Appendix for CARMS: Categorical-Antithetic-REINFORCE Multi-Sample Gradient Estimator

A Background

A.1 Copulas

Although any multivariate random variable with uniform marginal distributions is commonly referred to as a copula, strictly speaking, a copula is specifically its CDF. We are interested in copulas with strong negative dependence between any pair of its variables, which we refer to as antithetic in this work. They are also referred to as countermonotonic copulas when investigated in other work [Lee and Ahn, 2014, McNeil and Nešlehová, 2009]. However, only in the bivariate case do these copulas achieve the theoretical lower bound for negative dependence, the Fréchet–Hoeffding bound:

$$\mathcal{C}(u_1, ..., u_N) \geq \max\Big\{1 - N + \sum_{i=1}^{N} u_i, 0\Big\}.$$

Furthermore, it has been shown [McNeil and Nešlehová, 2009] that no copula can achieve this lower bound for more than two dimensions. However, the Dirichlet copula used in this work is in the family of Archimedean copulas, and Lee and Ahn [2014] show that within this family, the Dirichlet copula has the strongest countermonotonicity. Its CDF is defined as:

$$\mathcal{C}(u_1, ..., u_N) = \left(\max\left\{0, u_1^{\frac{1}{N-1}} + ... + u_N^{\frac{1}{N-1}} - (N-1)\right\}\right)^{N-1}.$$

A.2 Copula sampling

We use the Dirichlet copula described in ARMS [Dimitriev and Zhou, 2021], which transforms a Dirichlet vector with concentration $\boldsymbol{\alpha} = \mathbf{1}$ to a copula sample. Besides its countermonotonic properties, we choose this copula because both its univariate and bivariate marginal CDFs are analytically tractable. The former is required for transforming the Dirichlet vector $\boldsymbol{d} \sim \text{Dir}(\boldsymbol{\alpha})$ into uniform variables: $\boldsymbol{u} = 1 - (1 - \boldsymbol{d})^{N-1}$. For Eq. 8, we want to analytically calculate the bivariate joint for categorical sampling. This requires the bivariate CDF, which has the form:

$$\begin{aligned} P(\boldsymbol{u}_i < p, \boldsymbol{u}_j < q) &= P(\boldsymbol{d}_i < 1 - (1-p)^{1/(N-1)}, \boldsymbol{d}_j < 1 - (1-q)^{1/(N-1)}) \\ &= p + q - 1 + \max(0, (1-p)^{1/(N-1)} + (1-q)^{1/(N-1)} \end{aligned}$$

A.3 Pair equivalent definition of LOORF

Because the LOORF for N samples for any distribution can be decomposed into all pairs, we can reuse the theorem, and we reproduce the short algebra needed to show it below:

$$\begin{aligned} g_{\text{loorf}}(\boldsymbol{b}) &= \frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{b}_i) - \frac{1}{n-1}\sum_{j\neq i} f(\boldsymbol{b}_j)\right)\nabla_{\boldsymbol{\phi}} \ln p(\boldsymbol{b}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{b}_i)\nabla_{\boldsymbol{\phi}} \ln p(\boldsymbol{b}_i) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\nabla_{\boldsymbol{\phi}} \ln p(\boldsymbol{b}_i)\sum_{j\neq i} f(\boldsymbol{b}_j) \\ &= \frac{1}{n(n-1)}\sum_{i\neq j}\Big(f(\boldsymbol{b}_i)\nabla_{\boldsymbol{\phi}} \ln p(\boldsymbol{b}_i) - f(\boldsymbol{b}_i)\nabla_{\boldsymbol{\phi}} \ln p(\boldsymbol{b}_j)\Big) \\ &= \frac{1}{n(n-1)}\sum_{i\neq j}\frac{1}{2}(f(\boldsymbol{b}_i) - f(\boldsymbol{b}_j))(\nabla_{\boldsymbol{\phi}} \ln p(\boldsymbol{b}_i) - \nabla_{\boldsymbol{\phi}} \ln p(\boldsymbol{b}_j)) \\ &= \frac{1}{n(n-1)}\sum_{i\neq j} g_{\text{pod}}(b_i, b_j). \end{aligned}$$

B Proofs

B.1 Proof of Theorem 2

Define $\Delta f_{ij} = f(\boldsymbol{z}_i) - f(\boldsymbol{z}_j)$. Using the assumption that $P(\boldsymbol{z} = \hat{\boldsymbol{\imath}}, \boldsymbol{z}' = \hat{\boldsymbol{\jmath}}) \geq P(\boldsymbol{z} = \hat{\boldsymbol{\imath}})P(\boldsymbol{z}' = \hat{\boldsymbol{\jmath}})$, and starting with CARTS we have:

$$\begin{aligned}
\text{Var}[g_{\text{CARTS}}] &= \sum_{i \neq j} P(\boldsymbol{z} = \hat{\boldsymbol{\imath}}, \boldsymbol{z}' = \hat{\boldsymbol{\jmath}}) \left(\frac{1}{2} \Delta f_{ij} (\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}}) \frac{P(\boldsymbol{z} = \hat{\boldsymbol{\imath}})P(\boldsymbol{z}' = \hat{\boldsymbol{\jmath}})}{P(\boldsymbol{z} = \hat{\boldsymbol{\imath}}, \boldsymbol{z}' = \hat{\boldsymbol{\jmath}})} \right)^2 \\
&= \sum_{i \neq j} P(\boldsymbol{z} = \hat{\boldsymbol{\imath}})P(\boldsymbol{z}' = \hat{\boldsymbol{\jmath}}) \left(\frac{1}{2} \Delta f_{ij} (\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}}) \right)^2 \frac{P(\boldsymbol{z} = \hat{\boldsymbol{\imath}})P(\boldsymbol{z}' = \hat{\boldsymbol{\jmath}})}{P(\boldsymbol{z} = \hat{\boldsymbol{\imath}}, \boldsymbol{z}' = \hat{\boldsymbol{\jmath}})} \\
&< \sum_{i \neq j} P(\boldsymbol{z} = \hat{\boldsymbol{\imath}})P(\boldsymbol{z}' = \hat{\boldsymbol{\jmath}}) \left(\frac{1}{2} \Delta f_{ij} (\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}}) \right)^2 = \text{Var}[g_{\text{LOORF}}]
\end{aligned}$$

Below is an example that satisfies the conditions. Let $\boldsymbol{p} = (0.6, 0.3, 0.1)$, for which we show the independent vs one possible antithetic pmf:

$$\text{pmf}_{\text{indep}} = \boldsymbol{p}\boldsymbol{p}^T = \begin{bmatrix} 0.36 & 0.18 & 0.06 \\ 0.18 & 0.09 & 0.03 \\ 0.06 & 0.03 & 0.01 \end{bmatrix} \qquad \text{pmf}_{\text{anti}} = \begin{bmatrix} 0.30 & 0.24 & 0.06 \\ 0.24 & 0.02 & 0.04 \\ 0.06 & 0.04 & 0.01 \end{bmatrix}.$$

Because every off-diagonal element of pmf_{anti} is larger or equal to the corresponding independent pmf value, the variance is guaranteed to be no larger:

$$\begin{aligned}
\text{Var}(g_{\text{CARTS}}) &= 2 \left(0.24 \Delta f_{12} [1,1,0]^T \left(\frac{0.18}{0.24} \right)^2 + 0.06 \Delta f_{13} [1,0,1]^T + 0.04 \Delta f_{23} [0,1,1]^T \left(\frac{0.03}{0.04} \right)^2 \right) \\
&= 2 \left(0.18 \Delta f_{12} [1,1,0]^T \frac{0.18}{0.24} + 0.06 \Delta f_{13} [1,0,1]^T + 0.03 \Delta f_{23} [0,1,1]^T \frac{0.03}{0.04} \right) \\
&\leq 2 \left(0.18 \Delta f_{12} [1,1,0]^T + 0.06 \Delta f_{13} [1,0,1]^T + 0.03 \Delta f_{23} [0,1,1]^T \right) = \text{Var}(g_{\text{2LOORF}}).
\end{aligned}$$

B.2 Proof of Lemma 4

First, note that the $(ij)^{\text{th}}$ element of $\boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{O}}$ is $\left(\sum_{j \neq i} \boldsymbol{\mathcal{R}}_{\boldsymbol{ij}} \right) - \boldsymbol{\mathcal{R}}_{ij}$. Then we can explicitly write out the summation, and distribute it into all pairs, to arrive at the form of Theorem 3:

$$\begin{aligned}
g_{\text{CARMS}} &= \frac{1}{N} f(\boldsymbol{Z})^T (\boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{O}}) \Big(\boldsymbol{Z} - \boldsymbol{1}_N \sigma(\boldsymbol{\phi})^T \Big) \\
&= \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{Z})_i^T \Big(\sum_{j \neq i} \boldsymbol{\mathcal{R}}_{ij} \Big) - \Big(\sum_{j \neq i} f(\boldsymbol{Z})_j^T \boldsymbol{\mathcal{R}}_{ij} \Big) (\boldsymbol{Z}_i - \sigma(\boldsymbol{\phi})) \\
&= \tfrac{1}{N(N-1)} \sum_{i \neq j} \Big(f(\boldsymbol{Z})_i - f(\boldsymbol{Z})_j \Big) (\boldsymbol{z}_i - \boldsymbol{z}_j) \boldsymbol{z}_i^T \boldsymbol{\mathcal{R}}_{ij} \boldsymbol{z}_j
\end{aligned}$$

B.3 Proof of Equation 8

$$\begin{aligned}
P(\boldsymbol{z} = \hat{\boldsymbol{\imath}}, \boldsymbol{z}' = \hat{\boldsymbol{\jmath}}) &= P\left(u \in [\boldsymbol{l}_i, \boldsymbol{r}_i], u' \in [\boldsymbol{l}_j, \boldsymbol{r}_j] \right) \\
&= P\left(u \leq \boldsymbol{r}_i, u' \in [\boldsymbol{l}_j, \boldsymbol{r}_j] \right) - P\left(u \leq \boldsymbol{l}_i, u' \in [\boldsymbol{l}_j, \boldsymbol{r}_j] \right) \\
&= P\left(u \leq \boldsymbol{r}_i, u' \leq \boldsymbol{r}_j \right) - P\left(u \leq \boldsymbol{r}_i, u' \leq \boldsymbol{l}_j \right) \\
&\qquad - P\left(u \leq \boldsymbol{l}_i, u' \leq \boldsymbol{r}_j \right) + P\left(u \leq \boldsymbol{l}_i, u' \leq \boldsymbol{l}_j \right) \\
&= \Phi_{\mathcal{C}}(\boldsymbol{r}_i, \boldsymbol{r}_j) - \Phi_{\mathcal{C}}(\boldsymbol{r}_i, \boldsymbol{l}_j) - \Phi_{\mathcal{C}}(\boldsymbol{l}_i, \boldsymbol{r}_j) + \Phi_{\mathcal{C}}(\boldsymbol{l}_i, \boldsymbol{l}_j).
\end{aligned}$$

B.4 Proof that the bivariate PMF for the pair (1, C) is non-zero

We show that joint probability of the first and last category $P(\boldsymbol{z} = \hat{\mathbf{1}}, \boldsymbol{z}' = \hat{\boldsymbol{C}})$ is always positive, because the Dirichlet copula has an area of non-zero density in a subset of this region. If we take $\epsilon = \min(p_1, p_C) > 0$, then:

$$P(\boldsymbol{z} = \hat{\mathbf{1}}, \boldsymbol{z}' = \hat{\boldsymbol{C}}) = P(u < p_1, u' > 1 - p_C) \geq P(u < \epsilon, u' > 1 - \epsilon)$$
$$= P(u < \epsilon) - P(u < \epsilon, u' < 1 - \epsilon) = \epsilon - \max\left\{0, \epsilon^{\frac{1}{N-1}} + (1-\epsilon)^{\frac{1}{N-1}} - 1\right\}^{N-1}.$$

If the second term is zero, it trivially holds that the probability is non-zero since $\epsilon > 0$. Otherwise, note that for any integer $N > 1$:

$$\epsilon > 0 \iff 1 - \epsilon < 1 \iff (1-\epsilon)^{\frac{1}{N-1}} < 1 \iff (1-\epsilon)^{\frac{1}{N-1}} - 1 < 0$$
$$\iff \epsilon^{\frac{1}{N-1}} + (1-\epsilon)^{\frac{1}{N-1}} - 1 < \epsilon^{\frac{1}{N-1}} \iff \left(\epsilon^{\frac{1}{N-1}} + (1-\epsilon)^{\frac{1}{N-1}} - 1\right)^{N-1} < \epsilon$$
$$\iff \epsilon - \left(\epsilon^{\frac{1}{N-1}} + (1-\epsilon)^{\frac{1}{N-1}} - 1\right)^{N-1} > 0,$$

which concludes the proof.

C Additional results

Table 3: Final test log likelihood of VAEs using different estimators, where the stochastic layer contains C=3, 5, or 10 categories, with $\lfloor 200/C \rfloor$ latent variables and C samples per gradient step, respectively. Results are reported on three datasets: Dynamic MNIST, Fashion MNIST, and Omniglot over 5 runs.

Categories			CARMS-I	CARMS-G	LOORF	UNORD	ARSM
Dynamic MNIST	Linear	3	-104.82 ± 0.25	-104.9 ± 0.25	-105.17 ± 0.24	-104.73 ± 0.24	-106.85 ± 0.57
		5	-103.12 ± 0.13	-102.9 ± 0.18	-103.16 ± 0.16	-103.06 ± 0.13	-105.73 ± 0.54
		10	-103.0 ± 0.04	-102.88 ± 0.06	-103.19 ± 0.06	-103.35 ± 0.08	-106.41 ± 0.59
	Nonlinr	3	-95.73 ± 0.32	-95.44 ± 0.31	-95.98 ± 0.25	-96.09 ± 0.25	-101.33 ± 0.54
		5	-93.83 ± 0.17	-93.27 ± 0.13	-93.67 ± 0.18	-93.69 ± 0.13	-99.61 ± 0.48
		10	-92.97 ± 0.07	-93.26 ± 0.11	-93.22 ± 0.04	-93.94 ± 0.15	-98.73 ± 0.44
Fashion MNIST	Linear	3	-247.46 ± 0.22	-247.77 ± 0.18	-247.82 ± 0.19	-247.92 ± 0.21	-249.5 ± 0.46
		5	-244.16 ± 0.13	-244.02 ± 0.1	-244.27 ± 0.1	-244.55 ± 0.13	-246.69 ± 0.48
		10	-242.6 ± 0.03	-242.69 ± 0.04	-243.1 ± 0.04	-243.27 ± 0.06	-245.43 ± 0.37
	Nonlinr	3	-235.81 ± 0.18	-235.89 ± 0.17	-236.47 ± 0.11	-236.06 ± 0.13	-240.46 ± 0.21
		5	-234.37 ± 0.09	-234.27 ± 0.04	-234.7 ± 0.05	-234.81 ± 0.09	-239.83 ± 0.34
		10	-233.39 ± 0.05	-233.81 ± 0.05	-234.02 ± 0.04	-234.41 ± 0.03	-238.55 ± 0.18
Omniglot	Linear	3	-115.15 ± 0.13	-115.31 ± 0.15	-115.46 ± 0.14	-115.33 ± 0.16	-116.92 ± 0.43
		5	-114.88 ± 0.1	-114.86 ± 0.12	-114.89 ± 0.11	-114.86 ± 0.1	-116.56 ± 0.37
		10	-116.31 ± 0.05	-116.41 ± 0.05	-116.55 ± 0.06	-116.92 ± 0.11	-118.79 ± 0.39
	Nonlinr	3	-114.54 ± 0.31	-114.4 ± 0.23	-114.6 ± 0.35	-114.5 ± 0.27	-119.0 ± 0.57
		5	-113.18 ± 0.12	-113.39 ± 0.16	-113.44 ± 0.21	-113.69 ± 0.18	-119.41 ± 0.41
		10	-112.71 ± 0.13	-112.7 ± 0.03	-112.93 ± 0.07	-113.8 ± 0.17	-120.05 ± 0.42

Table 4: Final test log likelihood of a categorical network for conditional estimation using different gradient estimators, where the stochastic layer contains C=3, 5, or 10 categories, with $\lfloor 200/C \rfloor$ latent variables and C samples per gradient step, respectively. Results are reported on three datasets: Dynamic MNIST, Fashion MNIST, and Omniglot over 5 runs.

Categories		CARMS-I	CARMS-G	LOORF	UNORD	ARSM
Dynamic MNIST	3	59.61 ± 0.17	59.89 ± 0.14	59.78 ± 0.2	59.72 ± 0.06	60.92 ± 0.2
	5	59.43 ± 0.03	59.69 ± 0.1	59.59 ± 0.18	59.5 ± 0.09	60.32 ± 0.16
	10	59.93 ± 0.22	60.01 ± 0.16	59.95 ± 0.09	59.93 ± 0.19	60.25 ± 0.15
Fashion MNIST	3	134.92 ± 0.12	135.02 ± 0.07	135.16 ± 0.05	135.18 ± 0.08	136.17 ± 0.37
	5	134.81 ± 0.1	134.94 ± 0.13	135.05 ± 0.06	135.1 ± 0.11	135.78 ± 0.13
	10	135.35 ± 0.13	135.44 ± 0.24	135.61 ± 0.18	135.47 ± 0.1	135.81 ± 0.2
Omniglot	3	72.19 ± 0.11	72.45 ± 0.08	72.33 ± 0.15	72.34 ± 0.1	72.73 ± 0.12
	5	72.32 ± 0.14	72.56 ± 0.14	72.43 ± 0.11	72.59 ± 0.08	72.79 ± 0.24
	10	72.79 ± 0.12	72.88 ± 0.06	72.9 ± 0.06	72.94 ± 0.17	72.98 ± 0.15

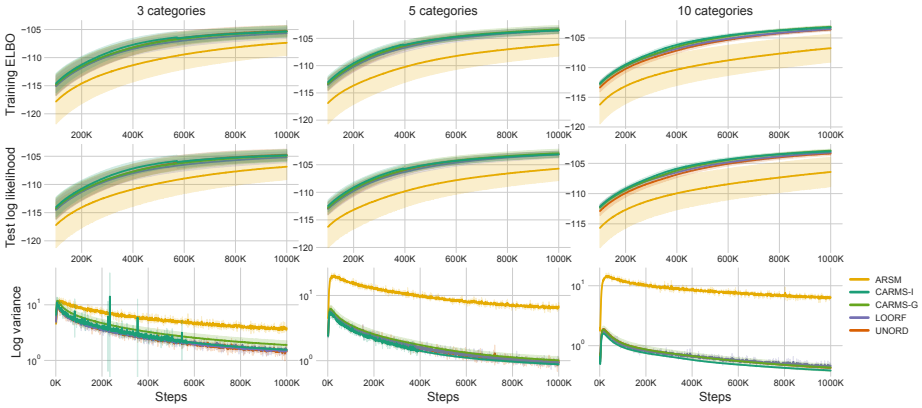

Figure 4: Training a linear categorical VAE with different estimators on Dynamic MNIST using ELBO. Columns correspond to $C \in \{3, 5, 10\}$ categories with C samples per gradient step, respectively. Rows correspond to the 100 sample training and test log likelihood, and the variance of the gradient with respect to the logits of the encoder network.

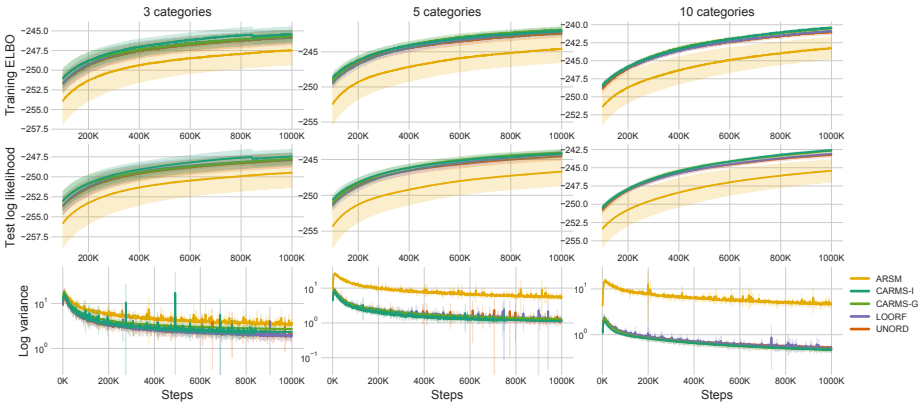

Figure 5: Training a linear categorical VAE with different estimators on Fashion MNIST using ELBO. Columns correspond to $C \in \{3, 5, 10\}$ categories with C samples per gradient step, respectively. Rows correspond to the 100 sample training and test log likelihood, and the variance of the gradient with respect to the logits of the encoder network.

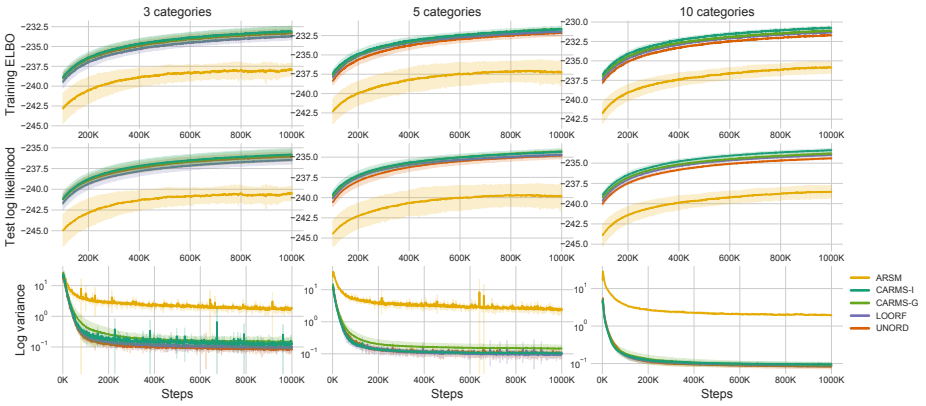

Figure 6: Training a nonlinear categorical VAE with different estimators on Fashion MNIST using ELBO. Columns correspond to $C \in \{3, 5, 10\}$ categories with C samples per gradient step, respectively. Rows correspond to the 100 sample training and test log likelihood, and the variance of the gradient with respect to the logits of the encoder network.

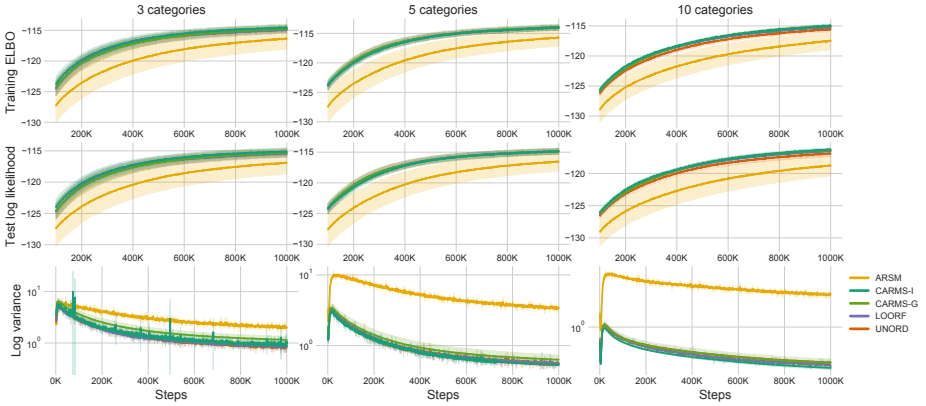

Figure 7: Training a linear categorical VAE with different estimators on Omniglot using ELBO. Columns correspond to $C \in \{3, 5, 10\}$ categories with C samples per gradient step, respectively. Rows correspond to the 100 sample training and test log likelihood, and the variance of the gradient with respect to the logits of the encoder network.

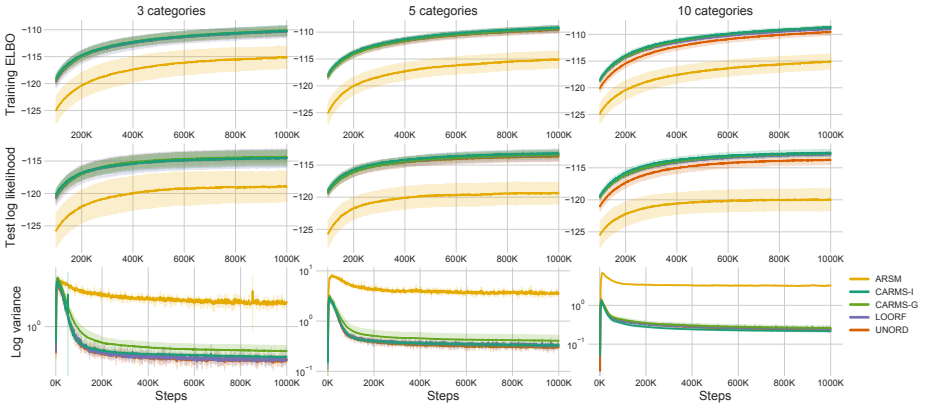

Figure 8: Training a nonlinear categorical VAE with different estimators on Omniglot using ELBO. Columns correspond to $C \in \{3, 5, 10\}$ categories with C samples per gradient step, respectively. Rows correspond to the 100 sample training and test log likelihood, and the variance of the gradient with respect to the logits of the encoder network.