A. Derivations

A.1. Derivation of InfoNCE, I_{NCE}

We start from Barber and Agakov's variational lower bound on MI (Barber & Agakov, 2003). $I(x;y)$ can be bounded as follows:

$$I(x;y) = \mathbb{E}_{p(x,y)} \log \frac{p(y|x)}{p(y)} \geq \mathbb{E}_{p(x,y)} \log \frac{q(y|x)}{p(y)}, \tag{11}$$

where q is an arbitrary distribution. We show that the InfoNCE bound (Oord et al., 2018) corresponds to a particular choice for the variational distribution q followed by the application of the Jensen inequality. Specifically, $q(y|x)$ is defined by independently sampling a set of examples $\{y_1, \ldots, y_K\}$ from a proposal distribution $\pi(y)$ and then choosing y from $\{y_1, \ldots, y_K\}$ in proportion to the importance weights $w_y = \frac{e^{\psi(x,y)}}{\sum_k e^{\psi(x,y_k)}}$, where ψ is a function that takes x and y and outputs a scalar. In the context of representation learning, ψ is usually a dot product between some representations of x and y, e.g. $f(x)^T f(y)$ (Oord et al., 2018). The unnormalized density of y given a specific set of samples $y_{2:K} = \{y_2, \ldots, y_K\}$ and x is:

$$q(y|x, y_{2:K}) = \pi(y) \cdot \frac{K \cdot e^{\psi(x,y)}}{e^{\psi(x,y)} + \sum_{k=2}^{K} e^{\psi(x,y_k)}}, \tag{12}$$

where we introduce a factor K which provides "normalization in expectation". By normalization in expectation, we mean that taking the expectation of $q(y|x, y_{2:K})$ with respect to resampling of the alternatives $y_{2:K}$ from $\pi(y)$ produces a normalized density (see Sec. A.1.1 for a derivation):

$$\bar{q}(y|x) = \mathbb{E}_{\pi(y_{2:K})}[q(y|x, y_{2:K})], \tag{13}$$

where $\pi(y_{2:K}) = \prod_{k=2}^{K} \pi(y_k)$. The InfoNCE bound (Oord et al., 2018) is then obtained by setting the proposal distribution as the marginal distribution, $\pi(y) \equiv p(y)$ and applying Jensen's inequality, giving:

$$\begin{aligned} I(x,y) &\geq \mathbb{E}_{p(x,y)} \log \frac{\mathbb{E}_{p(y_{2:K})} q(y|x, y_{2:K})}{p(y)} \geq \mathbb{E}_{p(x,y)} \left[\mathbb{E}_{p(y_{2:K})} \log \frac{p(y)\, K \cdot w_y}{p(y)} \right] \\ &= \mathbb{E}_{p(x,y)} \left[\mathbb{E}_{p(y_{2:K})} \log \frac{K \cdot e^{\psi(x,y)}}{e^{\psi(x,y)} + \sum_{k=2}^{K} e^{\psi(x,y_k)}} \right] \\ &= \mathbb{E}_{p(x,y_1)p(y_{2:K})} \left[\log \frac{e^{\psi(x,y)}}{\frac{1}{K} \sum_{k=1}^{K} e^{\psi(x,y_k)}} \right] = I_{NCE}(x;y|\psi, K) \leq \log K, \end{aligned} \tag{14}$$

where the second inequality was obtained using Jensen's inequality.

A.1.1. Derivation of normalized distribution

We follow Cremer et al. (2017) to show that $q(y|x) = \mathbb{E}_{y_{2:K}\sim\pi(y)}[q(y|x, y_{2:K})]$ is a normalized distribution:

$$
\begin{aligned}
\int_x q(y|x)\,dy &= \int_y \mathbb{E}_{y_{2:K}\sim\pi(y)} \left(\pi(y) \frac{e^{\psi(x,y)}}{\frac{1}{K}\left(\sum_{k=2}^{K} e^{\psi(x,y_k)} + e^{\psi(x,y)}\right)} \right) dy \\
&= \int_y \pi(y) \mathbb{E}_{y_{2:K}\sim\pi(y)} \left(\frac{e^{\psi(x,y)}}{\frac{1}{K}\left(\sum_{k=2}^{K} e^{\psi(x,y_k)} + e^{\psi(x,y)}\right)} \right) dy \\
&= \mathbb{E}_{\pi(y)} \mathbb{E}_{\pi(y_{2:K})} \left(\frac{e^{\psi(x,y)}}{\frac{1}{K}\left(\sum_{k=2}^{K} e^{\psi(x,y_k)} + e^{\psi(x,y)}\right)} \right) \\
&= \mathbb{E}_{\pi(y_{1:K})} \left(\frac{e^{\psi(x,y)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\psi(x,y_k)}} \right) \\
&= K \cdot \mathbb{E}_{\pi(y_{1:K})} \left(\frac{e^{\psi(x,y_1)}}{\sum_{k=1}^{K} e^{\psi(x,y_k)}} \right) \\
&= \sum_{i=1}^{K} \mathbb{E}_{\pi(y_{1:K})} \frac{e^{\psi(x,y_i)}}{\sum_{k=1}^{K} e^{\psi(x,y_k)}} \\
&= \mathbb{E}_{\pi(y_{1:K})} \frac{\sum_{i=1}^{K} e^{\psi(x,y_i)}}{\sum_{k=1}^{K} e^{\psi(x,y_k)}} = 1
\end{aligned}
\tag{15}
$$

A.2. Proof of Proposition 1

Proposition 1 (Conditional InfoNCE). *I_{CNCE} is a lower-bound on $I(x; y|x')$ and verifies the properties below:*

$$
I_{CNCE}(x; y|x', \phi, K) = \mathbb{E}\left[\log \frac{e^{\phi(x',x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\phi(x',x,y_k)}} \right], \tag{6}
$$

1. $I_{CNCE} \leq I(x; y|x')$.

2. $\phi^* = \arg\sup_\phi I_{CNCE} = \log \frac{p(y|x',x)}{p(y|x')} + c(x, x')$.

3. $\lim_{K\to\infty} I_{CNCE}(x; y|x', \phi^*, K) = I(x; y|x')$.

Proof. We begin with 1., the derivation is as follows:

$$
\begin{aligned}
I(x; y|x') = \mathbb{E}_{p(x',x,y)} \log \frac{p(y|x', x)}{p(y|x')} &\geq \mathbb{E}_{p(x',x,y)} \log \frac{\bar{q}(y|x', x)}{p(y|x')} && (16) \\
&= \mathbb{E}_{p(x',x,y)} \log \frac{\mathbb{E}_{p(y_{2:K}|x')} q(y|x', x, y_{2:K})}{p(y|x')} && (17) \\
&\geq \mathbb{E}_{p(x',x,y)} \mathbb{E}_{p(y_{2:K}|x')} \log \frac{p(y|x')\, K \cdot w_y}{p(y|x')} && (18) \\
&= \mathbb{E}_{p(x',x,y)} \mathbb{E}_{p(y_{2:K}|x')} \log \frac{K \cdot e^{\phi(x',x,y)}}{\sum_{k=1}^{K} e^{\phi(x',x,y_k)}} && (19) \\
&= \mathbb{E}_{p(x',x,y)} \mathbb{E}_{p(y_{2:K}|x')} \log \frac{e^{\phi(x',x,y)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\phi(x',x,y_k)}} && (20) \\
&= I_{CNCE}(x; y|x', \phi, K), && (21)
\end{aligned}
$$

where in Eq. 18 we used Jensen's inequality and $p(y|x')$ as our proposal distribution for the variational approximation $\bar{q}(y|x', x)$.

For 2., we rewrite I_{CNCE} by grouping the expectation w.r.t x':

$$\mathbb{E}_{p(x')}\left[\mathbb{E}_{p(x,y_1|x')p(y_{2:K}|x')}\left[\log \frac{e^{\psi(x',x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\psi(x',x,y_k)}}\right]\right]. \tag{22}$$

Given that both distributions in the inner-most expectation condition on the same x', this term has the same form as I_{NCE} and therefore the optimal solution is $\phi^*_{x'} = \log \frac{p(y|x,x')}{p(y|x')} + c_{x'}(x)$ (Ma & Collins, 2018). The optimal ϕ for I_{CNCE} is thus obtained by choosing $\phi(x',x,y) = \phi^*_{x'}$ for each x', giving $\phi^* = \log \frac{p(y|x,x')}{p(y|x')} + c(x,x')$.

For proving 3., we substitute the optimal critic and take the limit $K \to \infty$. We have:

$$\lim_{K\to\infty} \mathbb{E}_{p(x',x,y_1)p(y_{2:K}|x')}\left[\log \frac{\frac{p(y|x',x)}{p(y|x')}}{\frac{1}{K}\left(\frac{p(y_1|x',x)}{p(y_1|x')} + \sum_{k=2}^{K} \frac{p(y_k|x',x)}{p(y_k|x')}\right)}\right], \tag{23}$$

From the Strong Law of Large Numbers, we know that as $\frac{1}{K-1}\sum_{k=1}^{K-1} \frac{p(y_k|x',x)}{p(y_k|x')} \to \mathbb{E}_{p(y|x')} \frac{p(y|x',x)}{p(y|x')} = 1$, as $K \to \infty$ a.s., therefore (relabeling $y = y_1$):

$$I_{CNCE} \sim_{K\to\infty} \mathbb{E}_{p(x',x,y)}\left[\log \frac{\frac{p(y|x',x)}{p(y|x')}}{\frac{1}{K}\left(\frac{p(y|x',x)}{p(y|x')} + K - 1\right)}\right] \tag{24}$$

$$\sim_{K\to\infty} \mathbb{E}_{p(x',x,y)}\left[\log \frac{p(y|x',x)}{p(y|x')} + \log \frac{K}{\left(\frac{p(y|x',x)}{p(y|x')} + K - 1\right)}\right] \tag{25}$$

$$\sim_{K\to\infty} I(x,y|x'), \tag{26}$$

where the last equality is obtained by noting that the second term $\to 0$. □

A.3. Proof for Proposition 2

Proposition 2 (**Variational I_{CNCE}**). *For any variational approximation $q_\xi(y|x')$ in lieu of $p(y|x')$, with $p(\cdot|x') \ll q_\xi(\cdot|x')$ for any x', we have:*

$$I_{VAR}(x,y|x',\phi,\xi,K) = \tag{7}$$

$$\mathbb{E}\left[\log \frac{e^{\phi(x',x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\phi(x',x,y_k)}}\right] - \mathbb{E}\left[KL\left(p(y|x') \,\|\, q_\xi\right)\right],$$

1. $I_{VAR} \leq I(x;y|x')$.
2. *If* $q_\xi(y|x') = p(y|x')$, $I_{VAR} = I_{CNCE}$.
3. $\lim_{K\to\infty} \sup_\phi I_{VAR}(x;y|x',\phi,\xi,K) = I(x;y|x')$.

Proof. For 1., we proceed as follows:

$$\begin{aligned}
I(x;y|x') &\geq \mathbb{E}_{p(x,y)}\left[\log \frac{q(y|x',x)q_\xi(y|x')}{p(y|x')q_\xi(y|x')}\right] \\
&= \mathbb{E}_{p(x,y)}\left[\log \frac{q(y|x',x)}{q_\xi(y|x')}\right] - \mathbb{E}_{p(x)}\left[KL(p(y|x')\|q_\xi(y|x'))\right] \\
&\geq \mathbb{E}_{p(x,y_1)q_\xi(y_{2:K}|x')}\left[\log \frac{e^{\phi(x',x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\phi(x',x,y_k)}}\right] - \mathbb{E}_{p(x)}\left[KL(p(y|x') \,\|\, q_\xi(y|x'))\right], \\
&= I_{VAR}(x,y|x',\phi,\xi,K)
\end{aligned} \tag{27}$$

where the last step has been obtained as in Eq. 18.

Proving 2. is straightforward by noting that if $q_\xi = p$, $KL(p(y|x')||q_\xi(y|x')) = 0$ and the first term corresponds to I_{CNCE}.

Proving 3. goes as follows:

$$\sup_\phi \mathbb{E}_{p(x,x',y_1)q_\xi(y_{2:K}|x')}\left[\log \frac{e^{\phi(x',x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\phi(x',x,y_k)}}\right] - \mathbb{E}_{p(x')}\left[KL\left(p(y|x') \,\|\, q_\xi(y|x')\right)\right] \tag{28}$$

$$= E_{p(x',x,y_1)q_\xi(y_{2:K}|x')}\left[\log \frac{p(y_1|x',x)}{q_\xi(y_1|x')} - \log \frac{p(y_1|x')}{q_\xi(y_1|x')} - \log \frac{1}{K}\sum_{k=1}^{K} \frac{p(y_k|x,x')}{q_\xi(y_k|x')}\right] \tag{29}$$

$$= I(x,y|x') - E_{p(x',x,y_1)q_\xi(y_{2:K}|x')}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p(y_k|x,x')}{q_\xi(y_k|x')}\right] \tag{30}$$

$$\rightarrow_{K\rightarrow\infty} I(x,y|x'). \tag{31}$$

This is obtained by noting that (1) for any K and q_ξ, $\arg\sup_\phi I_{VAR} = \log \frac{p(y|x',x)}{q_\xi(y|x)} + c(x,x')$ (because the KL doesn't depend on ϕ) and (2) the second term in the last line goes to 0 for $K \rightarrow \infty$ (a straightforward application of the Strong Law of Large Numbers shows that for samples $y_{2:K}$ drawn from $q_\xi(y_{2:K}|x')$, we have: $\frac{1}{K}\sum_{k=2}^{K} \frac{p(y_k|x,x')}{q_\xi(y_k|x')} \rightarrow_{K\rightarrow\infty} 1$).

□

A.4. Proofs for I_{IS}

We will be using the following lemma.

Lemma 1. *For any x', x and y, and any sequence ϕ_K such that $||\phi_K - \phi||_\infty \rightarrow_{K\rightarrow\infty} 0$:*

$$\lim_{K\rightarrow\infty} \mathbb{E}_{p(y_{2:K})} \log \frac{Ke^{\phi_K(x',x,y)}}{e^{\phi_K(x',x,y)} + (K-1)\sum_{k=2}^{K} w_k e^{\phi_K(x',x,y_k)}} \tag{32}$$

$$= \lim_{K\rightarrow\infty} \mathbb{E}_{p(y_{2:K}|x')} \log \frac{Ke^{\phi(x',x,y)}}{e^{\phi(x',x,y)} + \sum_{k=2}^{K} e^{\phi(x',x,y_k)}}, \tag{33}$$

where $w_k = \frac{\exp \psi^*(x',y_k)}{\sum_{k=2}^{K} \exp \psi^*(x',y_k)}$, for $\psi^*(x',y_k) = \arg\sup_\psi I_{NCE}(x',y|\psi,K) = \log \frac{p(y_k|x')}{p(y_k)}$.

Proof. We see that almost surely, for $y_{2:K} \sim p(\cdot)$:

$$\sum_{k=2}^{K} w_k e^{\phi_K(x',x,y_k)} = \frac{\frac{1}{K-1}\sum_{k=2}^{K} \frac{p(y_k|x')}{p(y_k)} e^{\phi_K(x',x,y_k)}}{\frac{1}{K-1}\sum_{k=2}^{K} \frac{p(y_k|x')}{p(y_k)}} \rightarrow_{K\rightarrow\infty} \mathbb{E}_{p(y|x')} e^{\phi(x',x,y)}, \tag{34}$$

where we applied the Strong Law of Large Numbers to the denominator.

For the numerator, we write:

$$\frac{1}{K-1}\sum_{k=2}^{K} \frac{p(y_k|x')}{p(y_k)} e^{\phi_K(x',x,y_k)} = \frac{1}{K-1}\sum_{k=2}^{K} \frac{p(y_k|x')}{p(y_k)} e^{\phi(x',x,y_k)}$$
$$+ \frac{1}{K-1}\sum_{k=2}^{K} \frac{p(y_k|x')}{p(y_k)} (e^{\phi_K(x',x,y_k)} - e^{\phi(x',x,y_k)})$$

and note that the first term is the standard IS estimator using $p(y_k)$ as proposal distribution and tends to $\mathbb{E}_{p(y|x')} e^{\phi(x',x,y)}$ from the Strong Law of Large Numbers, while the second term goes to 0 as ϕ_K tends to ϕ uniformly.

This gives $\lim_{K\rightarrow\infty} \mathbb{E}_{p(y_{2:K})} \log \frac{Ke^{\phi_K(x',x,y)}}{e^{\phi_K(x',x,y)} + (K-1)\sum_{k=2}^{K} w_k e^{\phi_K(x',x,y_k)}} = \log \frac{e^{\phi(x',x,y)}}{\mathbb{E}_{p(y|x')} e^{\phi(x',x,y)}}$.

Following the same logic (without the importance-sampling) demonstrates that:

$$\lim_{K\rightarrow\infty} \mathbb{E}_{p(y_{2:K}|x')} \log \frac{Ke^{\phi(x',x,y)}}{e^{\phi(x',x,y)} + \sum_{k=2}^{K} e^{\phi(x',x,y_k)}} = \log \frac{e^{\phi(x',x,y)}}{\mathbb{E}_{p(y|x')} e^{\phi(x',x,y)}},$$

which concludes the proof. □

Proposition 3 (**Importance Sampled I_{CNCE}**). *Assuming $\psi^* = \arg\sup_\psi I_{\mathit{NCE}}(x', y)$ and $w_k = \frac{\exp \psi^*(x', y_k)}{\sum_{k=2}^{M} \exp \psi^*(x', y_m)}$, we have the following two properties, where:*

$$I_{\mathit{IS}}(x, y|x', \phi, K) = \\ \mathbb{E}\left[\log \frac{e^{\phi(x', x, y_1)}}{\frac{1}{K}(e^{\phi(x', x, y_1)} + (K-1)\sum_{k=2}^{K} w_k e^{\phi(x', x, y_k)})}\right], \tag{9}$$

1. *$\lim_{K\to\infty} \sup_\phi I_{\mathit{IS}}(x; y|x', \phi, K) = I(x; y|x')$,*

2. *$\lim_{K\to\infty} \arg\sup_\phi I_{\mathit{IS}} = \log \frac{p(y|x', x)}{p(y|x')} + c(x, x')$.*

Proof. By applying Lemma 1 with $\phi_K = \phi$, we know that for any ϕ:

$$\lim_{K\to\infty} I_{\mathit{IS}}(x; y|x', \phi, K) = \lim_{K\to\infty} \mathbb{E}_{p(x', x, y)p(y_{2:K}|x')} \log \frac{K e^{\phi(x', x, y)}}{e^{\phi(x', x, y)} + \sum_{k=2}^{K} e^{\phi(x', x, y_k)}}.$$

In particular, the RHS of the equality corresponds to $\lim_{K\to\infty} I_{\mathit{CNCE}}(x, y|x', \phi, K)$. That quantity is smaller than $I(x, y|x')$, with equality for $\phi = \phi^*$. This guarantees that:

$$\lim_{K\to\infty} \sup_\phi I_{\mathit{IS}}(x; y|x', \phi, K) \geq \lim_{K\to\infty} I_{\mathit{IS}}(x; y|x', \phi^*, K) = I(x, y|x'). \tag{35}$$

We now prove the reverse inequality. We let $2\epsilon = \lim_{K\to\infty} \sup_\phi I_{\mathit{IS}}(x; y|x', \phi, K) - I(x, y|x')$, and assume towards a contradiction that $\epsilon > 0$. We know that:

$$\exists K_0, \quad \forall K \geq K_0, \quad \sup_\phi I_{\mathit{IS}}(x; y|x', \phi, K) \geq I(x, y|x') + \epsilon.$$

Now, $\forall K \geq K_0$, let ϕ_K be such that:

$$I_{\mathit{IS}}(x; y|x', \phi_K, K) \geq \sup_\phi I_{\mathit{IS}}(x; y|x', \phi, K) - \frac{\epsilon}{2},$$

and thus: $\forall K \geq K_0, I_{\mathit{IS}}(x; y|x', \phi_K, K) \geq I(x, y|x') + \frac{\epsilon}{2}$.

Since $\phi_K \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|\times|\mathcal{Y}|}$, $\{\phi_K\}_{K \geq K_0}$ contains a subsequence that converges to a certain $\phi_\infty \in \bar{\mathbb{R}}^{|\mathcal{X}|\times|\mathcal{X}|\times|\mathcal{Y}|}$. Without loss of generality, we assume that $\forall K, \forall x', \forall x, \mathbb{E}_{p(y)}[\phi_K(x', x, y)] = 0$ which implies that $\mathbb{E}_{p(y)}[\phi_\infty(x', x, y)] = 0$ (similarly to I_{NCE}, I_{IS} is invariant to functions of (x', x) added to ϕ).

In particular, this guarantees that $||\phi_\infty||_\infty < \infty$. Otherwise, we would have $\phi_\infty(x', x, y) = -\infty$ for a given y, which would then imply $I_{\mathit{IS}}(x; y|x', \phi_\infty, K) = -\infty$ and give a contradiction.

We can now apply Lemma 1 to $\{\phi_K\}$ and ϕ_∞ to show that $\lim_{K\to\infty} I_{\mathit{IS}}(x; y|x', \phi_K, K) = \lim_{K\to\infty} I_{\mathit{CNCE}}(x, y|x', \phi_\infty, K)$, and get a contradiction: the first term is larger than $I(x, y|x') + \frac{\epsilon}{2}$ while the second is smaller than $I(x, y|x')$. □

A.5. Proof for I_{BO}

Proposition 4 (**Boosted Critic Estimation**). *Assuming $\psi^* = \arg\sup_\psi I_{\mathit{NCE}}(x', y)$, the following holds, with:*

$$I_{\mathit{BO}}(x, y|x', \phi, K) = \mathbb{E}\left[\log \frac{e^{\psi^*(x', y_1) + \phi(x', x, y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\psi^*(x', y_k) + \phi(x', x, y_k)}}\right], \tag{10}$$

1. *$I_{\mathit{BO}} \leq I(x, x'; y)$,*

2. *$\phi^* = \arg\sup_\phi I_{\mathit{BO}} = \log \frac{p(y|x', x)}{p(y|x')} + c(x, x')$.*

Proof. To prove 1., it suffices to follow the proof for I_{NCE} (Sec. A.1). To prove 2., we set $\eta(x', x, y) = \psi^*(x', y) + \phi(x', x, y_1)$. Ma & Collins (2018) show that $\eta^*(x', x, y) = \log \frac{p(y|x',x)}{p(y)} + c^\eta(x', x)$, for any K. Knowing that $\psi^*(x', y) = \log \frac{p(y|x')}{p(y)} + c^\psi(x')$ is a constant in the maximization problem, simple algebra shows that $\phi^*(x', x, y) = \log \frac{p(y|x',x)}{p(y|x')} + c(x', x)$. □

A.6. Synthetic Experiments

Here, we provide details for Sec. 5.1. In this experiment, each x, x' and y are 20-dimensional. For each dimension, we sampled $(x_i, {x'}_i, y_i)$ from a correlated Gaussian with mean 0 and covariance matrix $\texttt{cov}_i$. For a given value of MI, $\texttt{mi} = \{5, 10, 15, 20\}$, we sample covariance matrices $\texttt{cov}_i = \texttt{sample\_cov}(\texttt{mi}_i)$, such that $\sum_i \texttt{mi}_i = \texttt{mi}, \texttt{mi}_i > 0$ chosen at random. We optimize the bounds by stochastic gradient descent (Adam, learning rate $5 \cdot 10^{-4}$). All encoders f are multi-layer perceptrons with a single hidden layer and ReLU activation. Both hidden and output layer have size 100.

InfoNCE computes:

$$\mathbb{E}_p\left[\log \frac{e^{f([x,x'])^T f(y)}}{e^{f([x,x'])^T f(y)} + \sum_{k=2}^{K} e^{f([x,x'])^T f(y_k)}}\right] + \log K, \quad y_{2:K} \sim p(y),$$

where the proposal is the marginal distribution $p(y)$, E is chosen to be a dot product between representations, $\mathbb{E}_p$ denotes expectation w.r.t. the known joint distribution $p(x, x', y)$ and is approximated with Monte-Carlo, $[x, x']$ denotes concatenation and f is a 1-hidden layer MLP.

DEMI computes:

$$\begin{aligned}\mathbb{E}_{p(x',x,y)p(y_{2:K/2})}&\left[\log \frac{e^{f(x')^T f(y)}}{e^{f(x')^T f(y)} + \sum_{k=2}^{K/2} e^{f(x')^T f(y_k)}}\right] + \\ &\mathbb{E}_{p(x',x,y)p(y_{2:K/2}|x')}\left[\log \frac{e^{f([x',x])^T f(y)}}{e^{f([x',x])^T f(y)} + \sum_{k=2}^{K/2} e^{f([x',x])^T f(y_k)}}\right] + 2\log K/2\end{aligned} \tag{36}$$

where $f(x)$ is just $f([x, \mathbf{0}])$ in order to re-use MLP parameters for the two terms. The negative samples of the conditional MI term come from the conditional distribution $p(y|x')$, which is assumed to be known in this controlled setting. We maximize both lower bounds with respect to the encoder f. We report pseudo-code for `sample_cov` in Listing 2, used to generate 3×3 covariance matrices for a fixed $I(\{x, x'\}; y)$ and uniformly sampled $\alpha = I(x; y)/I(\{x, x'\}; y)$.

B. Experiments on Dialogue

B.1. DEMI Details

The optimization of the DEMI requires the specification of a critic. Following previous work (Oord et al., 2018; Hjelm et al., 2019), we implement the critic by a dot product between representations of the past $f(x)$ and those of the future $f(y)$. We obtain f_x, f_y by running a forward pass with the GPT2 model on the words from the past and the future separately and by taking the state of the last layer of the GPT2 corresponding to the last token in the past and the future respectively.

For all DEMI terms, given the past, the model is trained to pick the ground-truth future among a set of N future candidates. This candidate set includes the ground-truth future and $N - 1$ negative futures drawn from different proposal distributions. To compute $I_{NCE}(x; y)$, we consider the ground truth future of each sample in the batch as a negative candidate for the other samples in the same batch. Using this approach, the number of candidates N is equated to the batch size. This ensures that negative samples are sampled from the marginal distribution $p(y)$. To compute the conditional MI boud $I_{CNCE}(x; y|x')$, we sample negative futures $p(y|x')$ by conditioning the GPT2 model on the most recent utterance in the past x'.

B.2. Dataset

Wizard of Wikipedia (Dinan et al., 2019) consists of 20 365 dialogues where each dialogue in the conversation is about a specific topic. There are two participants in the conversation: the wizard and the apprentice. The apprentice is a curious learner who is eager to know more about a particular topic. However, the wizard is a knowledgeable expert who tries to inform the apprentice about the topic. In our experiments, we used the valid data "unseen valid" that includes topics that do not overlap with the train data and the test data. Detailed statistics of the dataset are presented in Table 4.

```
def sample_cov(mi):
  alpha = random.uniform(0.1, 0.9)
  params = random.normal(0, $\mathbb{I}_6$)
  # use black box optimizer (Nealder-Mead) to determine opt_params
  opt_param = $\arg\min_x$ residual(params,mi,$\alpha$)
  return project_posdef(opt_params)

def project_posdef(x):
  # project x $\in \mathbb{R}^6$ to a positive definite 3x3 matrix
  cov = zeros(3, 3)
  cov[tril_indices(3)] = x
  cov /= column_norm(cov)
  return dot(cov, cov.T)

def analytical_mi(cov):
  # compute analytical MI of 3 covariate Gaussian variables
  cov_01 = cov[:2, :2]
  cov_2 = cov[2:3, 2:3]
  mi_xp_xpp_y = 0.5 * (log(det(cov_01)) + log(det(cov_2)) - log(det(cov)))
  cov_1 = cov[1:2, 1:2]
  cov_23 = cov[1:, 1:]
  mi_xp_y = 0.5 * (log(det(cov_1)) + log(det(cov_2)) - log(det(cov_23)))
  return mi_xp_xpp_y, mi_xp_y

def residual(x, mi, $\alpha$):
  # penalize difference between analytical mi and target mi, $\alpha$mi
  cov = project_posdef(x)
  mi_xp_y, mi_xp_y = analytical_mi(cov)
  return (mi_xp_xpp_y - mi) ** 2 + (mi_xp_y - $\alpha$ * mi) ** 2
```

Listing 2: Pseudo-code for covariance sampling in the synthetic experiment.

Table 3: A sample dialogue between speaker A and speaker B from the Wizard of Wikipedia dataset. The four rows from top to bottom are: (1) x: the "past" dialogue up to utterance k (2) y: the ground-truth utterance for the next turn $k+1$ (3) $y_{1:N}$: future candidates sampled from the "restricted context" future distribution $p(y|x')$. These candidates correspond to the set of **hard** negatives that are closely related to the conversation. (4) $y'_{1:N}$: future candidates sampled randomly from the dataset. We can see that candidates $y_{1:N}$ are semantically close but incoherent w.r.t to the dialogue history as they were conditioned solely on the immediate past utterance x'. However, we can notice that candidates $y'_{1:N}$ are semantically distant from x as they were sampled randomly from the data distribution. The highlighted text in green correspond to the topic of the conversation. Speaker B mentions that they have never done either parachuting or skydiving. B_1 corresponds to the utterance generated based on the restricted context x'. The utterance is on-topic but completely contradictory to what speaker B has said in the past. On the other hand B'_1 is randomly sampled from other dialogues. We can observe that the utterance is clearly irrelevant to the conversation.

x	A:	I like parachuting or skydiving .
	$\boldsymbol{B}$:	I've never done either but they sound terrifying, not a fan of heights.
	A:	But it is interesting game. This first parachute jump in history was made by Andre Jacques.
	$\boldsymbol{B}$:	Oh really ? Sounds like a french name, what year did he do it ?
	A:	It done in October 22 1797. They tested his contraption by leaping from a hydrogen balloon.
	$\boldsymbol{B}$:	Was he successful or did he kick the bucket off that stunt?
	A:	I think its a success. The military developed parachuting tech.
$y \sim p(y\|x')$	B_{gt}	Yeah nowadays they are a lot more stable and well made.
$y_{1:N} \sim p(y\|x')$	B_1:	That is great. I've been skydiving for days now . How is it ?
	B_2:	Oh I have never flown but I'm glad to know.
	B_3:	I've been dying for it since I was a kid.
	B_4:	Yes, that is why NASA had an advanced mechanics tech for months.
	B_5:	I went parachuting last Sunday and enjoyed it.
$y'_{1:N} \sim p(y)$	B'_1:	I think science fiction is an amazing genre for anything
	B'_2:	Can you imagine the world without internet access ?
	B'_3:	I am just finishing my university course and I will be a qualified pharmacist.
	B'_4:	I don't know how to be romantic. I have trouble expressing emotional attraction.
	B'_5:	I think Krav Maga is a martial art sport. That 's the reason I picked it .

Table 4: Statistics of the Wizard of Wikipedia dataset

	# Train	# Valid	# Test
Number of utterances	166 787	8806	8782
Number of dialogues	18 430	967	968
Number of topics	1247	300	58
Average turns per dialog	9	9	9

Table 5: Results for perplexity, sequence-level metric, token-level metrics, BLEU and diversity metrics on the test data of the Wizard of Wikipedia dataset. Results demonstrate that the proposed InfoNCE and DEMI bounds achieve lower perplexity, reduce next-token repetition and increase the number of unique next-tokens compared to the baselines GPT2, GPT2-MMI and TransferTransfo. Note that our results are not directly comparable with Li et al. (2020) as their model is trained from scratch on a not publicly available Reddit-based corpus.

Model	ppl	seq-rep-avg	rep	wrep	uniq	dist-1	dist-2	BLEU	Entropy-4
GPT2	19.24	0.064	0.130	0.132	7393	0.064	0.392	0.775	0.095
TransferTransfo	19.33	0.078	0.134	0.132	7735	0.058	0.386	0.752	0.084
GPT2-MMI	19.35	0.070	0.129	0.135	7623	0.052	0.384	0.740	0.092
InfoNCE	18.88	0.065	0.126	0.131	8432	0.065	0.390	0.799	0.107
DEMI	**18.66**	**0.050**	**0.120**	**0.128**	**8666**	**0.070**	**0.405**	**0.810**	**0.108**
Ground Truth	–	0.052	0.095	–	9236	0.069	0.416	–	0.110

B.3. Experimental Setup

Given memory constraints, all the proposed models are trained with a batch size of 5 per GPU, considering up to three utterances for the future and five utterances in the past. All the models are trained on 2 NVIDIA V100s. The models early-stop in the 4th epoch. We use the Adam optimizer with a learning rate of 6.25×10^{-5}, which we linearly decay to zero during training. Dropout is set to 10% on all layers. InfoNCE/DEMI terms are weighted with a factor 0.1 in the loss function. We varied the factor from 0.1 to 1 and 0.1 was chosen based on the best results on the validation set. During inference, we use nucleus sampling (Holtzman et al., 2020) with $p = 0.9$ for all models.

B.4. Additional Automated metrics

Repetition The word repetition metrics aim at testing the model's performance in generating responses while avoiding artificial repetitions. We employ the repetition metrics presented in Welleck et al. (2020): **seq-rep-**n, **rep**, **wrep** and **uniq**. These metrics are defined based on the amount of repetitions in the generations. **seq-rep-**n measures the portion of duplicate n-grams in a generated sequence:

$$\textbf{seq-rep-}n = 1 - \frac{|\text{unique n-grams}(w_{1:N})|}{|\text{n-grams}|} \tag{37}$$

where $w_{1:N}$ is the generated utterance. We report **seq-rep-avg** which averages over $n \in \{2, 3, 4, 5, 6\}$. **rep** measures the fraction of tokens that occur in previous tokens, **uniq** counts the number of unique tokens on the validation set. Please refer to (Welleck et al., 2020; Li et al., 2020) for more information about these metrics.

Distinct-n The metric is derived from Li et al. (2016). It is defined as the number of unique n-grams, normalized by the total number of n-grams of tested sentences.

Entropy-n We employ the entropy metric from Zhang et al. (2018) which aims to fix the problem of frequency difference of n-grams in Distinct-n by reflecting how evenly the empirical n-gram distribution is for each given sentence.

Results on the test set and the valid set are presented in Table 5 and Table 6 respectively.

Table 6: Results for perplexity, sequence-level metric, token-level metrics, BLEU and diversity metrics on the valid data of the Wizard of Wikipedia dataset.

Model	**ppl**	**seq-rep-avg**	**rep**	**wrep**	**uniq**	**dist-1**	**dist-2**	**BLEU**	**Entropy-4**
GPT2	19.21	0.066	0.136	0.134	7259	0.083	0.432	0.780	0.106
TransferTransfo	19.32	0.074	0.134	0.133	7223	0.082	0.431	0.750	0.106
GPT2-MMI	19.30	0.065	0.126	0.130	7351	0.082	0.425	0.650	0.106
InfoNCE	18.85	0.065	0.126	0.126	7300	0.091	0.046	0.800	0.107
DEMI	**18.70**	**0.054**	**0.130**	**0.120**	**7375**	**0.077**	**0.047**	**0.820**	**0.108**
Ground Truth	–	0.052	0.095	–	9236	0.069	0.416	–	0.110

Table 7: Selected responses to the same context from different methods fine-tuned on the Wizard of Wikipedia datasets.

Context	**Generated Responses**
A: I recently took up archery and I think I love it . Do you know anything about Archery ? **B:** Archery is a sport / skill of using a bow to propel arrows. **A:** And a great sport it is . Do you know where archery originated from ?	**DEMI**: Archery originated from ancient Mesopotamia, and is actually practiced internationally in over 200 countries. **InfoNCE**: I am not sure but many traditions and principles arose during the Mesolithic and Paleolithic era. **TransferTransfo** : Yep , you just use it for skill and using it to shoot arrows. **GPT2**: I don’t know, but I know that the old French called it archer’s art.
A: What is your motivation? **B:** That’s a tough question. I ’d say my own progress. I like seeing how much I improve with something. You ? **A:** I am retired now. Are you familiar with the work of Mehr and Meyer, well known psychologists ? **B:** They said that " Motivation is a word that is part of the popular culture as few other psychological concepts are . " Do you agree with what they said ?	**DEMI**: Yes, definitely! You should check out Mark Twain’s famous motivational book ! There are theories that point out how important it is to meditate. **InfoNCE**: They said that in psychology research, theories and evidence, participants who enjoyed continuous experience could be more likely to be successful. **TransferTransfo**: Absolutely, I disagree. Are you an author or a commentator? **GPT2**: That’s awesome. So , what types of differences exist between “good " motivation and bad?

	DEMI_wins	DEMI_CI	baseline_wins	baseline_CI	pairwise_CI	p
Baseline						
GPT2	0.48726	(0.44, 0.53]	0.28662	(0.25, 0.32]	(0.13, 0.27] *	<0.001
GPT2-MMI	0.65833	(0.6, 0.71]	0.16250	(0.12, 0.21]	(0.4, 0.58] *	<0.001
TransferTransfo	0.46888	(0.43, 0.51]	0.30043	(0.26, 0.34]	(0.09, 0.24] *	<0.001
InfoNCE	0.41711	(0.38, 0.46]	0.36748	(0.33, 0.41]	(-0.03, 0.13]	0.0905
gold_response	0.22679	(0.19, 0.26]	0.54325	(0.5, 0.59]	(-0.39, -0.25] *	<0.001

Table 8: Which response is more *relevant*?

	DEMI_wins	DEMI_CI	baseline_wins	baseline_CI	pairwise_CI	p
Baseline						
GPT2	0.45084	(0.41, 0.49]	0.32636	(0.29, 0.37]	(0.05, 0.2] *	<0.001
GPT2-MMI	0.61734	(0.56, 0.67]	0.18393	(0.14, 0.23]	(0.34, 0.53] *	<0.001
TransferTransfo	0.43617	(0.4, 0.48]	0.35000	(0.31, 0.39]	(0.01, 0.16] *	0.0028
InfoNCE	0.44630	(0.41, 0.49]	0.34515	(0.31, 0.38]	(0.03, 0.17] *	<0.001
gold_response	0.22164	(0.19, 0.26]	0.56608	(0.52, 0.61]	(-0.41, -0.28] *	<0.001

Table 9: Which response is more *humanlike*?

	DEMI_wins	DEMI_CI	baseline_wins	baseline_CI	pairwise_CI	p
Baseline						
GPT2	0.56157	(0.52, 0.6]	0.21444	(0.18, 0.25]	(0.28, 0.42] *	<0.001
GPT2-MMI	0.68750	(0.63, 0.74]	0.12292	(0.09, 0.16]	(0.48, 0.65] *	<0.001
TransferTransfo	0.51931	(0.48, 0.56]	0.24571	(0.21, 0.28]	(0.21, 0.34] *	<0.001
InfoNCE	0.41288	(0.37, 0.45]	0.33580	(0.3, 0.38]	(0.0, 0.15] *	0.0059
gold_response	0.32384	(0.28, 0.36]	0.46624	(0.43, 0.51]	(-0.22, -0.07] *	<0.001

Table 10: Which response is more *interesting*?

B.5. Human Evaluation

We closely follow the protocol used in Zhang et al. (2020). Systems were paired and each response pair was presented to 3 judges in random order. Judges expressed their preference on a 3 point Likert scale. We use a majority vote for each response pair to decide whether a specific baseline, the pivot (DEMI), or neither, performed better. We then bootstrap the set of majority votes to obtain a 99% confidence interval (CI) on the expected difference between the baseline and DEMI. If this confidence interval contains 0, the difference is deemed insignificant. We also compute p-values from the confidence intervals[5].

In the following tables, the "pivot" is always the system given by DEMI. Pairings where the pairwise confidence interval is marked with "*" have a significant difference.

[5] https://www.bmj.com/content/343/bmj.d2304