsymbol{\eta}_1) + D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_{lc}), D(\boldsymbol{\eta}_2 \,\|\, \boldsymbol{\eta}_{rc}) + D(\boldsymbol{\eta}_{rc} \,\|\, \boldsymbol{\eta}_2)\}. \tag{S.13}$$

□