Appendix

A. Proofs

Eq 2. LOORF unbiasedness and alternate form

It is easy to see that for independent samples LOORF is unbiased, because $\mathbb{E}[\nabla_\phi \ln p(b)] = 0$. Using the linearity property of expectations and the fact that for independent variables $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, we have:

$$\begin{aligned}\mathbb{E}\left[\left(f(b_i) - \frac{1}{n-1}\sum_{j\neq i} f(b_j)\right)\nabla_\phi \ln p(b_i)\right] &= \mathbb{E}\big[f(b_i)\nabla_\phi \ln p(b_i)\big] - \frac{1}{n-1}\sum_{j\neq i}\mathbb{E}\big[f(b_j)\big]\mathbb{E}[\nabla_\phi \ln p(b_i)] \\ &= \mathbb{E}\big[f(b_i)\nabla_\phi \ln p(b_i)\big] = \nabla_{\boldsymbol{\phi}}\mathcal{E}(\boldsymbol{\phi}).\end{aligned}$$

The alternate form of LOORF is useful because we can easily calculate the mean and subtract it from each sample. For the i^{th} sample, we have:

$$\begin{aligned}&\left(f(b_i) - \frac{1}{n-1}\sum_{j\neq i} f(b_j)\right)\nabla_\phi \ln p(b_i) = \frac{n}{n-1}\left(\frac{n-1}{n}f(b_i) - \frac{1}{n}\sum_{j\neq i} f(b_j)\right)\nabla_\phi \ln p(b_i) \\ &= \frac{n}{n-1}\left(\Big(1-\frac{1}{n}\Big)f(b_i) - \frac{1}{n}\sum_{j\neq i} f(b_j)\right)\nabla_\phi \ln p(b_i) = \frac{n}{n-1}\left(f(b_i) - \frac{1}{n}\sum_{i=1}^{n} f(b_j)\right)\nabla_\phi \ln p(b_i).\end{aligned}$$

Eq 4. PoD for binary variables

First, we show that LOORF for $n = 2$ samples is indeed PoD:

$$\begin{aligned}g_{\text{loorf}}(b_1, b_2) &= \frac{1}{2}\Big(\big(f(b_1) - f(b_2)\big)(\nabla_\phi \ln p(b_1)) + \big(f(b_2) - f(b_1)\big)(\nabla_\phi \ln p(b_2))\Big) \\ &= \frac{1}{2}\big(f(b_1) - f(b_2)\big)\big(\nabla_\phi \ln p(b_1) - \nabla_\phi \ln p(b_2)\big) = g_{\text{pod}}(b_1, b_2).\end{aligned}$$

When both variables are Bern(p), the score function is $\nabla_\phi \ln p_\phi(b) = b - \sigma(\phi)$, which simplifies the estimator:

$$g_{\text{pod}}(b, b') = \big(f(b) - f(b')\big)\big(b - \sigma(\phi) - b' + \sigma(\phi)\big) = \big(f(b) - f(b')\big)\big(b - b'\big).$$

For the expected value, first note that symmetric marginals $P(b = 1) = P(b' = 1)$ imply:

$$P(b = 1, b' = 0) = P(b = 1) - P(b = 1, b' = 1) = P(b' = 1) - P(b = 1, b' = 1) = P(b = 0, b' = 1),$$

and also that the estimator is zero when $b = b'$. With these two things in mind, we have:

$$\begin{aligned}\mathbb{E}\big[g_{\text{pod}}(b, b')\big] &= \frac{1}{2}P(b = 1, b' = 0)\big(f(1) - f(0)\big)(1 - 0) + \frac{1}{2}P(b = 0, b' = 1)\big(f(0) - f(1)\big)(0 - 1) \\ &= \frac{1}{2}\big(P(b = 1, b' = 0) + P(b = 0, b' = 1)\big)\big(f(1) - f(0)\big) = \big(f(1) - f(0)\big)P(b = 1, b' = 0).\end{aligned}$$

If b and b' are independent $P(b = 1, b' = 0) = p(1 - p)$, which indeed coincides with the analytical gradient:

$$\begin{aligned}\nabla_\phi \mathcal{E}(\phi) = \nabla_\phi \mathbb{E}[f(b)] &= \nabla_\phi\Big(\sigma(\phi)f(1) + (1 - \sigma(\phi))f(0)\Big) = \sigma(\phi)(1 - \sigma(\phi))f(1) - \sigma(\phi)(1 - \sigma(\phi))f(0) \\ &= \big(f(1) - f(0)\big)\sigma(\phi)(1 - \sigma(\phi)) = \big(f(1) - f(0)\big)p(1 - p),\end{aligned}$$

where we used the fact that $\nabla_\phi \sigma(\phi) = \sigma(\phi)(1 - \sigma(\phi))$.

Theorem 2. Variance of ARTS

The easiest way to compare the variance of ARTS to PoD is to just compute both. Define $\Delta f = f(1) - f(0)$, and the shorthand for the true gradient: $\nabla_\phi = \nabla_\phi \mathcal{E}(\phi) = \Delta f p(1-p)$. Since ARTS reduces to PoD for $\rho = 0$, and we know their expected values (∇_ϕ), the variance is:

$$\mathbb{E}\big[g^2_{\text{arts}}\big] = \mathbb{E}\left[\left(\frac{1}{2}\big(f(b) - f(b')\big)(b-b')\right)^2 \left(\frac{2p(1-p)}{P(b \neq b')}\right)^2\right] = \frac{1}{P(b \neq b')}\big(\Delta f p(1-p)\big)^2 = \frac{\nabla^2_\phi}{P(b \neq b')}$$
$$\implies \text{Var}(g^2_{\text{arts}}) = \mathbb{E}\big[g^2_{\text{arts}}\big] - \mathbb{E}\big[g_{\text{arts}}\big]^2 = \nabla^2\left(\frac{1}{P(b \neq b')} - 1\right).$$

Rewriting this using the correlation will make the relationship clear:

$$\frac{1}{1-\rho} = \frac{2p(1-p)}{P(b \neq b')} \implies \frac{1}{P(b \neq b')} = \frac{1}{2p(1-p)(1-\rho)}$$
$$\rho < 0 \implies 2p(1-p)(1-\rho) > 2p(1-p) \implies \frac{1}{2p(1-p)(1-\rho)} < \frac{1}{2p(1-p)},$$

which implies, for $\rho < 0$:

$$\text{Var}(g^2_{\text{arts}}) = \nabla^2\left(\frac{1}{2p(1-p)(1-\rho)} - 1\right) < \nabla^2\left(\frac{1}{2p(1-p)} - 1\right) = \text{Var}(g^2_{\text{pod}}).$$

It is clear from above that the variance is an increasing (typo in original paper) function of ρ, and the lowest value it can achieve is the lower limit on a correlation of two Bernoulli variables. This depends on maximizing the probability $P(b=1, b'=0)$ as seen below, for which we have: $P(b=1,b'=0) \leq P(b=1)$ as well as $P(b=1,b'=0) \leq P(b'=0)$, which implies $P(b=1,b'=0) \leq \min(p, 1-p)$. Putting it all together:

$$\rho = \frac{P(b=1,b'=1) - p^2}{p(1-p)} = \frac{p - P(b=1,b'=0) - p^2}{p(1-p)} = \frac{p - \min(p,1-p) - p^2}{p(1-p)} = -\min\left(\frac{p}{1-p}, \frac{1-p}{p}\right).$$

The lower limit is achieved when $P(b=1,b'=0) = P(u < p, (1-u) > p) = P(u < p, u < (1-p)) = \min(p, 1-p)$, with $u \sim \text{Unif}(0,1)$, which corresponds to the DisARM/U2G sampling method. Lastly, the debiasing term is:

$$\frac{1}{1-\rho} = \frac{p(1-p)}{P(b=1,b'=0)} = \frac{p(1-p)}{\min(p,1-p)} = \max(p, 1-p)$$

Eq 6. LOORF is all PoD pairs

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

B. Theorem 4. Dirichlet Copula Derivation

The two things we need to derive are the univariate and bivariate CDF. For both, we will make use of the Dirichlet aggregation property (Ng et al., 2011): if $\boldsymbol{d} = (d_1, ..., d_n) \sim \text{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, .., \alpha_n)$, then:

$$(d_1, ..., d_i + d_j, ..., d_n) \sim \text{Dir}(\alpha_1, ..., \alpha_i + \alpha_j, ..., \alpha_n).$$

This is easily seen to be true from the Gamma sampling procedure:

$$g_i \sim \text{Gamma}(\alpha_i, \theta),\ d_i = \frac{g_i}{\sum_j g_j} \implies (d_1, ..., d_n) \sim \text{Dir}(\alpha_1, ..., \alpha_n),$$

because adding any two elements has no influence on the distribution of the others, and the fact that $g_i + g_j \sim \text{Gamma}(\alpha_i + \alpha_j, 1)$.

B.1. Univariate CDF

Using the aggregation property, we see that the marginal distribution of d_i is:

$$(d_i, \sum_{j\neq i} d_j) \sim \text{Dir}(\alpha_i, \sum_{j\neq i} \alpha_j) = \text{Beta}(\alpha_i, \sum_{j\neq i} \alpha_j).$$

The CDF of the Beta distribution is the regularized incomplete Beta function $I_x(a, b)$, which has a simple closed-form value when $a = 1$ or $b = 1$:

$$I_x(a, b) = \frac{B(x, a, b)}{B(a, b)} = \frac{\int_0^x t^{a-1}(1-t)^{b-1}dt}{\int_0^1 t^{a-1}(1-t)^{b-1}dt}, \quad I_x(1, b) = 1 - (1-x)^b, \quad I_x(a, 1) = x^a.$$

Since $1 - d_i \sim \text{Beta}(\sum_{j\neq i} \alpha_j, \alpha_i)$, given a sample $\boldsymbol{d} \sim \text{Dir}(\mathbf{1}_n)$ we can obtain two different copula samples using the probability integral transform from Eq. 3. If $\boldsymbol{\alpha} = \mathbf{1}_n$, then the two marginal CDFs are $I_x(1, n-1)$ and $I_x(n-1, 1)$ for d_i and $1 - d_i$, respectively.

B.2. Bivariate CDF

The bivariate CDF is necessary for calculating $\rho = \text{Corr}(\tilde{b}_i, \tilde{b}_j)$ because:

$$\rho = \frac{E[b_i b_j] - p^2}{p(1-p)} = \frac{P(b_i = 1, b_j = 1 - p^2}{p(1-p)} = \frac{P(\tilde{u}_i < p, \tilde{u}_j < p) - p^2}{p(1-p)}.$$

The transformation $\tilde{u}_i = 1 - (1 - d_i)^{n-1}$ implies $d_i = 1 - (1 - \tilde{u}_i)^{\frac{1}{n-1}}$, which makes the unknown term:

$$P(\tilde{u}_i < p, \tilde{u}_j < p) = P(d_i < 1 - (1-p)^{1/(n-1)}, d_j < 1 - (1-p)^{1/(n-1)}),$$

so the calculation reduces to finding the bivariate Dirichlet distribution of (d_i, d_j). Using the aggregation property again, the density of (d_i, d_j) is equivalent in both of these cases:

$$(d_1, ..., d_i, ..., d_j, ..., d_n) \sim \text{Dir}(\alpha_1, ..., \alpha_i, ..., \alpha_j, ..., \alpha_n) \iff (d_i, d_j, \sum_{k\neq i,j} d_k) \sim \text{Dir}(\alpha_i, \alpha_j, \sum_{k\neq i,j} \alpha_k).$$

When $\boldsymbol{\alpha} = \mathbf{1}_n$, the density of $\text{Dir}(1, 1, n-2)$ is: $f(x, y) = (n-1)(n-2)(1-x-y)^{n-3}$. Note that $P(x < a, y < b) = P(x > a, y > b) - 1 + P(x < a) + P(y < b)$, so we can equivalently solve for the survival function $P(x > a, y > b)$. Let $a = b = q = 1 - (1-p)^{1/(n-1)}$. To calculate the survival function $P(x > q, y > q)$ we need to be careful about the limits of integration. The conditions are:

$$q < x < 1, \quad q < y < 1, \quad y < 1 - x, \quad x < 1 - q \quad \implies \quad q < y < 1 - x, \quad q < x < 1 - q.$$

The last condition implies $q < 1 - q \implies 1 - 2q > 0$. When this holds, the integral becomes:

$$\int_p^{1-q}\int_p^{1-x} (n-1)(n-2)(1-x-y)^{n-3}dydx = (n-1)(n-2)\int_p^{1-q} -\frac{(1-x-y)^{n-2}}{n-2}\Big|_{y=q}^{1-x} dx$$

$$= (n-1)\int_q^{1-q}(1-x-q)^{n-2}dx = -(n-1)\frac{(1-x-q)^{n-1}}{n-1}\Big|_{x=q}^{1-q} = (1-2q)^{n-1}.$$

Therefore, $P(x > q, y > q) = \max(1 - 2q, 0)^{n-1}$, which implies $P(\tilde{u}'_i < p, \tilde{u}'_j < p) = \max(0, 2p^{1/(n-1)} - 1)^{n-1}$. For the other copula uniforms $\tilde{\boldsymbol{u}} = 1 - \tilde{\boldsymbol{u}}'$, we have:

$$\begin{aligned} P(u_i < p, u_j < p) &= P(1 - \tilde{u}_i < p, 1 - \tilde{u}_j < p) = P(\tilde{u}_i > 1 - p, \tilde{u}_j > 1 - p) \\ &= 1 - P(\tilde{u}_i < 1 - p) - P(\tilde{u}_j < 1 - p) + P(\tilde{u}_i < 1 - p, \tilde{u}_j < 1 - p) \\ &= 2p - 1 + \max(0, 2(1 - p)^{1/(n-1)} - 1). \end{aligned}$$

Putting it all together:

$$P(\tilde{u}_i < p, \tilde{u}_j < p) = 2p - 1 + \max(0, 2(1 - p)^{1/(n-1)} - 1)^{n-1} \quad \implies \quad \rho = \frac{\max(0, 2(1 - p)^{\frac{1}{n-1}} - 1)^{n-1} - (1 - p)^2}{p(1 - p)}$$

$$P(\tilde{u}'_i < p, \tilde{u}'_j < p) = \max(0, 2p^{1/(n-1)} - 1)^{n-1} \quad \implies \quad \rho' = \frac{\max(0, 2p^{1/(n-1)} - 1)^{n-1} - p^2}{p(1 - p)}.$$

C. Dirichlet Copula Joint Density

To use the Dirichlet copula it is not necessary to know the joint densities of $\tilde{\boldsymbol{u}}$ or $\tilde{\boldsymbol{u}}'$, but it is possible to derive either of them using the multivariate change of variables theorem $p_U(u_1, \ldots, u_n) = |\det(J)| p_D(d_1, \ldots, d_n)$, where $|\det(J)|$ denotes the absolute value of the determinant of the Jacobian, whose elements are $J_{ij} = \partial d_i / \partial u_j$.

For ease of notation denote the copula sample $\boldsymbol{u} = \tilde{\boldsymbol{u}}'$, which uses the elementwise transformation $u_i = f(d_i) = (1 - d_i)^{n-1}$, with inverse $d_i = f^{-1}(u_i) = 1 - u_i^{1/(n-1)}$. Since d_i depends only on u_i the determinant of the Jacobian is diagonal and thus has a simple form:

$$|\det(J)| = \left| \prod_{i=1}^{n} \frac{\partial}{\partial u_i} \left(1 - u_i^{1/(n-1)}\right) \right| = (n - 1)^{-n} \prod_{i=1}^{n} u_i^{\frac{1}{n-1} - 1}.$$

For the Dirichlet density simplex condition, we have:

$$1 = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (1 - u_i^{1/(n-1)}) = n - \sum_{i=1}^{n} u_i^{1/(n-1)} \quad \implies \quad \sum_{i=1}^{n} u_i^{1/(n-1)} = n - 1$$

Putting it all together, the density is:

$$p(u_1, ..., u_n) = \frac{(n - 1)!}{(n - 1)^n} \prod_{i=1}^{n} u_i^{\frac{1}{n-1} - 1}, \quad \text{such that} \quad \sum_{i=1}^{n} u_i^{1/(n-1)} = n - 1.$$

D. Additional Results

Table 3. Test log likelihoods for ELBO optimized VAEs using different estimators. Results are reported on three datasets: Dynamic MNIST, Fashion MNIST, and Omniglot, and for 4, 6, 8, and 10 samples. Results are averaged over five runs, with the best performing methods in bold.

		SAMPLES	ARMS-D	ARMS-N	LOORF	DISARM	RELAX
DYNAMIC MNIST	LINEAR	4	**-111.57 ± 0.13**	**-111.47 ± 0.16**	-111.67 ± 0.04	-112.71 ± 0.07	-112.57 ± 0.34
		6	**-110.47 ± 0.02**	**-110.4 ± 0.12**	**-110.42 ± 0.03**	-111.58 ± 0.02	-110.94 ± 0.14
		8	**-109.73 ± 0.06**	-110.08 ± 0.06	-109.88 ± 0.07	-111.28 ± 0.16	-110.08 ± 0.10
		10	**-109.52 ± 0.07**	-109.61 ± 0.14	-109.61 ± 0.04	-110.56 ± 0.11	-109.64 ± 0.15
	NONLINR	4	-100.02 ± 0.24	-100.05 ± 0.17	**-99.63 ± 0.07**	-101.65 ± 0.32	-101.72 ± 0.18
		6	-99.67 ± 0.20	**-98.90 ± 0.01**	-99.21 ± 0.23	-100.42 ± 0.18	-100.31 ± 0.3
		8	**-98.89 ± 0.36**	**-98.86 ± 0.09**	-99.22 ± 0.41	-99.82 ± 0.25	-99.91 ± 0.07
		10	**-98.71 ± 0.22**	**-98.31 ± 0.44**	**-98.35 ± 0.54**	-99.73 ± 0.27	-99.79 ± 0.20
FASHION MNIST	LINEAR	4	**-254.65 ± 0.17**	**-254.76 ± 0.05**	-254.89 ± 0.07	-256.12 ± 0.04	-255.65 ± 0.08
		6	-254.04 ± 0.22	**-253.78 ± 0.08**	-254.14 ± 0.18	-255.11 ± 0.09	-254.40 ± 0.18
		8	**-253.43 ± 0.13**	**-253.24 ± 0.31**	-253.56 ± 0.14	-254.73 ± 0.07	**-253.49 ± 0.10**
		10	-253.38 ± 0.04	**-253.19 ± 0.17**	**-253.28 ± 0.01**	-253.87 ± 0.29	**-253.28 ± 0.13**
	NONLINR	4	**-238.25 ± 0.46**	**-238.36 ± 0.11**	**-238.42 ± 0.06**	-239.19 ± 0.09	-239.45 ± 0.15
		6	**-238.08 ± 0.25**	**-238.01 ± 0.15**	-238.31 ± 0.22	-238.59 ± 0.11	-238.87 ± 0.39
		8	-238.01 ± 0.14	**-237.71 ± 0.22**	-237.98 ± 0.18	-238.25 ± 0.18	-238.25 ± 0.20
		10	**-237.79 ± 0.16**	**-237.94 ± 0.02**	-238.19 ± 0.01	-238.24 ± 0.13	-238.02 ± 0.27
OMNIGLOT	LINEAR	4	**-118.61 ± 0.08**	-118.73 ± 0.03	**-118.63 ± 0.09**	-119.66 ± 0.26	-119.11 ± 0.06
		6	**-118.00 ± 0.02**	-118.03 ± 0.06	-118.12 ± 0.18	-118.87 ± 0.13	-118.24 ± 0.05
		8	**-117.60 ± 0.05**	-117.66 ± 0.12	-117.74 ± 0.10	-118.41 ± 0.10	-117.71 ± 0.02
		10	**-117.33 ± 0.14**	**-117.39 ± 0.04**	-117.51 ± 0.02	-118.06 ± 0.01	**-117.40 ± 0.05**
	NONLINR	4	-116.14 ± 0.55	**-115.88 ± 0.15**	-116.03 ± 0.39	-117.45 ± 0.24	-118.54 ± 0.51
		6	**-115.33 ± 0.12**	**-115.27 ± 0.24**	**-115.05 ± 0.35**	-116.45 ± 0.12	-118.05 ± 0.27
		8	**-114.71 ± 0.10**	-114.79 ± 0.05	**-114.52 ± 0.25**	-115.80 ± 0.17	-118.12 ± 0.15
		10	**-114.39 ± 0.16**	**-114.52 ± 0.40**	-114.60 ± 0.36	-115.15 ± 0.34	-118.50 ± 0.06

Table 4. Test log likelihoods for different estimators optimizing the multi sample bound, with the best performing methods in bold.

		SAMPLES	ARMS	DISARM	VIMCO
DYNAMIC MNIST	LINEAR	4	**-111.82 ± 0.09**	**-111.87 ± 0.07**	-112.52 ± 0.23
		6	**-111.27 ± 0.01**	**-111.15 ± 0.17**	-111.66 ± 0.44
		8	**-110.11 ± 0.18**	-110.50 ± 0.11	-110.57 ± 0.31
		10	**-109.74 ± 0.12**	-109.89 ± 0.25	-110.17 ± 0.28
	NONLINR	4	**-98.96 ± 0.03**	-99.30 ± 0.14	-98.98 ± 0.09
		6	**-97.51 ± 0.27**	-98.35 ± 0.20	**-97.65 ± 0.09**
		8	**-97.05 ± 0.20**	**-97.23 ± 0.31**	-97.33 ± 0.04
		10	**-96.08 ± 0.22**	-96.65 ± 0.11	-96.55 ± 0.02
FASHION MNIST	LINEAR	4	**-254.34 ± 0.10**	-254.48 ± 0.20	-254.86 ± 0.13
		6	**-253.51 ± 0.07**	**-253.41 ± 0.18**	-253.76 ± 0.32
		8	**-252.78 ± 0.13**	**-252.72 ± 0.16**	-252.90 ± 0.09
		10	-252.58 ± 0.07	**-251.90 ± 0.10**	-252.70 ± 0.15
	NONLINR	4	**-237.31 ± 0.18**	-238.16 ± 0.10	**-237.28 ± 0.25**
		6	**-236.22 ± 0.52**	**-236.62 ± 0.25**	**-236.33 ± 0.58**
		8	-235.64 ± 0.02	**-236.37 ± 0.04**	-235.96 ± 0.08
		10	**-235.27 ± 0.04**	-235.65 ± 0.27	-235.72 ± 0.06
OMNIGLOT	LINEAR	4	**-119.56 ± 0.09**	**-119.28 ± 0.31**	-120.71 ± 0.16
		6	-119.34 ± 0.37	**-119.09 ± 0.08**	-119.65 ± 0.09
		8	**-119.00 ± 0.11**	**-118.90 ± 0.16**	-119.08 ± 0.21
		10	-118.82 ± 0.10	**-118.57 ± 0.08**	-118.93 ± 0.17
	NONLINR	4	**-115.79 ± 0.35**	-116.91 ± 0.40	**-115.87 ± 0.35**
		6	**-114.56 ± 0.25**	-115.31 ± 0.42	**-114.68 ± 0.10**
		8	**-114.09 ± 0.09**	-114.45 ± 0.30	**-113.89 ± 0.22**
		10	-113.40 ± 0.05	**-114.19 ± 0.16**	-113.55 ± 0.05

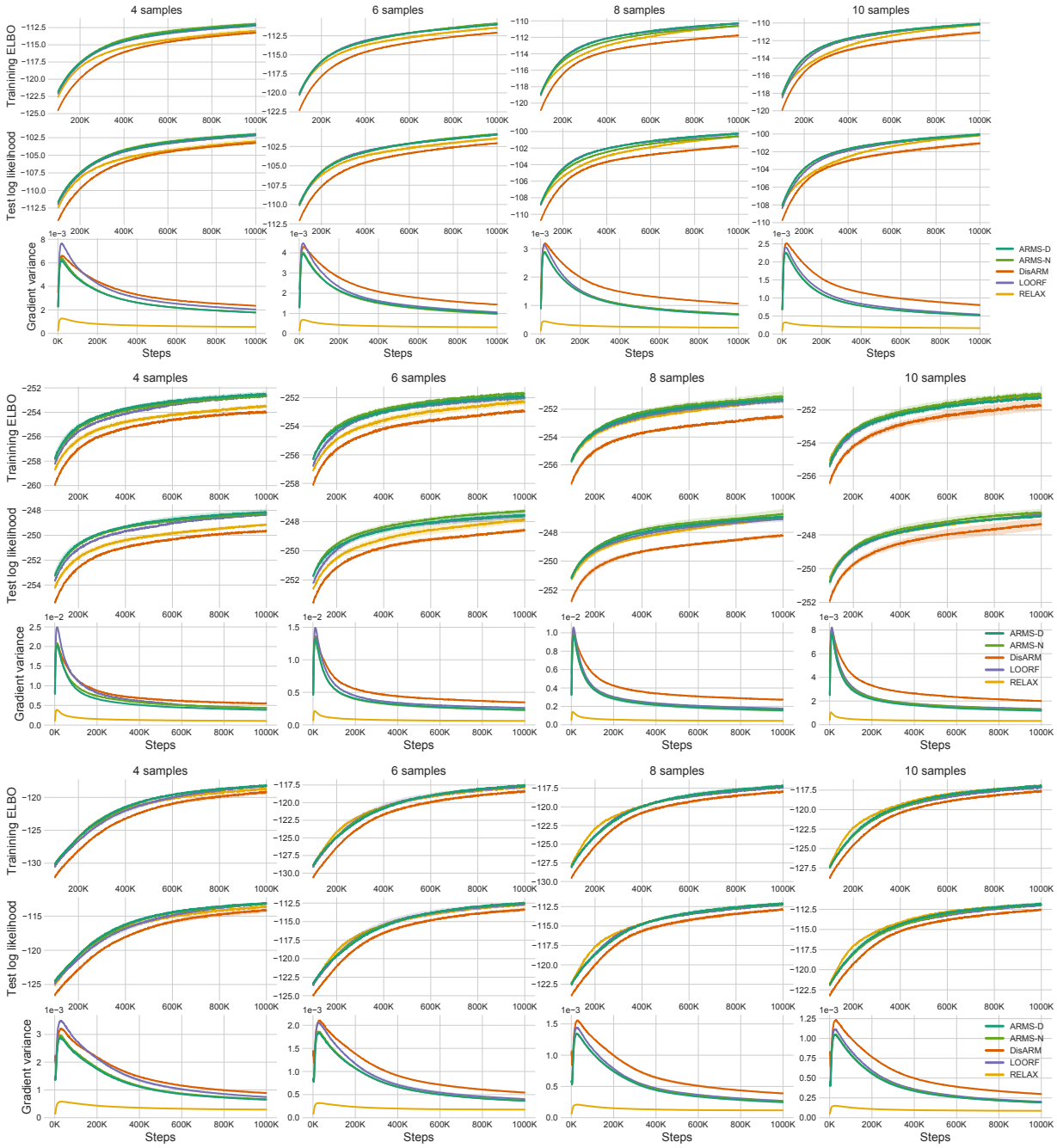

Figure 5. For each dataset, shown is training a *linear* discrete VAE using the *ELBO*. Each group of three rows, from top to bottom, represent Dynamic MNIST, Fashion MNIST, and Omniglot, respectively. Within each triplet of rows, they correspond to the training ELBO, test log likelihood, and the variance of the gradient updates averaged over all parameters. Columns correspond to $n \in \{4, 6, 8, 10\}$ samples used per step.

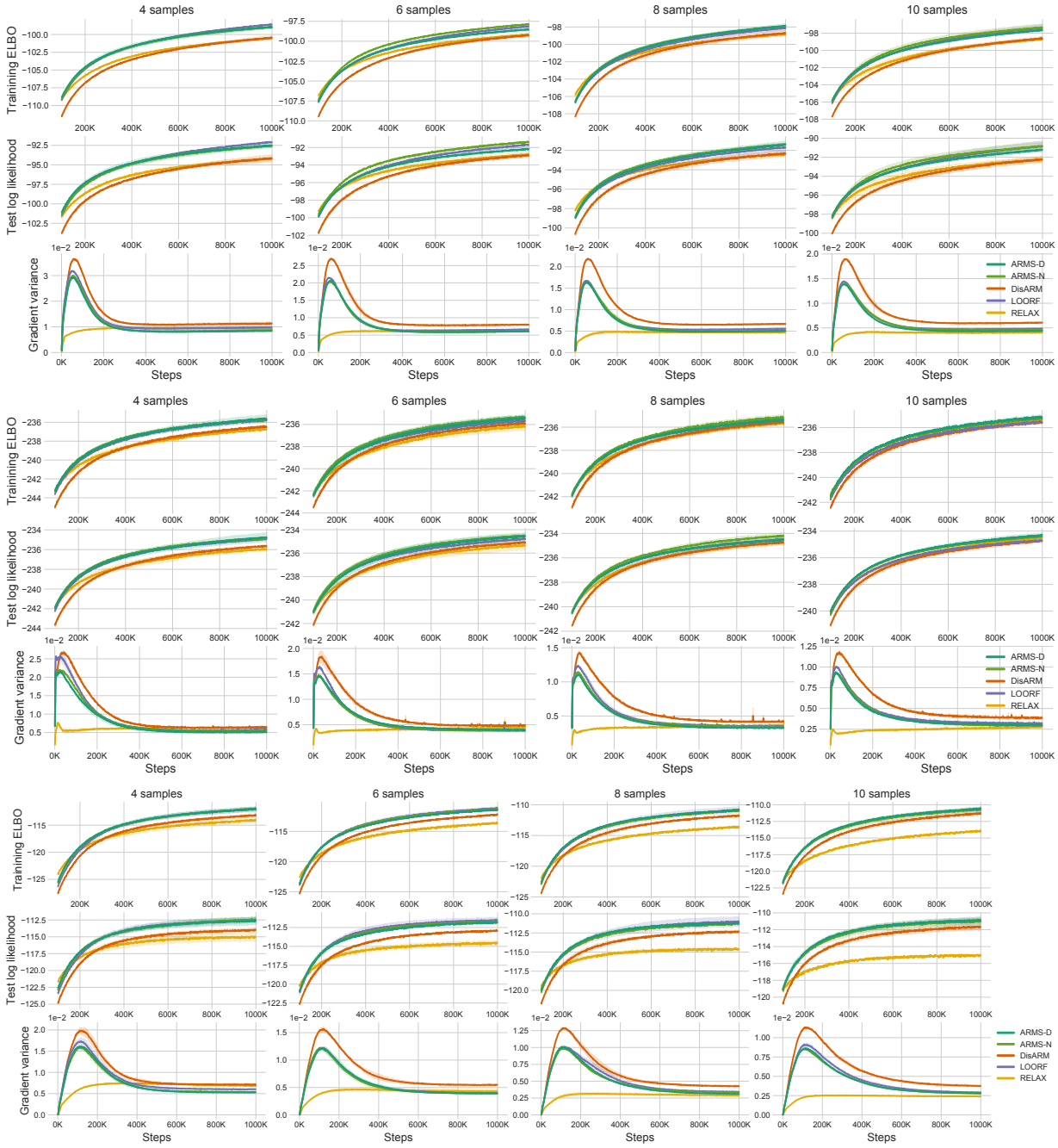

Figure 6. For each dataset, shown is training a *nonlinear* discrete VAE using the *ELBO*. Each group of three rows, from top to bottom, represent Dynamic MNIST, Fashion MNIST, and Omniglot, respectively. Within each triplet of rows, they correspond to the training ELBO, test log likelihood, and the variance of the gradient updates averaged over all parameters. Columns correspond to $n \in \{4, 6, 8, 10\}$ samples used per step.

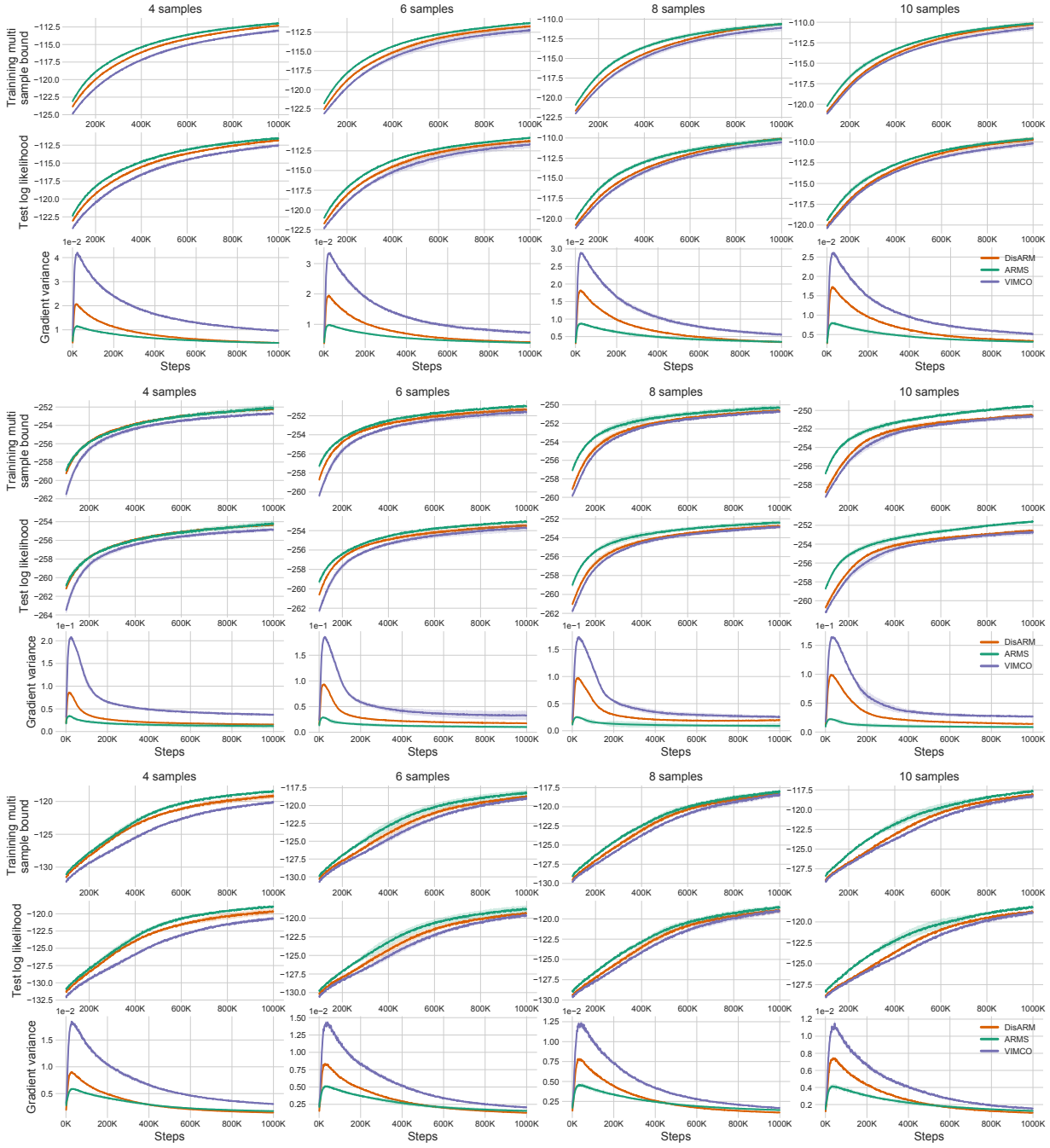

Figure 7. For each dataset, shown is training a *linear* discrete VAE using the *multi sample bound*. Each group of three rows, from top to bottom, represent Dynamic MNIST, Fashion MNIST, and Omniglot, respectively. Within each triplet of rows, they correspond to the training multi sample bound, test log likelihood, and the variance of the gradient updates averaged over all parameters. Columns correspond to $n \in \{4, 6, 8, 10\}$ samples used per step.

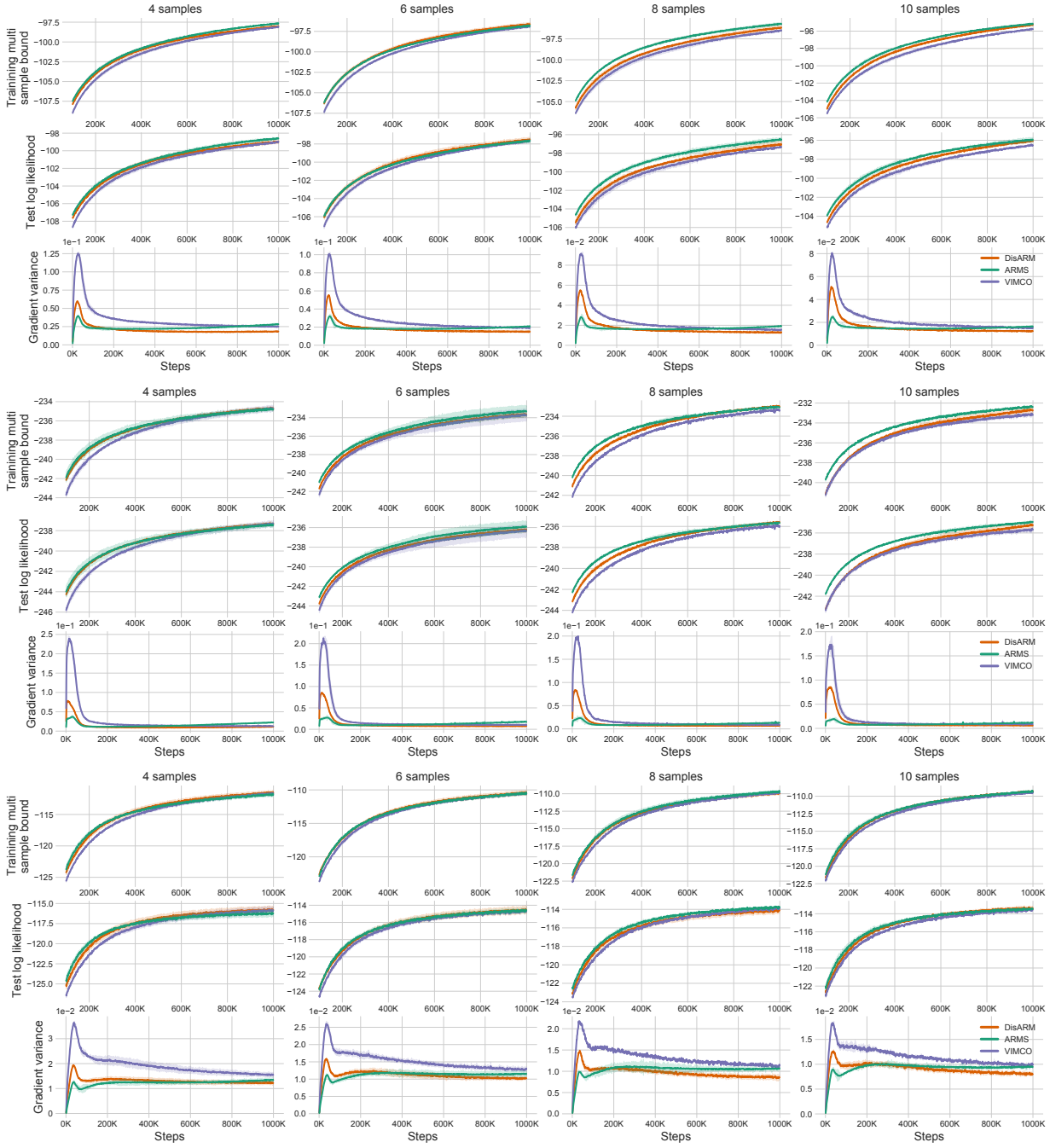

Figure 8. For each dataset, shown is training a *nonlinear* discrete VAE using the *multi sample bound*. Each group of three rows, from top to bottom, represent Dynamic MNIST, Fashion MNIST, and Omniglot, respectively. Within each triplet of rows, they correspond to the training multi sample bound, test log likelihood, and the variance of the gradient updates averaged over all parameters. Columns correspond to $n \in \{4, 6, 8, 10\}$ samples used per step.