APPENDIX

A ON THE UNIDENTIFIABILITY OF NONLINEAR ICA

The purpose of this section is to briefly review the proof of unidentifiablity of nonlinear ICA as [22]: In this section we assume the most general conventional form of nonlinear ICA where the generative model follows:

$$\mathbf{x} = \mathbf{f}(\boldsymbol{s}) \tag{22}$$

where $\boldsymbol{s}$ are the independent sources and $\mathbf{x}$ are mixed signals. In the following, we show how to construct a function $\mathbf{g} : R^n \to R^n$ so that the components $\mathbf{y} = \mathbf{g}(\mathbf{x})$ are independent. More importantly, we show that this construction is by no means unique.

A.1 EXISTENCE

The proposed method in [22] is a generalization of the famous Gram-Schmidt orthogonalization. Given m independent variables, $y_1, \ldots, y_m$ and a variable x, one constructs a new variable $y_{m+1} = g(y_1, \ldots, y_m, x)$ so that the set $y_1 s, \ldots, y_{m+1}$ is mutually independent. The construction process is defined recursively as follows. Assume we have m independent random variables $y_1, \ldots, y_m$ with uniform distribution in $[0, 1]^m$. x is any random variable and $a_1, \ldots, a_m, b$ are some nonrandom scalars. Next, we define

$$\begin{aligned} g\left(a_1, \ldots, a_m, b; p_{y,x}\right) &= p(x \leq b | y_1 = a_1, \ldots, y_m = a_m) \\ &= \frac{\int_{-\infty}^{b} p_{y,x}\left(a_1, \ldots, a_m, \xi\right) \mathrm{d}\xi}{p_y\left(a_1, \ldots, a_m\right)} \end{aligned} \tag{23}$$

Theorem 1 of [22] says that the random variable defined as $y_{m+1} = g(y_1, \ldots, y_m, x)$ is independent from the $y_1 \ldots, y_m$ and $y_1, \ldots, y_{m+1}$ are uniformly distributed in the unit cube $[0, 1]^{m+1}$.

A.2 NON-UNIQUENESS

In the previous section, it was shown that there exists a mapping $\mathbf{g}$ that transforms any random vector $\mathbf{x}$ into a uniformly distributed random vector $\mathbf{y} = \mathbf{g}(\mathbf{x})$. Here, we show that the construction of $\mathbf{g}$ is not unique and this non-Uniqueness can be caused by several factors.

- A linear transformation $\mathbf{x}'$ can precede the nonlinear map $\mathbf{f}$ and then compute the independent components $\mathbf{y}' = \mathbf{g}'(\mathbf{x}')$ where $\mathbf{g}'$ is computed as describe in the previous section. The new map $\mathbf{g}'$ gives a new decomposition of $\mathbf{x}$ into independent components $\mathbf{y}'$ which can not be trivially reduced to $\mathbf{y}$.
- An element-wise function $\mathbf{h}$ can apply on the independent sources $\boldsymbol{s}$ first to give new sources $\boldsymbol{s}'$ such that $s_i' = h_i(s_i)$. Constructing the solution g for these new scaled version of sources gives a new decomposition into independent components.
- Assume a class of measure-preserving automorphisms $\mathbf{h} : [0, 1]^n \to [0, 1]^n$. The mapping $\mathbf{h}$ does not change the probability distribution of a uniformly distributed random variable in n-dimensional hypercube. The composition $\mathbf{h} \circ \mathbf{g}$ gives another solution to nonlinear ICA. Therefore, the class of measure-preserving automorphisms gives a parameterization of the solutions to nonlinear ICA introducing a class of non-trivial indeterminacies.

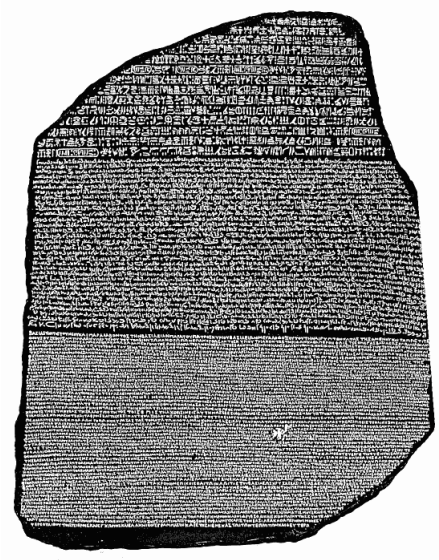

Figure 4: The Rosetta Stone, a stele found in 1799, inscribed with three versions of a decree issued at Memphis, Egypt in 196 BC. The top and middle texts are in Ancient Egyptian using hieroglyphic script and Demotic script, respectively, while the bottom is in Ancient Greek. (Source: Wikipedia)

If only independence among the components matters, it is possible to construct a mapping $\boldsymbol{y} = G(\boldsymbol{x})$ such that y_i is independent of y_j for $i \neq j$ and uniformely distributed in $[0, 1]^n$. This shows that at least one solution exists. The non-uniqueness of the solution can be shown by parameterising a class of infinitely many solutions. Once $\boldsymbol{y}$ is found with above conditions, any measure-preserving automorphism $f : [0, 1]^n \to [0, 1]^n$ can be used to parameterize G as $f \circ G$, suggesting that there

are infinitely many solutions to nonlinear ICA whose relations are nontrivial.

A.3 THE SCALAR INVERTIBLE FUNCTION GAUGE

Another indeterminacy is element-wise functions f_i applying on y_i which suggets another dimension of ambiguity. Non-Gaussianity cannot help here since we can construct any marginal distribution by combining the CDF of the observed variable with the inverse CDF of the target marginal distribution. This indeterminacy is in some sense unavoidable and is related to the fact that in linear ICA recovery of the sources is possible up to a scalar multiplicative ambiguity.

B WHY DOES CLASSIFICATION RESULT IN THE LOG RATIO?

Let us suppose that a variable X is drawn with equal probability from two distributions P_0 and P_1 with densities $p_0(x)$ and $p_1(x)$ respectively. We train a classifier $D : x \mapsto [0, 1]$ to estimate the posterior probability that a particular realization of X was drawn from P_0 with the cross entropy loss, i.e. the parameters of D are chosen to minimize

$$L(D) = \mathbb{E}_{X\sim P_0}\left[-\log D(X)\right] + \mathbb{E}_{X\sim P_1}\left[-\log(1 - D(X))\right].$$

As shown in, for instance, [14], the global optimum of this loss occurs when $D(x) = \frac{p_0(x)}{p_0(x)+p_1(x)}$, which can be rewritten as

$$D(x) = \frac{1}{1 + p_1(x)/p_0(x)} \tag{24}$$

$$= \frac{1}{1 + \exp(-\log(p_0(x)/p_1(x)))} \tag{25}$$

$$\tag{26}$$

Recall that in our setting, the function $r(x_1, x_2)$ is trained to classify between the two cases that (x_1, x_2) is drawn from the joint distribution $\mathbb{P}_{x_1,x_2}$ (*class* 0) or the product of marginals $\mathbb{P}_{x_1}\mathbb{P}_{x_2}$ (*class* 1). $r(x_1, x_2)$ is trained so that $\frac{1}{1+\exp(-r(x_1,x_2))}$ estimates the posterior probability of (x_1, x_2) belonging to class 0. By comparing to Equation 25, it can be seen that

$$\begin{aligned} r(x_1, x_2) &= \log\left(p(x_1, x_2)/p(x_1)p(x_2)\right) \\ &= \log p(x_1|x_2) - \log p(x_1) \\ &= \log p(x_2|x_1) - \log p(x_2) \end{aligned}$$

Note that in order for the classification trick of contrastive learning to be useful, the variables x_1 and x_2 cannot be deterministically related. If this is the case, the log-ratio is everywhere either 0 or ∞ and hence the learned features are not useful.

To see why this is the case, suppose that x_1, and x_2 are each N-dimensional vectors. If they are deterministically related, $p(x_1, x_2)$ puts mass on an N-dimensional submanifold of a $2N$-dimensional space. On the other hand, $p(x_1)p(x_2)$ will put mass on a $2N$-dim manifold since it is the product of two distributions each of which are N-dimensional.

In this case, the distributions $p(x_1, x_2)$ and $p(x_1)p(x_2)$ are therefore not absolutely continuous with respect to one another and thus the log-ratio is ill-defined: $p(x_1, x_2)/p(x_1)p(x_2) = \infty$ at any point (x_1, x_2) at which $p(x_1, x_2)$ puts mass and zero at points where $p(x_1)p(x_2)$ puts mass and $p(x_1, x_2)$ does not.

C THE SUFFICIENTLY DISTINCT VIEWS ASSUMPTION

We give the following two examples to provide intuition about the Sufficiently Distinct Views (SDV) assumption - one regarding a case in which it does not hold, and another one in which it does.

A simple case in which the assumption does not hold is when the conditional probability of $\boldsymbol{z}$ given $\boldsymbol{s}$ is Gaussian, as in

$$p(\boldsymbol{z}|\boldsymbol{s}) = \frac{1}{Z}\exp\left[-\sum_i (z_i - s_i)^2/(2\sigma_i^2)\right], \tag{27}$$

where Z is the normalization factor, $Z = (2\pi)^{n/2}\prod_i \sigma_i$. Since taking second derivatives of the log-probability with respect to s_i results in constants, it can be easily shown that there is no way to find $2D$ vectors $\boldsymbol{z}_j$, $j = 1, \ldots, 2D$, such that the corresponding $\boldsymbol{w}(\boldsymbol{s}, \boldsymbol{z}_j)$ (see Definition 1) are linearly independent.

The fact that the assumption breaks down in this case is reminiscent of the breakdown in the case of Gaussianity for linear ICA. Interestingly, in our work, the true latent

sources **are** allowed to be Gaussian. In fact, the distribution of $\boldsymbol{s}$ does not enter the expression above.

An example in which the SDV assumption does hold is a conditional pdf given by

$$p(\boldsymbol{z}|\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{s})} \exp\left[-\sum_i (z_i^2 s_i^2 + z_i^4 s_i^4)\right], \quad (28)$$

where $Z(\boldsymbol{s})$ is again a normalization function. Proving that this distribution satisfies the SDV assumption requires a few lines of computation. The idea is that $\boldsymbol{w}(\boldsymbol{s}, \boldsymbol{z})$ can be written as the product of a matrix and vector which are functions only of $\boldsymbol{s}$ and $\boldsymbol{z}$ respectively. Once written in this form, it is straightforward to show that the columns of the matrix are linearly independent for almost all values of $\boldsymbol{s}$ and that $2D$ linearly independent vectors can be realized by different choices of $\boldsymbol{z}$.

D PROOF OF THEOREM 1 AND COROLLARY 3

D.1 PROOF OF THEOREM 1

This proof is mainly inspired by the techniques employed by [23].

Proof. We have to show that, upon convergence, $h_i(\boldsymbol{x}_1)$ are s.t.

$$h_i(\boldsymbol{x}_1) \perp\!\!\!\perp h_j(\boldsymbol{x}_1), \forall i \neq j$$

We start by writing the difference in log-densities of the two classes:

$$\sum_i \psi_i(h_i(\boldsymbol{x}_1), \boldsymbol{x}_2) = \sum_i \alpha_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1), \boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) + - \sum_i \delta_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2))$$

We now make the change of variables

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{h}(\boldsymbol{x}_1) \\ \boldsymbol{v}(\boldsymbol{y}) &= \boldsymbol{f}_1^{-1}(\boldsymbol{h}^{-1}(\boldsymbol{y})) \\ \boldsymbol{t} &= \boldsymbol{f}_2^{-1}(\boldsymbol{x}_2)) \end{aligned}$$

and rewrite the first equation in the following form:

$$\sum_i \psi_i(y_i, \boldsymbol{x}_2) = \sum_i \alpha_i(v_i(\boldsymbol{y}), t_i) \quad (29)$$

$$- \sum_i \delta_i(t_i) \quad (30)$$

We take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of the LHS and RHS of equation 38. Adopting the conventions in 9 and 10 and

$$v_i^j(\boldsymbol{y}) = \partial v_i(\boldsymbol{y}) / \partial y_j \quad (31)$$

$$v_i^{jj'}(\boldsymbol{y}) = \partial^2 v_i(\boldsymbol{y}) / \partial y_j \partial y_{j'}\,, \quad (32)$$

we have

$$\sum_i \alpha_i''(v_i(\boldsymbol{y}), t_i) v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y}) + \alpha_i'(v_i(\boldsymbol{y}), t_i) v^{jj'}(\boldsymbol{y}) = 0\,,$$

where taking derivative w.r.t. y_j and y_j' for $j \neq j'$ makes LHS equal to zero, since the LHS has functions which depend only one y_i each. If we now rearrange our variables by defining vectors $\boldsymbol{a}_i(\boldsymbol{y})$ collecting all entries $v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, and vectors $\boldsymbol{b}_i(\boldsymbol{y})$ with the variables $v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, the above equality can be rewritten as

$$\sum_i \alpha_i''(v_i(\boldsymbol{y}), t_i) \boldsymbol{a}_i(\boldsymbol{y}) + \alpha_i'(v_i(\boldsymbol{y}), t_i)) \boldsymbol{b}_i(\boldsymbol{y}) = 0\,.$$

The above expression can be recast in matrix form,

$$\boldsymbol{M}(\boldsymbol{y}) \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = 0\,,$$

where $\boldsymbol{M}(\boldsymbol{y}) = (\boldsymbol{a}_1(\boldsymbol{y}), \ldots, \boldsymbol{a}_n(\boldsymbol{y}), \boldsymbol{b}_1(\boldsymbol{y}), \ldots, \boldsymbol{b}_n(\boldsymbol{y}))$ and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = (\alpha_1'', \ldots, \alpha_n'', \alpha_1', \ldots, \alpha_n')$. $\boldsymbol{M}(\boldsymbol{y})$ is therefore a $n(n-1)/2 \times 2n$ matrix, and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t})$ is a $2n$ dimensional vector.

To show that $\boldsymbol{M}(\boldsymbol{y})$ is equal to zero, we invoke the SDV assumption. This implies the existence of $2n$ linearly independent $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$. It follows that

$$\boldsymbol{M}(\boldsymbol{y})[\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_1), \ldots, \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_{2n})] = 0\,,$$

and hence $\boldsymbol{M}(\boldsymbol{y})$ is zero by elementary linear algebraic results. It follows that $v_i^j(\boldsymbol{y}) \neq 0$ for at most one value of j, since otherwise the product of two non-zero terms would appear in one of the entries of $\boldsymbol{M}(\boldsymbol{y})$, thus rendering it non-zero. Thus v_i is a function only of one y_j.

Observe that $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{s}$. We have just proven that $v_i(y_{\pi(i)}) = s_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(\boldsymbol{x}_1) = y_{\pi(i)} = v_i^{-1}(s_i)$ and hence the components of $\boldsymbol{h}(\boldsymbol{x}_1)$ recover the components of $\boldsymbol{s}$ up to the invertible component-wise ambiguity given by $\boldsymbol{v}$, and the permutation ambiguity.

□

D.2 PROOF OF COROLLARY 3

Proof. This follows exactly by repeating the proof of Theorem 1 where the roles of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are exchanged and the regression function in the statement of the corollary is used. □

E PROOF OF THEOREMS 4 AND 5

Theorem 4 is a special case of Theorem 5 by considering the case $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1) = \boldsymbol{s}$. We therefore prove only the more general Theorem 5.

Proof. We have to show that, upon convergence, $h_i(\boldsymbol{x}_1)$ and $k_i(\boldsymbol{x}_2)$ are such that

$$h_{1,i}(\boldsymbol{x}_1) \perp\!\!\!\perp h_{1,j}(\boldsymbol{x}_1), \forall i \neq j \tag{33}$$
$$h_{2,i}(\boldsymbol{x}_2) \perp\!\!\!\perp h_{2,j}(\boldsymbol{x}_2), \forall i \neq j \tag{34}$$
$$h_{1,i}(\boldsymbol{x}_1) \perp\!\!\!\perp h_{2,j}(\boldsymbol{x}_2), \forall i \neq j. \tag{35}$$

We start by exploiting Equations 14 and 15 to write the difference in log-densities of the two classes

$$\sum_i \psi_i(h_{1,i}(\boldsymbol{x}_1), h_{2,i}(\boldsymbol{x}_2))$$
$$= \sum_i \eta_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1), \boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) - \sum_i \theta_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1)) \tag{36}$$
$$= \sum_i \lambda_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2), \boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1)) - \sum_i \mu_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) \tag{37}$$

We now make the change of variables

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{h}_1(\boldsymbol{x}_1) \\ \boldsymbol{t} &= \boldsymbol{h}_2(\boldsymbol{x}_2) \\ \boldsymbol{v}(\boldsymbol{y}) &= \boldsymbol{f}_1^{-1}(\boldsymbol{h}_1^{-1}(\boldsymbol{y})) \\ \boldsymbol{u}(\boldsymbol{t}) &= \boldsymbol{f}_2^{-1}(\boldsymbol{h}_2^{-1}(\boldsymbol{t})) \end{aligned}$$

and rewrite equation 36 in the following form:

$$\sum_i \psi_i(y_i, t_i)$$
$$= \sum_i \eta_i(v_i(\boldsymbol{y}), u_i(\boldsymbol{t})) - \sum_i \theta_i(v_i(\boldsymbol{y})) \tag{38}$$

We first want to prove the condition in Equation 33. We will show this is true by proving that

$$v_i(\boldsymbol{y}) \equiv v_i(y_{\pi(i)}) \tag{39}$$

for some permutation of the indices π with respect to the indexing of the sources $\boldsymbol{s} = (s_1, \ldots, s_D)$.

We take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of the LHS and RHS of equation 38, yielding

$$\sum_i \eta_i''(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$$
$$+ \sum_i \eta_i'(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))v^{jj'}(\boldsymbol{y}) = 0$$

If we now rearrange our variables by defining vectors $\boldsymbol{a}_i(\boldsymbol{y})$ collecting all entries $v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, and vectors $\boldsymbol{b}_i(\boldsymbol{y})$ with the variables $v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, the above equality can be rewritten as

$$\sum_i \eta_i''(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))\boldsymbol{a}_i(\boldsymbol{y})$$
$$+ \eta_i'(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))\boldsymbol{b}_i(\boldsymbol{y}) = 0\,.$$

Again following [23], we recast the above formula in matrix form,

$$\boldsymbol{M}(\boldsymbol{y})\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = 0\,, \tag{40}$$

where $\boldsymbol{M}(\boldsymbol{y}) = (\boldsymbol{a}_1(\boldsymbol{y}), \ldots, \boldsymbol{a}_n(\boldsymbol{y}), \boldsymbol{b}_1(\boldsymbol{y}), \ldots, \boldsymbol{b}_n(\boldsymbol{y}))$ and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = (\eta_1'', \ldots, \eta_n'', \eta_1', \ldots, \eta_n')$. $\boldsymbol{M}(\boldsymbol{y})$ is therefore a $n(n-1)/2 \times 2n$ matrix, and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t})$ is a $2n$ dimensional vector.

To show that $\boldsymbol{M}(\boldsymbol{y})$ is equal to zero, we invoke the SDV assumption on $\boldsymbol{\eta}$. This implies the existence of $2n$ linearly independent $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$. It follows that

$$\boldsymbol{M}(\boldsymbol{y})[\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_1), \ldots, \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_{2n})] = 0\,,$$

and hence $\boldsymbol{M}(\boldsymbol{y})$ is zero by elementary linear algebraic results. It follows that $v_i^j(\boldsymbol{y}) \neq 0$ for at most one value of j, since otherwise the product of two non-zero terms would appear in one of the entries of $\boldsymbol{M}(\boldsymbol{y})$, thus rendering it non-zero. Thus v_i is a function only of one $y_j = y_{\pi(i)}$.

Observe that $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{s}$. We have just proven that $v_i(y_{\pi(i)}) = s_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(\boldsymbol{x}_1) = y_{\pi(i)} = v_i^{-1}(s_i)$ and hence the components of $\boldsymbol{h}(\boldsymbol{x}_1)$ recover the components of $\boldsymbol{s}$ up to the invertible component-wise ambiguity given by $\boldsymbol{v}$, and the permutation ambiguity.

For the condition in Equation 34, we need

$$u_i(\boldsymbol{t}) \equiv u_i(t_{\tilde{\pi}(i)})\,, \tag{41}$$

where the permutation $\tilde{\pi}$ doesn't need to be equal to π. By symmetry, exactly the same argument as used

to prove the condition in Equation 39 holds, by replacing $(\boldsymbol{v}, \boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\theta})$ with $(\boldsymbol{u}, \boldsymbol{t}, \boldsymbol{\lambda}, \boldsymbol{\mu})$, noting that the SDV assumption is also assumed for $\boldsymbol{\lambda}$.

We have shown that $\boldsymbol{y} = \boldsymbol{h}_1(\boldsymbol{x}_1)$ and $\boldsymbol{t} = \boldsymbol{h}_2(\boldsymbol{x}_2)$ estimate $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)$ up to two different gauges of all possible scalar invertible functions.

A remaining ambiguity could be that the two representations might be misaligned; that is, defining $\boldsymbol{z}_1 = \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{z}_2 = \boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)$, while

$$z_{1,i} \perp\!\!\!\perp z_{2,j} \forall i \neq j \tag{42}$$

we might have

$$y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \forall i \neq j\,,$$

where $\pi(i)$, $\tilde{\pi}(i)$ are two different permutations of the indices $i = 1, \ldots, n$. We want to show that this ambiguity is also resolved; that means, our goal is to show that

$$y_i \perp\!\!\!\perp t_j, \quad \forall i \neq j \tag{43}$$

We recall that, by definition, we have $v_i(y_{\pi(i)}) = z_{1,i}$ and $u_j(t_{\tilde{\pi}(j)}) = z_{2,j}$. Then, due to equation 42,

$$v_i(y_{\pi(i)}) \perp\!\!\!\perp u_j(t_{\tilde{\pi}(j)}) \;\; \forall i \neq j \tag{44}$$
$$\implies y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \;\; \forall i \neq j \tag{45}$$
$$\implies y_i \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(j)} \;\; \forall i \neq j\,, \tag{46}$$

where the implication 44-45 follows from invertibility of v_i and u_j, and the implication 45-46 follows from considering that, given that we know 45, we can define $l = \pi(j)$ and $k = \pi(i)$ and have

$$y_k \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(l)} \;\; \forall k \neq l.$$

Define

$$\tau = \tilde{\pi} \circ \pi^{-1}$$

and note that it is a permutation. Then

$$y_i \perp\!\!\!\perp t_{\tau(j)} \forall i \neq j \tag{47}$$

Fix any particular i. Our goal is to show that for any $j \neq i$ the independence relation in Equation 43 holds. There are two possibilities:

i $\tau(i) = i$

ii $\tau(i) \neq i$

In the first case, τ restricted to the set $\{1, \ldots, D\} \setminus \{i\}$ is still a permutation, and thus considering the independences of Equation 47 for all $j \neq i$ implies each of the independences of Equation 43 and we are done.

Let us consider the second case. Then,

$$\exists l \in \{1, \ldots, D\} \setminus \{i\} \text{ s.t. } l = \tau(i)\,.$$

We then need to prove

$$y_i \perp\!\!\!\perp t_l\,, \tag{48}$$

which is the only independence implied by Equation 43 which is not implied by Equation 47.

In order to do so, we rewrite equation 38, yielding

$$\begin{aligned}&\sum_m \psi_m(y_m, t_m) \\ &= \sum_m \eta_m(v_m(y_{\pi(m)}), u_m(t_{\tilde{\pi}(m)})) - \sum_m \theta_i(v_m(y_{\pi(m)}))\end{aligned} \tag{49}$$

We now take derivative with respect to y_i and t_l in 48; noting that $\tilde{\pi}^{-1}(l) = \pi^{-1}(i)$, we get

$$\begin{aligned}0 =& \frac{\partial^2}{\partial v_{\pi^{-1}(i)} \partial u_{\pi^{-1}(i)}} \eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) \\ &\times \frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) \frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l)\end{aligned} \tag{50}$$

Since $v_{\pi^{-1}(i)}(y_i)$ is a smooth and invertible function of its argument, the set of y_i such that $\frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) = 0$ has measure zero. Similarly, $\frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l) = 0$ on a set of measure zero.

It therefore follows that

$$\frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) \frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l) \neq 0$$

almost everywhere and hence that

$$\frac{\partial^2}{\partial v_{\pi^{-1}(i)} \partial u_{\pi^{-1}(i)}} \eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) = 0\,. \tag{51}$$

almost everywhere. We can thus conclude that

$$\begin{aligned}&\eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) = \\ &\eta^y_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i)) + \eta^t_{\pi^{-1}(i)}(u_{\pi^{-1}(i)}(t_l))\end{aligned}$$

This in turn implies that, for some functions A and B, we can write

$$\begin{aligned}&\log p(z_{1,\pi^{-1}(i)} | z_{2,\pi^{-1}(i)}) - \log p(z_{1,\pi^{-1}(i)}) \\ &= A(v_{\pi^{-1}(i)}(y_i)) + B(u_{\pi^{-1}(i)}(t_l))\end{aligned}$$

and therefore

$$\log p(z_{1,\pi^{-1}(i)}, z_{2,\pi^{-1}(i)}) = C(v_{\pi^{-1}(i)}(y_i)) + D(u_{\pi^{-1}(i)}(t_l))$$

for some functions C and D. This decomposition of the log-pdf implies

$$\begin{aligned} & z_{1,\pi^{-1}(i)} \perp\!\!\!\perp z_{2,\pi^{-1}(i)} \\ \implies & z_{1,\pi^{-1}(i)} \perp\!\!\!\perp z_{2,\tilde{\pi}^{-1}(l)} \\ \implies & v_{\pi^{-1}(i)}(y_i) \perp\!\!\!\perp u_{\tilde{\pi}^{-1}(l)}(t_l) \\ \implies & y_i \perp\!\!\!\perp t_l \,, \end{aligned}$$

where the last implication holds due to invertibility of $v_{\pi^{-1}(i)}$ and $u_{\tilde{\pi}^{-1}(l)}$.

We have thus concluded the proof.

□

F PROOF OF COROLLARY 6

Proof. Denoting by $\boldsymbol{d}_1^{(k)}$ the component-wise invertible ambiguity up to which $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})$ is recovered, we have that

$$\inf_{\boldsymbol{e}\in\boldsymbol{E}} \mathbb{E}_{\boldsymbol{x}_1}\left[\left\|\boldsymbol{s} - \boldsymbol{e}(\boldsymbol{h}_1^{(k)}(\boldsymbol{x}_1))\right\|_2^2\right] \tag{52}$$

$$= \inf_{\boldsymbol{e}\in\boldsymbol{E}} \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \boldsymbol{e}\circ\boldsymbol{d}_1^{(k)}\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{53}$$

$$= \inf_{\tilde{\boldsymbol{e}}\in\boldsymbol{E}} \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \tilde{\boldsymbol{e}}\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{54}$$

$$\leq \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \boldsymbol{e}^*\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{55}$$

The lower bound holds for any $\boldsymbol{e}^* \in \boldsymbol{E}$ by definition of infimum and in particular for $\boldsymbol{e}^* = \boldsymbol{g}_1|_{\boldsymbol{n}=0}^{-1}$, the existence of which is guaranteed by the assumptions on $\boldsymbol{g}_1$. Taking a Taylor expansion of $\boldsymbol{e}^*\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})$ around $\boldsymbol{n}_1^{(k)} = 0$ yields

$$\mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \boldsymbol{e}^*\circ\boldsymbol{g}_1(\boldsymbol{s}, 0) + \frac{\partial \boldsymbol{e}^*}{\partial \boldsymbol{g}_1}\frac{\partial \boldsymbol{g}_1(\boldsymbol{s}, 0)}{\partial \boldsymbol{n}_1^{(k)}}\cdot\boldsymbol{n}_1^{(k)} + \mathcal{O}(\|\boldsymbol{n}_1^{(k)}\|^2)\right\|_2^2\right]$$

$$= \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\frac{\partial \boldsymbol{e}^*}{\partial \boldsymbol{g}_1}\frac{\partial \boldsymbol{g}_1(\boldsymbol{s}, 0)}{\partial \boldsymbol{n}_1^{(k)}}\cdot\boldsymbol{n}_1^{(k)} + \mathcal{O}(\|\boldsymbol{n}_1^{(k)}\|^2)\right\|_2^2\right]$$

$\longrightarrow 0$ as $k \longrightarrow \infty$

where the last equality follows from fact that $\boldsymbol{e}^* = \boldsymbol{g}|_{\boldsymbol{n}=0}^{-1}$ and the convergence follows from the fact that $\boldsymbol{n}_1^{(k)} \longrightarrow 0$ as $k \to \infty$. □

G PROOF OF LEMMA 7

We will make crucial use of *Kolmogorov's strong law*:

Theorem 9. *Suppose that X_n is a sequence of independent (but not necessarily identically distributed) random variables with*

$$\sum_{n=1}^{\infty}\frac{1}{n^2}\mathrm{Var}[X_n] < \infty$$

Then,

$$\frac{1}{N}\sum_{n=1}^{N} X_n - \mathbb{E}[X_n] \xrightarrow{a.s.} 0$$

Fix $\boldsymbol{s}$ and consider $\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ as a random variable with randomness induced by $\boldsymbol{n}$. We will show that for almost all $\boldsymbol{s}$ this converges $\boldsymbol{n}$-almost surely to a constant, and hence $\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ converges almost surely to a function of $\boldsymbol{s}$.

The law of total expectation says that

$$\begin{aligned} & \mathrm{Var}_{\boldsymbol{s},\boldsymbol{n}_i}[\boldsymbol{e}_i\circ\boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i)] \\ & = \mathbb{E}_{\boldsymbol{s}}\left[V_i(\boldsymbol{s})\right] + \mathrm{Var}_{\boldsymbol{s}}\left[\mathbb{E}_{\boldsymbol{n}_i}[\boldsymbol{e}_i\circ\boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i)]\right] \\ & \geq \mathbb{E}_{\boldsymbol{s}}\left[V_i(\boldsymbol{s})\right]. \end{aligned}$$

Since by assumption $\mathrm{Var}_{\boldsymbol{s},\boldsymbol{n}_i}[\boldsymbol{e}_i\circ\boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i)] \leq K$, we have that

$$\mathbb{E}_{\boldsymbol{s}}\left[\sum_{i=1}^{\infty}\frac{V_i(\boldsymbol{s})}{i^2}\right] \leq \frac{K\pi^2}{6}$$

and therefore $\sum_{i=1}^{\infty}\frac{V_i(\boldsymbol{s})}{i^2} < \infty$ with probability 1 over $\boldsymbol{s}$, else the expectation above would be unbounded since $V_i(\boldsymbol{s}) \geq 0$.

We have further that for almost all $\boldsymbol{s}$,

$$\Omega_{\boldsymbol{e}}(\boldsymbol{s}) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} E_{\boldsymbol{e}_i}(\boldsymbol{s})$$

exists. Therefore, for almost all s the conditions of Kolmogorov's strong law are met by $\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ and so

$$\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}) - \mathbb{E}_{\boldsymbol{n}}[\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n})] \overset{\boldsymbol{n}-a.s.}{\longrightarrow} 0$$

Since $\mathbb{E}_{\boldsymbol{n}}[\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n})] \overset{\boldsymbol{n}-a.s.}{\longrightarrow} \Omega_{\boldsymbol{e}}(\boldsymbol{s})$, it follows that

$$\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}) \overset{\boldsymbol{n}-a.s.}{\longrightarrow} \Omega_{\boldsymbol{e}}(\boldsymbol{s}).$$

Since this holds with probability 1 over $\boldsymbol{s}$, we have that

$$\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}) \overset{\boldsymbol{n}-a.s.}{\longrightarrow} \Omega_{\boldsymbol{e}}(\boldsymbol{s}).$$

It follows that we can write

$$\begin{aligned} R_{\boldsymbol{e},i}^N(\boldsymbol{s}, \boldsymbol{n}) &= \boldsymbol{e}_i\circ\boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i) - \Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}) \\ &\xrightarrow{a.s.} R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) := \boldsymbol{e}_i\circ\boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i) - \Omega_{\boldsymbol{e}}(\boldsymbol{s}) \end{aligned}$$

H PROOF OF THEOREM 8

We will begin by showing that if $K \geq \text{Var}(\boldsymbol{s}) + C$ then $\{\boldsymbol{k}_i^{-1}\} \in \mathcal{G}_K$.

For $\boldsymbol{e}_i = \boldsymbol{k}_i^{-1}$, we have that

$$\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{s} + \boldsymbol{n}_i \xrightarrow{a.s.} \boldsymbol{s} = \Omega_{\boldsymbol{e}}^N(\boldsymbol{s})$$
$$R_i^N = \boldsymbol{s} + \boldsymbol{n}_i - \Omega_{\boldsymbol{e}}(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{a.s.} \boldsymbol{n}_i = R_{\boldsymbol{e},i}(\boldsymbol{n}_i)$$

where the convergences follow from application of Kolmogorov's strong law, using the fact that $\text{Var}(\boldsymbol{n}_i) \leq C$ for all i. Satisfaction of condition 17 follows from the fact that $\text{Var}_{\boldsymbol{s},\boldsymbol{n}_i}(\boldsymbol{s} + \boldsymbol{n}_i) \leq C + \text{Var}(\boldsymbol{s}) \leq K$. Since $\boldsymbol{s}$ is a well-defined random variable, $\Omega_{\boldsymbol{e}}(\boldsymbol{s}) < \infty$ with probability 1, satisfying condition 18. It follows from the mutual independence of $\boldsymbol{n}_i$ and $\boldsymbol{n}_j$ that $R_{\boldsymbol{e},i}$ and $R_{\boldsymbol{e},j}$ satisfy condition 19. Condition 20 follows from the fact that $\mathbb{E}[\boldsymbol{n}_i] = 0$ Condition 21 follows from $R_{\boldsymbol{e},i}$ being constant as a function of $\boldsymbol{s}$.

It therefore follows that $\{\boldsymbol{k}_i^{-1}\} \in \mathcal{G}_K$ for K sufficiently large.

We will next show that if $\{\boldsymbol{e}_i\} \in \mathcal{G}_K$ then there exist a matrix $\boldsymbol{\alpha}$ and vector $\boldsymbol{\beta}$ such that $\boldsymbol{e}_i = \boldsymbol{\alpha}\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}$ for all i. Since $\boldsymbol{e}_i$ acts coordinate-wise, it moreover follows that $\boldsymbol{\alpha}$ is diagonal.

First, we will show that each $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, i.e. there exist potentially different $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ such that $\boldsymbol{e}_i = \boldsymbol{\alpha}_i\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}_i$ for each i.

Then we will show that we must have $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for all i, j.

To see that $\boldsymbol{e}_i$ is affine, we make use of that fact that $R_{\boldsymbol{e},i}$ is constant as a function of $\boldsymbol{s}$. It follows that for any x and y

$$\begin{aligned}
\boldsymbol{e}_i \circ \boldsymbol{k}_i(x+y) &= R_{\boldsymbol{e},i}(x) + \Omega_{\boldsymbol{e}}(y) \\
&= R_{\boldsymbol{e},i}(x) + \Omega_{\boldsymbol{e}}(0) + R_{\boldsymbol{e},i}(0) + \Omega_{\boldsymbol{e}}(y) \\
&\quad - (R_{\boldsymbol{e},i}(0) + \Omega_{\boldsymbol{e}}(0)) \\
&= \boldsymbol{e}_i \circ \boldsymbol{k}_i(x) + \boldsymbol{e}_i \circ \boldsymbol{k}_i(y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)
\end{aligned}$$

It therefore follows that $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, since if we define

$$\begin{aligned}
L(x+y) &= \boldsymbol{e}_i \circ \boldsymbol{k}_i(x+y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0) \\
&= (\boldsymbol{e}_i \circ \boldsymbol{k}_i(x) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)) \\
&\quad + (\boldsymbol{e}_i \circ \boldsymbol{k}_i(y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)) \\
&= L(x) + L(y)
\end{aligned}$$

then L is linear and we can write $\boldsymbol{e}_i \circ \boldsymbol{k}_i(x)$ as the sum of a linear function and a constant:

$$\boldsymbol{e}_i \circ \boldsymbol{k}_i(x) = L(x) + \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)$$

Thus $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, and we have some (diagonal) matrix $\boldsymbol{\alpha}_i$ and vector $\boldsymbol{\beta}_i$ such that for any x

$$\begin{aligned}
&\boldsymbol{e}_i \circ \boldsymbol{k}_i(x) = \boldsymbol{\alpha}_i x + \boldsymbol{\beta}_i \\
\implies &\boldsymbol{e}_i(x) = \boldsymbol{\alpha}_i\boldsymbol{k}_i^{-1}x + \boldsymbol{\beta}_i.
\end{aligned}$$

Next we show that for the set of $\{\boldsymbol{e}_i = \boldsymbol{\alpha}_i\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}_i\}$, it must be the case that each $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$.

Observe that

$$\begin{aligned}
\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}) &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\alpha}_i\boldsymbol{s} + \boldsymbol{\alpha}_i\boldsymbol{n}_i + \boldsymbol{\beta}_i \\
&= \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\alpha}_i\right)\boldsymbol{s} + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\beta}_i + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{n}_i \\
\mathbb{E}_{\boldsymbol{n}}[\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n})] &= \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\alpha}_i\right)\boldsymbol{s} + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\beta}_i
\end{aligned}$$

Define

$$\boldsymbol{\alpha} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\alpha}_i$$
$$\boldsymbol{\beta} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\beta}_i$$

which exist by the assumption that $\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ converges as $N \to \infty$. Thus

$$\begin{aligned}
\Omega_{\boldsymbol{e}}(\boldsymbol{s}) &= \boldsymbol{\alpha}\boldsymbol{s} + \boldsymbol{\beta} \\
R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) &= (\boldsymbol{\alpha}_i - \boldsymbol{\alpha})\boldsymbol{s} + \boldsymbol{\alpha}_i\boldsymbol{n}_i + \boldsymbol{\beta}_i - \boldsymbol{\beta}
\end{aligned}$$

Now, suppose that there exist i and j such that such that $\boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}_j$. It follows that

$$\begin{aligned}
R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) &= (\boldsymbol{\alpha}_i - \boldsymbol{\alpha})\boldsymbol{s} + \boldsymbol{\alpha}_i\boldsymbol{n}_i + \boldsymbol{\beta}_i - \boldsymbol{\beta} \\
R_{\boldsymbol{e},j}(\boldsymbol{s}, \boldsymbol{n}_j) &= (\boldsymbol{\alpha}_j - \boldsymbol{\alpha})\boldsymbol{s} + \boldsymbol{\alpha}_j\boldsymbol{n}_j + \boldsymbol{\beta}_j - \boldsymbol{\beta}
\end{aligned}$$

There are two cases. If $\boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}$, then $R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i)$ is not a constant function of $\boldsymbol{s}$. But if $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}$, then $\boldsymbol{\alpha}_j \neq \boldsymbol{\alpha}$ and so $R_{\boldsymbol{e},j}(\boldsymbol{s}, \boldsymbol{n}_j)$ is not a constant function of $\boldsymbol{s}$. This is a contradiction, and so $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ for all i, j.

Suppose similarly that there exist $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$. If $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}$, then $\mathbb{E}[R_{\boldsymbol{e},i}(\boldsymbol{n}_i)] = \boldsymbol{\beta}_i - \boldsymbol{\beta}$ which is non-zero. If $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, then $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}$ and so $\mathbb{E}[R_{\boldsymbol{e},j}(\boldsymbol{n}_j)] = \boldsymbol{\beta}_j - \boldsymbol{\beta}$ is non-zero. This is a contradiction, and so $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for all i, j.

We have thus proven that set $\{\boldsymbol{e}_i\} \in \mathcal{G}_K$ is of the form $\boldsymbol{e}_i = \boldsymbol{\alpha}\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}$ for all i.