Proof of Theorem

I. The conversion form (1) to (2)

$$
\begin{aligned}
MI(\mathbf{g},\mathbf{s}) &= H(\mathbf{g}) - H(\mathbf{g}|\mathbf{s}) \\
&= -\int_{\mathfrak{g}} p(g)\log p(g)\mathrm{d}g + \int_{\mathfrak{g}}\int_{\mathfrak{s}} p(g,s)\log p(g|s)\mathrm{d}s\mathrm{d}g \\
&= \int_{\mathfrak{g}}\int_{\mathfrak{s}} p(g,s)(-\log p(g) + \log p(g|s))\mathrm{d}s\mathrm{d}g \\
&= \int_{\mathfrak{g}}\int_{\mathfrak{s}} p(g,s)\log\frac{p(g,s)}{p(g)p(s)}\mathrm{d}s\mathrm{d}g \\
&= KL(\mathbb{P}_{\mathbf{gs}}||\mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}),
\end{aligned} \tag{1}
$$

where $MI(\cdot)$ is the calculation function of MI, $H(\cdot)$ is the shannon entropy. $\mathbf{g}$ and $\mathbf{s}$ are random variables, $\mathfrak{g}$ and $\mathfrak{s}$ are the ranges of variables, and g and s are the random point of the ranges. $KL(\cdot)$ is the KL-divergence, $p(\cdot)$ is probability density function. $\mathbb{P}_{\mathbf{gs}}$ and $\mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}$ are the joint distribution and product of marginal distributions.

II. The proof of the Donsker-Varadhan representation

Theorem 1: Donsker-Varadhan representation.

$$
\begin{aligned}
KL(\mathbb{P}_{\mathbf{gs}}||\mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}) = \sup_{\Gamma:\Omega\to\mathbb{R}} & \mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}(\Gamma) \\
& - \log(\mathbb{E}_{\mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}}(\mathrm{e}^{\Gamma})),
\end{aligned} \tag{2}
$$

where Ω is the cartesian space $\mathfrak{g}\times\mathfrak{s}$, Γ is the set of all functions that map $\mathbf{g}$ and $\mathbf{s}$ to a real number. $\mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}(\Gamma)$ is the mean of Γ on $\mathbb{P}_{\mathbf{gs}}$, $\mathbb{E}_{\mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}}(\mathrm{e}^{\Gamma})$ is the mean of e^{Γ} on $\mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}$.

Proof 2.1: We first define the difference between both sides of (2) as:

$$
\begin{aligned}
\Delta = & KL(\mathbb{P}_{\mathbf{gs}}||\mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}) - \\
& \sup_{\Gamma:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}(\Gamma) - \log(\mathbb{E}_{\mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}}(\mathrm{e}^{\Gamma})).
\end{aligned} \tag{3}
$$

We further prove (2) by proving $\Delta \geq 0$. To achieve this goal, for any Γ, we define a Gibbs density $\mathbb{G}$ that satisfies $\mathrm{d}\mathbb{G} = \frac{1}{Z}\mathrm{e}^{\Gamma}\mathrm{d}\mathbb{Q}$, where $\mathbb{Q} = \mathbb{P}_{\mathbf{g}}\otimes\mathbb{P}_{\mathbf{s}}$ and $Z = \mathbb{E}_{\mathbb{Q}}[\mathrm{e}^{\Gamma}]$. We can get the following form:

$$
\begin{aligned}
\mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}[\Gamma] - \log\mathbb{E}_{\mathbb{Q}}[\mathrm{e}^{\Gamma}] &= \mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}[\Gamma] - \log Z \\
&= \mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}[\log\frac{\mathrm{d}\mathbb{G}}{\mathrm{d}\mathbb{Q}}].
\end{aligned} \tag{4}
$$

The following formula can be obtained by embedding the (4) into (3):

$$
\begin{aligned}
\Delta &= \mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}[\log\frac{\mathrm{d}\mathbb{P}_{\mathbf{gs}}}{\mathrm{d}\mathbb{Q}}] - \mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}[\log\frac{\mathrm{d}\mathbb{G}}{\mathrm{d}\mathbb{Q}}] \\
&= \mathbb{E}_{\mathbb{P}_{\mathbf{gs}}}\log[\frac{\mathrm{d}\mathbb{P}_{\mathbf{gs}}}{\mathrm{d}\mathbb{Q}}\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{G}}] \\
&= KL(\mathbb{P}_{\mathbf{gs}}||\mathbb{G}).
\end{aligned} \tag{5}
$$

Since the KullbackLeibler-divergence is non-negative, we can find that $\Delta \geq 0$. Therefore, (2) holds.

III. The solution for (9)

The constrained optimization problem of the MGB is:

$$
\begin{aligned}
\mathbf{g}^* = & \min_{\mathbf{g}}\frac{1}{2}||\mathbf{g}_{ASR} - \mathbf{g}||_2^2 \\
& s.t. \quad \tilde{\mathbf{G}}^{\mathrm{T}}\mathbf{g} \geq 0,
\end{aligned} \tag{6}
$$

where $\mathbf{g}_{ASR} = \frac{\partial\mathcal{L}_{ASR}}{\partial\boldsymbol{\theta}}$ is the ASR gradient, $\mathbf{g}^*$ is the balanced gradient, and $\tilde{\mathbf{G}} = [\frac{\partial\mathcal{L}_0}{\partial\boldsymbol{\theta}}, \frac{\partial\mathcal{L}_1}{\partial\boldsymbol{\theta}}, \ldots, \frac{\partial\mathcal{L}_{N-1}}{\partial\boldsymbol{\theta}}]$ is the set of the regularization gradients.

We can obtain the lagrangian function of (6), which is shown as follows:

$$
L(\mathbf{g},\boldsymbol{\gamma}) = \frac{1}{2}\mathbf{g}^{\mathrm{T}}\mathbf{g} - \mathbf{g}_{ASR}^{\mathrm{T}}\mathbf{g} - \boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}}\mathbf{g}, \tag{7}
$$

where $L(\mathbf{g},\boldsymbol{\gamma})$ is the lagrangian function and $\boldsymbol{\gamma}$ is the lagrange multiplier.

The lagrangian dual form of (6) is shown as follows:

$$
\begin{aligned}
f(\boldsymbol{\gamma}) = & \min_{\mathbf{g}} L(\mathbf{g},\boldsymbol{\gamma}) \\
& s.t. \quad \boldsymbol{\gamma} \geq 0,
\end{aligned} \tag{8}
$$

where $f(\boldsymbol{\gamma})$ is the lagrangian dual function.

We next find $\mathbf{g}^*$ that minimizes (7) by setting the derivative of (7) to zero:

$$
\begin{aligned}
\frac{L(\partial\mathbf{g},\boldsymbol{\gamma})}{\partial\mathbf{g}} &= \mathbf{g}^{\mathrm{T}} - \mathbf{g}_{ASR}^{\mathrm{T}} - \boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}} = 0, \\
&\Longrightarrow \mathbf{g}^* = \mathbf{g}_{ASR} + \tilde{\mathbf{G}}\boldsymbol{\gamma}^{\mathrm{T}}.
\end{aligned} \tag{9}
$$

We further put $\mathbf{g}^*$ in (8), and it can be written as:

$$
\begin{aligned}
f(\boldsymbol{\gamma}) = & \frac{1}{2}(\mathbf{g}_{ASR}^{\mathrm{T}}\mathbf{g}_{ASR} + 2\boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}}\mathbf{g}_{ASR} + \boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}}\tilde{\mathbf{G}}\boldsymbol{\gamma}^{\mathrm{T}}) \\
& - \mathbf{g}_{ASR}^{\mathrm{T}}\mathbf{g}_{ASR} - 2\boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}}\mathbf{g}_{ASR} - \boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}}\tilde{\mathbf{G}}\boldsymbol{\gamma}^{\mathrm{T}}) \\
= & -\frac{1}{2}\mathbf{g}_{ASR}^{\mathrm{T}}\mathbf{g}_{ASR} - \boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}}\mathbf{g}_{ASR} - \frac{1}{2}\boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}}\tilde{\mathbf{G}}\boldsymbol{\gamma}^{\mathrm{T}}
\end{aligned} \tag{10}
$$

The solution $\boldsymbol{\gamma}^* = \max_{\boldsymbol{\gamma};\boldsymbol{\gamma}>0} f(\boldsymbol{\gamma})$ to the dual form is obtained by:

$$
\begin{aligned}
\frac{\partial f(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}} &= -\mathbf{g}_{ASR}^{\mathrm{T}}\tilde{\mathbf{G}} - \boldsymbol{\gamma}\tilde{\mathbf{G}}^{\mathrm{T}}\tilde{\mathbf{G}} = 0, \\
&\Longrightarrow \boldsymbol{\gamma}^* = -\frac{\mathbf{g}_{ASR}^{\mathrm{T}}\tilde{\mathbf{G}}}{\tilde{\mathbf{G}}(t)^{\mathrm{T}}\tilde{\mathbf{G}}}.
\end{aligned} \tag{11}
$$

Finally, because $\boldsymbol{\gamma}^*$ must be grater than zeros, we obtain the balancing gradient by putting $\boldsymbol{\gamma}^*$ in (9):

$$
\mathbf{g}^*(t) = \mathbf{g}_{ASR} - \tilde{\mathbf{G}}[\frac{\mathbf{g}_{ASR}^{\mathrm{T}}\tilde{\mathbf{G}}(t)}{\tilde{\mathbf{G}}(t)^{\mathrm{T}}\tilde{\mathbf{G}}(t)}]_-^{\mathrm{T}}. \tag{12}
$$