I. Appendix: Proofs of Theoretical Results in the Main Paper

Proposition 1. We consider the case when F_{opt} is unique, and $F_{opt}^{new} \neq F_{opt}$. Also, let us have that $I(F_{opt}^{new}(X_{new})+\epsilon; Y_{new}) = H(Y_{new})$. We then have

$$I(X_1; F_{opt}^{new}(X_1) + \epsilon) > I(X_1; F_{opt}(X_1) + \epsilon) \tag{1}$$

Proof. First we re-iterate the expressions for the optimal encoders IB-wise in both cases as follows.

$$F_{opt}^{new} = \arg\min_F \Big(I(X_{new}; \widehat{Z}_{new}) - \beta I(\widehat{Z}_{new}; Y_{new}) \Big) \tag{2}$$

$$F_{opt} = \arg\min_F \Big(I(X_1; \widehat{Z}) - \beta I(\widehat{Z}; Y_1) \Big) \tag{3}$$

As (X_1, Y_1) is contained within (X_{new}, Y_{new}) and $I(F_{opt}^{new}(X_{new}) + \epsilon; Y_{new}) = H(Y_{new})$ which is its maximum value as $I(X;Y) \leq H(Y)$, it immediately follows that $I(F_{opt}^{new}(X_1) + \epsilon; Y_1) = H(Y_1)$. From the uniqueness of F_{opt} we have

$$\begin{aligned}
I(X_1; F_{opt}^{new}(X_1) + \epsilon) - \beta I(F_{opt}^{new}(X_1) + \epsilon; Y_1) >& I(X_1; F_{opt}(X_1) + \epsilon) - \beta I(F_{opt}(X_1) + \epsilon; Y_1) && (4)\\
I(X_1; F_{opt}^{new}(X_1) + \epsilon) - \beta H(Y_1) >& I(X_1; F_{opt}(X_1) + \epsilon) - \beta I(F_{opt}(X_1) + \epsilon; Y_1) && (5)\\
I(X_1; F_{opt}^{new}(X_1) + \epsilon) >& I(X_1; F_{opt}(X_1) + \epsilon) + \beta(H(Y_1) - I(F_{opt}(X_1) + \epsilon; Y_1)) && (6)\\
>& I(X_1; F_{opt}(X_1) + \epsilon), && (7)
\end{aligned}$$

where the last step follows from the fact that $I(F_{opt}(X_1) + \epsilon; Y_1) \leq H(Y_1)$. This proves the result. □

Proposition 2. Consider the case when the noise ϵ comes from a bounded domain distribution, and there exists an unknown deterministic function f such that $f(X_1) = Y_1$, $\forall X \in \mathcal{X}$. In this case, when $\beta < 1$, we have,

$$I(X_1; F_{opt}(X_1) + \epsilon) = 0 \;\;\&\;\; I(X_1; F_{opt}^{new}(X_1) + \epsilon) = 0. \tag{8}$$

When $\beta > 1$, we have,

$$\begin{aligned}
I(X_1; F_{opt}(X) + \epsilon) &= H(Y_1)\\
\&\;\; I(X_1; F_{opt}^{new}(X_1) + \epsilon) &= H(Y_1) + H(Y_2)
\end{aligned} \tag{9}$$

Proof. The conditions mentioned in the proposition imply that $H(Y_1|X_1) = 0$, and as Y_1 is a function of X_1, we can write $I(X_1; \widehat{Z}) = I(X_1, Y_1; \widehat{Z}) = I(Y_1; \widehat{Z}) + I(X_1; \widehat{Z}|Y_1)$. Then the IB optimization (3) can be re-written as follows.

$$\begin{aligned}
\Big(I(X_1; \widehat{Z}) - \beta I(\widehat{Z}; Y_1) \Big) &= I(Y_1; \widehat{Z}) + I(X_1; \widehat{Z}|Y_1) - \beta I(\widehat{Z}; Y_1) && (10)\\
&= I(X_1; \widehat{Z}|Y_1) + (1-\beta) I(\widehat{Z}; Y_1) && (11)
\end{aligned}$$

When $\beta < 1$, the IB optimization becomes $I(X_1; \widehat{Z}|Y_1) + (\gamma) I(\widehat{Z}; Y_1)$, where $\gamma > 0$. Thus, the expression will reach the true minimum if both $I(X_1; \widehat{Z}|Y_1)$ and $I(\widehat{Z}; Y_1)$ can be minimized simultaneously to 0. This is possible just by setting F_{opt} such that $F_{opt}(X_1) = 0$. In that case, it follows that $I(X_1; F_{opt}(X_1) + \epsilon) = 0$ and $I(X_1; F_{opt}^{new}(X_1) + \epsilon) = 0$ (repeating the argument for (2)).

When $\beta > 1$, the IB optimization becomes $I(X_1; \widehat{Z}|Y_1) + (\gamma) I(\widehat{Z}; Y_1)$, where $\gamma < 0$. Thus, ideally it is minimized when $I(X_1; \widehat{Z}|Y_1) = 0$ and $I(\widehat{Z}; Y_1)$ attains its maximum value $H(Y_1)$. In that case, we obtain $I(X_1; F_{opt}(X_1) + \epsilon) = H(Y_1)$, and similarly, repeating the argument for (2) we have $I(X_1; F_{opt}^{new}(X_1) + \epsilon) = H(Y_1) + H(Y_2)$. This proves our results. □

Theorem 1. We are given the flow: $Y \to X \to Z \to \widehat{Z} \to \widehat{Y}$, where $\widehat{Z} = F(X) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. We also have $\widehat{Y} = D(F(X) + \epsilon)$, and we consider the minimization of the reconstruction aware loss function as shown in (**??**). In that context, we can show that the optimization

$$\min_{F, D, F_{rec}} \mathbb{E}_X \left[\mathcal{L}_{CE} \left(D(F(X_i) + \epsilon) + \lambda \mathcal{L}_{MSE}(X, \widehat{X}) \right) \right] \tag{12}$$

is equivalent to

$$\max_F I_{\mathcal{V}_1}(\widehat{Z} \to X) + \beta I_{\mathcal{V}_2}(\widehat{Z} \to Y) \tag{13}$$

for some choice of function classes $\mathcal{V}_1$ and $\mathcal{V}_2$, which depend on the structure of D and F, and for some scalar $\beta > 0$.

Remark 1. Note that there is a typo in the main paper regards to the $\mathcal{V}$-information bottleneck. Instead of $I_{\mathcal{V}}(X \to \widehat{Z})$ it should be $I_{\mathcal{V}}(\widehat{Z} \to X)$. Also, the loss function objective also minimizes over F_{rec}, which is implied in the paper but for clarity we are including it in the minimization.

Proof. We describe the construction of $\mathcal{V}_1$ and $\mathcal{V}_2$. For $\mathcal{V}_1$, we design the function space using the reconstruction network F_{rec} as follows. Every function within $f_1 \in \mathcal{V}_1$ is such that:

$$f_1[Z](X) = \tau e^{-\frac{(F_{rec}(Z)-X)^2}{\sigma^2}}, \tag{14}$$

for some configuration of the reconstruction network F_{rec}, and where $\tau > 0$ is chosen such that $\int_X f_1[Z](X)dX = 1$. Note that then $f_1[Z]$ yields a probability distribution over X, and thus $\mathcal{V}_1$ satisfies all conditions to be used to measure $\mathcal{V}$-information as follows. We then note that $H_{\mathcal{V}_1}(X) = C_{\mathcal{V}_1}$ is a constant that only depends on the choice of the architecture for F_{rec} and the data distribution of X both of which aren't optimized in our loss, and thus

$$I_{\mathcal{V}_1}(\widehat{Z} \to X) = H_{\mathcal{V}_1}(X) - H_{\mathcal{V}_1}(X|Y) = C_{\mathcal{V}_1} - \inf_{f \in \mathcal{V}_1} \mathbb{E}_{(X,Y)\sim P_{XY}}[-\log f[X](Y)] \tag{15}$$

$$= C_{\mathcal{V}_1} - \inf_{f \in \mathcal{V}_1} \mathbb{E}_{(X,Y)\sim P_{XY}}[(F_{rec}(Z) - X)^2] = C_{\mathcal{V}_1} - \inf_{F_{rec}} \mathcal{L}_{MSE}(X, \widehat{X})] \tag{16}$$

As $H(X)$ is a constant, the above implies that minimizing $\mathcal{L}_{MSE}(X, \widehat{X})$ over F_{rec} and F is equivalent to maximizing $I_{\mathcal{V}_1}(\widehat{Z} \to X)$ over F. This is because $H_{\mathcal{V}_1}(X)$ only depends on the choice of architecture of F_{rec} and is independent of $\widehat{Z}$, and thus does not depend on D, F and F_{rec}. Next, for $\mathcal{V}_2$, we design the function space such that. Every $f_2 \in \mathcal{V}_2$ is such that $f_2[\widehat{Z}](Y)$ represents the softmax probabilities output from the decoder D for the label Y. Note that then $\mathcal{V}_2$ is a valid construction for estimating $\mathcal{V}$-information measures. Also, note that $H_{\mathcal{V}_2}(Y) = H(Y)$ for neural network D, as one can always set the biases such that they yield the exact prior probabilities of $P(Y)$. With this we see

$$I_{\mathcal{V}_2}(\widehat{Z} \to Y) = H_{\mathcal{V}_2}(Y) - H_{\mathcal{V}_1}(Y|Z) = H(Y) - \inf_{f \in \mathcal{V}_2} \mathbb{E}_{(X,Y)\sim P_{XY}}[-\log f[X](Y)] \tag{17}$$

$$= H(Y) - \tau' \inf_D \mathcal{L}_{CE}\left(D(F(X_i) + \epsilon)\right) \tag{18}$$

for some non-zero valued τ' that depends on the temperature of the softmax. As $H(Y)$ is a constant, the above implies that minimizing $\mathcal{L}_{CE}\left(D(F(X_i) + \epsilon)\right)$ over D and F is equivalent to maximizing $I_{\mathcal{V}_2}(\widehat{Z} \to Y)$ over F. As our loss objective optimizes F in addition to D and F_{rec} it follows that the optimization $\min_{F,D,F_{rec}} \mathbb{E}_X \left[\mathcal{L}_{CE}\left(D(F(X_i) + \epsilon) + \lambda \mathcal{L}_{MSE}(X, \widehat{X})\right)\right]$ is equivalent to $\max_F I_{\mathcal{V}_1}(\widehat{Z} \to X) + \beta I_{\mathcal{V}_2}(\widehat{Z} \to Y)$.

$\square$