A Proofs of Propositions

To simplify our notation, we remove subscripts such that X, A, Y denotes the transition X_t, A_t, X_{t+1}, and Z denotes the latent Z_{t+1}. Then the environment's dynamics is given by $\tau(Y|x,a)$, the agent's policy is given by $\pi(A|x)$, and the induced state visitation given by $\rho_\pi(X)$. The generator is denoted $p_\theta(Z|x,a,y)$, the reconstructor $f_\eta(x,a,z)$, and the critic $g_\nu(x,a,z)$. We start with several lemmas that will be useful, the first being a pointwise version of Barber and Agakov's variational lower bound:

Lemma 4 (Pointwise Barber-Agakov) Denote the pointwise mutual information:

$$\mathrm{PMI}_\theta(x,a;z) \coloneqq \log \frac{p_\theta(z|x,a)}{p_\theta(z)} \tag{20}$$

Then for any variational distribution q:

$$\mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x,a;Z) \geq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x,a)}{p_\theta(Z)} \tag{21}$$

Proof. Starting from the left hand side:

$$\begin{aligned}
\mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x,a;Z) &= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{p_\theta(Z|x,a)}{p_\theta(Z)} && (22)\\
&= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{p_\theta(Z|x,a)}{p_\theta(Z)} + \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x,a)}{q(Z|x,a)} && (23)\\
&= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x,a)}{p_\theta(Z)} + D_{\mathrm{KL}}\big(p_\theta(Z|x,a)\|q(Z|x,a)\big) && (24)\\
&\geq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x,a)}{p_\theta(Z)} && (25)
\end{aligned}$$

which completes the proof. □

Next, we define a generic contrastive expression with $K-1$ "negative" samples of Z, and show that taking its expectation with respect to those samples yields a valid (i.e. normalized) probability density:

Lemma 5 (Normalized Variational) Given independent samples $z_{1:K-1}$ from p_θ, define:

$$q(z|x,a,z_{1:K-1}) \coloneqq \frac{p_\theta(z)\cdot e^{g_\nu(x,a,z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,z_i)}\right)} \tag{26}$$

then the following defines a normalized density:

$$q(Z|x,a) \coloneqq \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} q(Z|x,a,Z_{1:K-1}) \tag{27}$$

Proof. The expectation integrates to one:

$$\begin{aligned}
\int_{\mathcal{Z}} q(z|x,a)dz &= \int_{\mathcal{Z}} \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \frac{p_\theta(z)\cdot e^{g_\nu(x,a,z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)} dz && (28)\\
&= \mathbb{E}_{\substack{Z\sim p_\theta \\ Z_{1:K-1}\sim p_\theta^{K-1}}} \frac{e^{g_\nu(x,a,Z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,Z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)} && (29)\\
&= K\cdot \mathbb{E}_{Z_{1:K}\sim p_\theta^{K}} \frac{e^{g_\nu(x,a,Z_1)}}{\sum_{i=1}^{K} e^{g_\nu(x,a,Z_i)}} && (30)\\
&= \mathbb{E}_{Z_{1:K}\sim p_\theta^{K}} \frac{\sum_{j=1}^{K} e^{g_\nu(x,a,Z_j)}}{\sum_{i=1}^{K} e^{g_\nu(x,a,Z_i)}} = 1 && (31)
\end{aligned}$$

which completes the proof. □

These two results allow us to show that the information Z contains on a tuple x, a—with respect to the generator parameterized as θ—is lower-bounded by the x, a-conditioned contrastive loss between "positive" samples $Z \sim p_\theta(\cdot|x, a)$ from the posterior and "negative" samples $Z \sim p_\theta$ from the prior:

Lemma 6 (State-Action Lower Bound) The x, a-wise mutual information satisfies:

$$\begin{aligned}&\mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x,a;Z)\\ &\qquad\qquad \geq \mathbb{E}_{\substack{Z\sim p_\theta(\cdot|x,a)\\ Z_{1:K-1}\sim p_\theta^{K-1}}} \log \frac{e^{g_\nu(x,a,Z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,Z)}+\sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)}\end{aligned} \tag{32}$$

Proof. Use Lemmas 4 and 5, then Jensen's inequality:

$$\begin{aligned}&\mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x,a;Z)\\
&\qquad\qquad \geq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x,a)}{p_\theta(Z)} &(33)\\
&\qquad\qquad = \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \frac{q(Z|x,a,Z_{1:K-1})}{p_\theta(Z)} &(34)\\
&\qquad\qquad \geq \mathbb{E}_{\substack{Z\sim p_\theta(\cdot|x,a)\\ Z_{1:K-1}\sim p_\theta^{K-1}}} \log \frac{q(Z|x,a,Z_{1:K-1})}{p_\theta(Z)} &(35)\\
&\qquad\qquad = \mathbb{E}_{\substack{Z\sim p_\theta(\cdot|x,a)\\ Z_{1:K-1}\sim p_\theta^{K-1}}} \log \frac{e^{g_\nu(x,a,Z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,Z)}+\sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)} &(36)\end{aligned}$$

which completes the proof. □

Next, we show that our invariance loss (Objective 2) for a tuple x, a, z is equal to the pointwise mutual information in the limit of infinitely large negative batches, assuming an optimal critic parameter:

Lemma 7 (Pointwise Asymptotic Equality) Define the pointwise invariance loss:

$$\mathcal{L}^K_{\theta,\nu}(x,a,z) := \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \log \frac{e^{g_\nu(x,a,z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,z)}+\sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)} \tag{37}$$

and the optimal critic parameter:

$$\nu^* := \arg\max_\nu \mathbb{E}_{\substack{X\sim\rho_\pi\\ A\sim\pi(\cdot|X)\\ Y\sim\tau(\cdot|X,A)\\ Z\sim p_\theta(\cdot|X,A,Y)}} \mathcal{L}^K_{\theta,\nu}(X,A,Z) \tag{38}$$

Then $\lim_{K\to\infty} \mathcal{L}^K_{\theta,\nu^*}(x,a,z) = \mathrm{PMI}_\theta(x,a;z)$.

Proof. The $\mathbb{E}[\mathcal{L}^K_{\theta,\nu}(X,A,Z)]$ term is just the InfoNCE loss between variables Z and X, A, so we know that ν^* satisfies $g_{\nu^*}(x,a,z) = \log \frac{p_\theta(z|x,a)}{p_\theta(z)} + c(x,a)$. Substituting this back into $\mathcal{L}^K_{\theta,\nu}(x,a,z)$:

$$\begin{aligned}&\lim_{K\to\infty} \mathcal{L}^K_{\theta,\nu^*}(x,a,z) &(39)\\
&\quad = \lim_{K\to\infty} \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \log \frac{e^{g_{\nu^*}(x,a,z)}}{\frac{1}{K}\left(e^{g_{\nu^*}(x,a,z)}+\sum_{i=1}^{K-1} e^{g_{\nu^*}(x,a,Z_i)}\right)} &(40)\\
&\quad = \lim_{K\to\infty} \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \log \frac{\frac{p_\theta(z|x,a)}{p_\theta(z)}}{\frac{1}{K}\left(\frac{p_\theta(z|x,a)}{p_\theta(z)}+\sum_{i=1}^{K-1}\frac{p_\theta(Z_i|x,a)}{p_\theta(Z_i)}\right)} &(41)\\
&\quad = \lim_{K\to\infty} \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \left[\log \frac{p_\theta(z|x,a)}{p_\theta(z)} - \log \frac{\frac{p_\theta(z|x,a)}{p_\theta(z)}+\sum_{i=1}^{K-1}\frac{p_\theta(Z_i|x,a)}{p_\theta(Z_i)}}{K}\right] &(42)\end{aligned}$$

$$= \log \frac{p_\theta(z|x,a)}{p_\theta(z)} - \lim_{K\to\infty} \log \frac{\frac{p_\theta(z|x,a)}{p_\theta(z)} + K - 1}{K} = \text{PMI}_\theta(x,a;z) \tag{43}$$

which completes the proof. □

This gives us what we need to derive Proposition 1, which we restate using our subscript-less notation:

Proposition 8 (Optimal Invariance) The (state-action) invariance bonus satisfies:

$$\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x,a) \le \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \text{PMI}_\theta(x,a;Z) \tag{44}$$

and for the optimal ν^* the bound is asymptotically tight as $K\to\infty$:

$$\lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x,a) = \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \text{PMI}_\theta(x,a;Z) \tag{45}$$

Proof. Use Lemma 6 for the first part:

$$\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x,a) \coloneqq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \mathcal{L}^K_{\theta,\nu}(x,a,Z) \tag{46}$$

$$= \mathbb{E}_{\substack{Z\sim p_\theta(\cdot|x,a)\\ Z_{1:K-1}\sim p_\theta^{K-1}}} \log \frac{e^{g_\nu(x,a,Z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,Z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)} \tag{47}$$

$$\le \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \text{PMI}_\theta(x,a;Z) \tag{48}$$

and use Lemma 7 for the second part:

$$\lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x,a) = \lim_{K\to\infty} \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \mathcal{L}^K_{\theta,\nu^*}(x,a,Z) \tag{49}$$

$$= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \lim_{K\to\infty} \mathcal{L}^K_{\theta,\nu^*}(x,a,Z) \tag{50}$$

$$= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \text{PMI}_\theta(x,a;Z) \tag{51}$$

which completes the proof. □

Next, we show that Proposition 2 is true, which we similarly restate using our subscript-less notation:

Proposition 9 (Optimal Reconstruction) Denote with $\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x,a)$ the reconstruction bonus as in Objective 1, $\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x,a)$ the invariance bonus as in Objective 2, and their weighted sum for any λ:

$$J(\theta,\eta,\nu;\lambda) \coloneqq \mathbb{E}_{\substack{X\sim\rho_\pi\\ A\sim\pi(\cdot|X)}} \left[\frac{1}{\lambda}\mathcal{R}^{\text{rec.}}_{\theta,\eta}(X,A) + \lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(X,A)\right] \tag{52}$$

Then its minimax optimal value is zero:

$$\min_{\theta,\eta} \max_{\nu} J(\theta,\eta,\nu;\lambda) = 0 \tag{53}$$

Proof. Take any MDP, as in Figure 1(a). By reparameterization, we know there exists an equivalent graphical representation under which Z is exogenous, as in Figure 1(b). Assuming realizability, let η^* be such that $f_{\eta^*}=f$, and let θ^* be such that $p_{\theta^*}(Z|x,a,y)=p_{\eta^*}(Z|x,a,y)$ for any x,a,y. First, by construction we have that $Z\perp X,A$, so the mutual information between Z and X,A must be zero:

$$\mathbb{E}_{\substack{X\sim\rho_\pi\\ A\sim\pi(\cdot|X)}} \left[\lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(X,A)\right] = \mathbb{E}_{\substack{X\sim\rho_\pi\\ A\sim\pi(\cdot|X)\\ Z\sim p_\theta(\cdot|X,A)}} \text{PMI}_\theta(X,A;Z) \tag{54}$$

$$= \mathbb{I}_\theta[X,A;Z] = 0 \tag{55}$$

for optimal critic parameter ν^*, where the first equality uses Proposition 1. Second, by consistency of counterfactuals $f_{\eta^*}(x,a,Z)=y$ for any $Z\sim p_{\theta^*}(\cdot|x,a,y)$, so the reconstruction term is also zero. It is easy to verify the optimal critic is a maximizer, and the optimal generator/reconstructor minimizers, which completes the proof. □

Finally, we recall the following basic relationship:

Lemma 10 (Conditional Mutual Information) Conditioned on any x, a, we have that:

$$\mathbb{I}_\theta[Y;Z|x,a] = \mathbb{H}[Y|x,a] + \mathbb{H}_\theta[Y|x,a,Z] \tag{56}$$

Proof. Starting from the left hand side:

$$\mathbb{I}_\theta[Y;Z|x,a] \coloneqq \mathbb{E}_{Z\sim p_\theta} D_{\mathrm{KL}}\big(p_\theta(Y|x,a,Z)\|\tau(Y|x,a)\big) \tag{57}$$

$$= \mathbb{E}_{\substack{Z\sim p_\theta \\ Y\sim p_\theta(\cdot|x,a,Z)}} \log p_\theta(Y|x,a,Z) - \mathbb{E}_{\substack{Z\sim p_\theta \\ Y\sim p_\theta(\cdot|x,a,Z)}} \tau(Y|x,a) \tag{58}$$

$$= -\textstyle\int_{\mathcal{Z}} p_\theta(z)\mathbb{H}_\theta[Y|x,a,z]dz - \mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a) \\ Z\sim p_\theta(\cdot|x,a,Y)}} \tau(Y|x,a) \tag{59}$$

$$= \mathbb{H}[Y|x,a] - \mathbb{H}_\theta[Y|x,a,Z] \tag{60}$$

which completes the proof. □

Now, in our structural causal model, by construction Z captures all sources of noise—that is, there is no residual noise in each outcome Y. However, for the purposes of optimization, while learning η we let the residual error be captured by a Gaussian "log-likelihood" (note that λ plays the role of "$2\sigma^2$"):

$$\log p_\eta(Y|x,a,z) \coloneqq -\frac{1}{2}\log(\lambda\pi) - \frac{1}{\lambda}\big(Y - f_\eta(x,a,z)\big)^2 \tag{61}$$

and note that θ also induces a log-likelihood of the "ground-truth" conditional:

$$\log p_\theta(Y|x,a,z) \coloneqq \log \frac{p_\theta(z|x,a,Y)\tau(Y|x,a)\pi(a,x)\rho_\pi(x)}{\int_{\mathcal{Y}} p_\theta(z|x,a,y)\tau(y|x,a)\pi(a|x)\rho_\pi(x)dy} \tag{62}$$

Now, recall the reconstruction loss and (state-action) reconstruction bonus:

$$\mathcal{L}_\eta(x,a,z,y) \coloneqq \Big\| y - f_\eta(x,a,z)\Big\|_2^2 \tag{63}$$

$$\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x,a) \coloneqq \mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a) \\ Z\sim p_\theta(\cdot|x,a,Y)}} \mathcal{L}_\eta(x,a,Z,Y) \tag{64}$$

as well as the invariance loss and (state-action) invariance bonus:

$$\mathcal{L}^K_{\theta,\nu}(x,a,z) \coloneqq \mathbb{E}_{\substack{(X_1,\ldots,X_{K-1})\sim\prod_{i=1}^{K-1}\rho_\pi \\ (A_1,\ldots,A_{K-1})\sim\prod_{i=1}^{K-1}\pi(\cdot|X_i) \\ (Y_1,\ldots,Y_{K-1})\sim\prod_{i=1}^{K-1}\tau(\cdot|X_i,A_i) \\ (Z_1,\ldots,Z_{K-1})\sim\prod_{i=1}^{K-1}p_\theta(\cdot|X_i,A_i,Y_i)}} \log \frac{e^{g_\nu(x,a,z)}}{\frac{1}{K}\Big(e^{g_\nu(x,a,z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\Big)} \tag{65}$$

$$\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x,a) \coloneqq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathcal{L}^K_{\theta,\nu}(x,a,Z) \tag{66}$$

Moreover, recall the hindsight intrinsic reward function:

$$\mathcal{R}_{\theta,\eta,\nu^*}(x,a) \coloneqq \frac{1}{\lambda}\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x,a) + \lim_{K\to\infty}\mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x,a) \tag{67}$$

We can now show that Theorem 3 is true, which we similarly restate using our subscript-less notation:

Theorem 11 (Optimistic Exploration) Let λ satisfy the inequality $\frac{1}{2}\log(\lambda\pi) \leq \mathbb{H}_\theta[Y|x,a,Z] + D_{\mathrm{KL}}\big(p_\theta(Z|x,a)\|p_\theta(Z)\big)$, with π here being the mathematical constant (not the agent's policy). Then:

$$\mathcal{R}_{\theta,\eta,\nu^*}(x,a) \geq D_{\mathrm{KL}}\big(\tau(Y|x,a)\|\tau_{\theta,\eta}(Y|x,a)\big) \tag{68}$$

where $\tau_{\theta,\eta}(Y|x,a) \coloneqq \mathbb{E}_{Z\sim p_\theta} p_\eta(Y|x,a,Z)$ denotes the learned environment model. Furthermore, for optimal model parameters θ^*, η^* we have that the intrinsic reward $\mathcal{R}_{\theta^*,\eta^*,\nu^*}(x,a) = 0$ for all x, a.

Proof. Use Proposition 8, then the constraint on λ, then Lemma 10:

$$\begin{aligned}
\mathcal{R}_{\theta,\eta,\nu^*}(x,a) &\coloneqq \tfrac{1}{\lambda}\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x,a) + \lim_{K\to\infty}\mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x,a) && (69)\\
&= \mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \tfrac{1}{\lambda}\big(Y - f_\eta(x,a,Z)\big)^2 + \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\text{PMI}_\theta(x,a;Z) && (70)\\
&= \mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \tfrac{1}{\lambda}\big(Y - f_\eta(x,a,Z)\big)^2 + D_{\text{KL}}\big(p_\theta(Z|x,a)\|p_\theta(Z)\big) && (71)\\
&\geq -\mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \log p_\eta(Y|x,a,Z) - \mathbb{H}_\theta[Y|x,a,Z] && (72)\\
&= -\mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \log p_\eta(Y|x,a,Z) + \mathbb{I}_\theta[Y;Z|x,a] - \mathbb{H}[Y|x,a] && (73)\\
&= -\mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \log p_\eta(Y|x,a,Z) \leftarrow \textbf{remaining}\text{ stochasticity}\\
&\quad + \mathbb{E}_{Y\sim\tau(\cdot|x,a)} D_{\text{KL}}\big(p_\theta(Z|x,a,Y)\|p_\theta(Z|x,a)\big) \leftarrow \textbf{hindsight}\text{ information}\\
&\quad - \mathbb{E}_{Y\sim\tau(\cdot|x,a)}\big[-\log\tau(Y|x,a)\big] \leftarrow \textbf{total}\text{ stochasticity} && (74)\\
&\geq -\mathbb{E}_{Y\sim\tau(\cdot|x,a)}\Big[\mathbb{E}_{Z\sim p_\theta(\cdot|x,a,Y)}\log p_\eta(Y|x,a,Z)\\
&\quad - D_{\text{KL}}\big(p_\theta(Z|x,a,Y)\|p_\theta(Z|x,a)\big) + D_{\text{KL}}\big(p_\theta(Z|x,a,Y)\|p_\eta(Z|x,a,Y)\big)\Big]\\
&\quad + \mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\tau(Y|x,a) && (75)\\
&= -\mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\mathbb{E}_{Z\sim p_\theta}p_\eta(Y|x,a,Z) + \mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\tau(Y|x,a) && (76)\\
&= -\mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\tau_{\theta,\eta}(Y|x,a) + \mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\tau(Y|x,a) && (77)\\
&= D_{\text{KL}}\big(\tau(Y|x,a)\|\tau_{\theta,\eta}(Y|x,a)\big) && (78)
\end{aligned}$$

which completes the proof. □

The intuition is as follows: Assuming realizability, at convergence "hindsight information" and "total stochasticity" cancel (i.e. neither more nor less), and the "remaining stochasticity" term goes to zero.