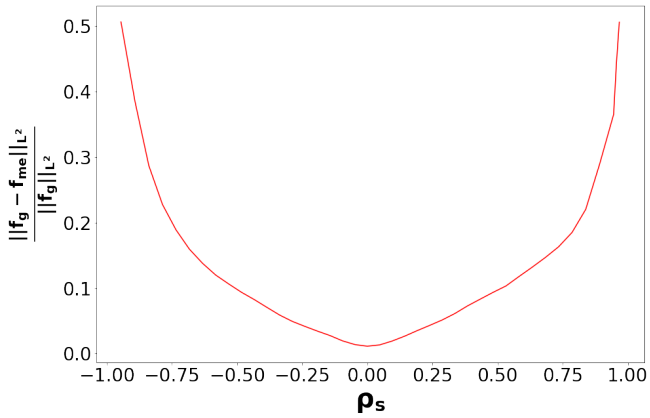

Figure 4: Normalized L_2 distance between two bivariate distributions with standard normal marginals, one with Gaussian copula and the other with maximum-entropy copula under Spearman correlation constraints, for various Spearman correlation values ρ_s.

A Further Details

A.1 Additional Experiment

It can be shown that the highest R^2 and the lowest RMSE achievable by a regression model using $\boldsymbol{x}$ to predict y read respectively

$$\bar{R}^2\left(P_{\boldsymbol{x},y}\right) = 1 - e^{-2I(y;\boldsymbol{x})}$$

and

$$R\bar{M}SE\left(P_{\boldsymbol{x}},y\right) = e^{-I(y;\boldsymbol{x})}\sqrt{\mathbb{V}\mathrm{ar}\left(y\right)}.$$

We use MIND to estimate the highest performance achievable in the Kaggle competition 'House Prices: Advanced Regression Techniques'. The aim is to predict the price at which houses were sold from various continuous and categorical variables. The results are summarized in Table 3, when all variables are used, when the 10 explanatory variables OverallQual, GrLivArea, YearBuilt, TotalBsmtSF, OverallCond, LotArea, BsmtFinSF1, BldgType, KitchenQual, MSZoning are used, and when the first 5 of the foregoing list are used.

A.2 Further Details on Handling Categorical Data

Categorical and non-ordinal variables should be ordinarily encoded as customary, and ordinal data should be treated as continuous variables. The only practical requirements for the validity of this approach are i) to use a ranking function that assigns different ranks to all inputs including ties (e.g. scipy's 'rankdata' function with method 'ordinal', or leveraging PyTorch's or Tensorflow's 'argsort'.), and ii) to avoid encoding methods that may result in linearly dependent coordinates (e.g. one-hot-encoding on a binary non-ordinal categorical variable). When a suitable ranking function is not available a small random jitter may be added to ordinal variables to remove ties.

This approach is mathematically valid thanks to the quantization characterization of the mutual information (Cover (1999), Definition 8.54). In effect, if we denote $\mathcal{P}$ (resp. $\mathcal{Q}$) a partition of the domain $\mathcal{X}$ (resp. $\mathcal{Y}$) of $\boldsymbol{x}$ (resp. $\boldsymbol{y}$), and $[\boldsymbol{x}]_{\mathcal{P}} \in \mathbb{N}$ (resp. $[\boldsymbol{y}]_{\mathcal{Q}} \in \mathbb{N}$) a discrete random variable indicating which element of $\mathcal{P}$ (resp. $\mathcal{Q}$) $\boldsymbol{x}$ (resp. $\boldsymbol{y}$) belongs to, then we have

$$I\left(\boldsymbol{y};\boldsymbol{x}\right) = \sup_{\mathcal{P},\mathcal{Q}} I\left([\boldsymbol{y}]_{\mathcal{Q}};[\boldsymbol{x}]_{\mathcal{P}}\right), \tag{12}$$

where the rightmost mutual information is between discrete random variables.

This characterization implies that the mutual information is always invariant by 1-to-1 transformations, of which ordinal encoding is one, whether coordinates are all continuous, all categorical, or a mix.[5] It can also be seen that adding a negligible random jitter to an ordinal random variable will not materially change the mutual information,[6] but will turn the ordinal variable into a continuous one so that results developed for continuous variables may apply. Strictly speaking, by the data processing inequality (Cover (1999), Theorem 2.8.1), adding a random jitter to ordinal variables increases the mutual information but, as the jitter standard deviation goes to zero, the difference becomes negligible, even though we still enjoy the benefits of working with continuous variables, without downside. Considering that MIND only depends on variables through their ranks, we may do without adding a jitter, so long as the ranking algorithm does not attribute the same rank to ties.

Another approach for handling categorical variables would be to use Table 2, and to use the entropy decomposition formula (Equation (1)) to estimate differential entropies and conditional entropies. However, this approach can be far less data-efficient as it requires splitting the dataset into as many subsets as the number of distinct tuples of categorical variable values, so as to evaluate conditional differential entropies.

[5] Indeed, for any 1-to-1 transformation f and partition $\mathcal{P}$ of $\mathcal{X}$, we may always find a partition $\mathcal{P}'$ of the image space $f\left(\mathcal{X}\right)$ such that $[\boldsymbol{x}]_{\mathcal{P}} = [f\left(\boldsymbol{x}\right)]_{\mathcal{P}'}$.

[6] If we denote g the operation consisting of adding a negligible random jitter to an ordinal variable x_i, then by reducing the jitter's standard deviation, for any partition $\mathcal{P}$ of the domain of x_i, we may always find a partition $\mathcal{P}'$ of the image space such that $\mathbb{P}\left([x_i]_{\mathcal{P}} = [g\left(x_i\right)]_{\mathcal{P}'}\right) = 1 - e$ for e arbitrarily small.

	y **is continuous**	y **is categorical**
$\boldsymbol{x}$ **is continuous**	$I(\boldsymbol{y};\boldsymbol{x}) = h(\boldsymbol{y}) + h(\boldsymbol{x}) - h(\boldsymbol{y};\boldsymbol{x})$	$I(y;\boldsymbol{x}) = h(\boldsymbol{x}) - \sum_{i\in\mathcal{Y}} h(\boldsymbol{x}\vert y=i)P_y(i)$
$\boldsymbol{x}$ **is categorical**	$I(\boldsymbol{y};\boldsymbol{x}) = h(\boldsymbol{y}) - \sum_{i\in\mathcal{X}} h(\boldsymbol{y}\vert\boldsymbol{x}=i)P_{\boldsymbol{x}}(i)$	$I(y;\boldsymbol{x}) = H(y) + H(\boldsymbol{x}) - H(y;\boldsymbol{x})$
$\boldsymbol{x}$ **has continous coordinates** $\boldsymbol{x}_c$ **and categorical coordinates** $\boldsymbol{x}_d$	$I(\boldsymbol{y};\boldsymbol{x}) = h(\boldsymbol{y}) + \sum_{i\in\mathcal{X}_d} \left[h(\boldsymbol{x}_c\vert\boldsymbol{x}_d=i) - h\left(\boldsymbol{y},\boldsymbol{x}_c\vert\boldsymbol{x}_d=i\right)\right] P_{\boldsymbol{x}_d}(i)$	$I(y;\boldsymbol{x}) = I(y;\boldsymbol{x}_d) + \sum_{i\in\mathcal{X}_d} P_{\boldsymbol{x}_d}(i)h(\boldsymbol{x}_c\vert\boldsymbol{x}_d=i) - \sum_{j\in\mathcal{Y}} h\left(\boldsymbol{x}_c\vert\boldsymbol{x}_d=i, y=j\right) P_{\boldsymbol{x}_d,y}(i,j)$

Table 2: Expression of the mutual information $I(y;\boldsymbol{x})$ as a function of the Shannon entropy $H(.)$, and/or the differential entropy $h(.)$, depending on whether y and/or $\boldsymbol{x}$ has continuous and/or categorical coordinates. Expressions of the type $h(\boldsymbol{x}|y=i)$ are to be understood as the differential entropy of the continuous conditional distribution $\boldsymbol{x}|y=i$.

$I(y;\boldsymbol{x})$	$\bar{R}^2$	$R\bar{M}SE$	d	n
1.50	0.95	18,531	80	1460
0.76	0.78	36,979	10	1460
0.65	0.73	41,007	5	1460

Table 3: Mutual information and highest performances achievable in the 'House Prices: Advanced Regression Techniques' Kaggle challenge using all, 10 or 5 explanatory variables.

B Proofs

B.1 Proof of Proposition 2.1

Any continuous 1-to-1 univariate transformation g_i is either increasing or decreasing. Increasing transformations leave copulas invariant. Moreover, any continuous decreasing function g_i on $\mathbb{R}$ can be written as $f(-x)$ where f is a continuous increasing function, so that we may focus on proving the statement for the transformation $g_i : z \to -z$. Let us denote $\bar{\boldsymbol{z}} = (z_1, \dots, -z_i, \dots, z_d)$, $\bar{c}$ its copula density, and c the copula density of $\boldsymbol{z} = (z_1, \dots, z_d)$. We have

$$\bar{c}\left(u_1, \dots, u_d\right) = c\left(u_1, \dots, 1-u_i, \dots, u_d\right).$$

A simple change of variables shows that $h\left(\boldsymbol{u}_z\right) = h\left(\boldsymbol{u}_{\bar{z}}\right)$.

B.2 Proof of Theorem 3.1

Let P be a d-dimensional copula distribution, and U the uniform distribution on $[0,1]^d$. We note that $h(P) = -KL(P||U)$. Thus the optimization problem (MIND) is equivalent to looking for the I-projection of U on the space $\mathcal{E}$ of copula distributions satisfying the linear constraint $E_P\left[\phi_m(\boldsymbol{u})\right] = \boldsymbol{\alpha}_m$, as defined in Csiszár (1975).

Existence and uniqueness: If there exists a copula distribution P satisfying the constraints and admitting an entropy, then $\mathcal{E}$ is not empty. $\mathcal{E}$ is convex as every convex combination of copulas satisfying the linear constraint $E_P\left[\phi_m(\boldsymbol{u})\right] = \boldsymbol{\alpha}_m$ is itself a copula that satisfies said constraint.

We say that a space of continuous distributions supported on $[0,1]^d$ is variation closed when it is closed in the topology of the variation distance $|P-Q| = \int |p-q|dU$, where p and q are the Radon-Nikodym derivatives of P and Q with respect to U (i.e. their pdfs).

Lemma B.1. *$\mathcal{E}$ is variation-closed.*

Proof. Let $P_n \in \mathcal{E}$ be a sequence converging in variation to a distribution P. We need to show that P also satisfies the linear constraints and has uniform marginals. Convergence in variation implies that for every test function f

$$\int_{[0,1]^d} f(\boldsymbol{u})p_n(\boldsymbol{u})d\boldsymbol{u} \to \int_{[0,1]^d} f(\boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u}. \tag{13}$$

Taking $f = \phi_m$ proves that the limit distribution P satisfies the linear constraints. We now need to prove that it has uniform marginals.

Let us consider $\boldsymbol{u}_{-i} = (\dots, u_{i-1}, u_{i+1}, \dots)$ the vector $\boldsymbol{u}$ without its i-th coordinate, and let us choose a test function f that only depends on u_i. By Fubini's theorem we have

$$\begin{aligned}\int_{[0,1]^d} f(\boldsymbol{u})p_n(\boldsymbol{u})d\boldsymbol{u} &= \int f(u_i)\underbrace{\left(\int p_n(\boldsymbol{u})d\boldsymbol{u}_{-i}\right)}_{=1}du_i \\ &= \int_{[0,1]} f(u_i)du_i \\ &= \int_{[0,1]^d} f(u_i)p(\boldsymbol{u})d\boldsymbol{u}\end{aligned}$$

where the last equality is due to the Equation (13). Putting the last two equality together, we get

$$\int f(u_i)\left(1 - \int p(\boldsymbol{u})d\boldsymbol{u}_{-i}\right)du_i = 0$$

for every univariate test function f, which implies that every marginal of P is uniform. □

$\mathcal{E}$ being convex, non-empty, and variation closed, U admits a unique I-projection on $\mathcal{E}$ (see Theorem 2.1

in Csiszár (1975)), or equivalently, (MIND) admits a unique solution.

Functional form of the pdf: Let us denote $\mathcal{F}$ the space of distributions supported on $[0,1]^d$ that satisfy the linear constraint $E_P[\phi_m(\boldsymbol{u})] = \boldsymbol{\alpha}_m$. Clearly, $\mathcal{E} \subset \mathcal{F}$ as the only difference between the two sets is that $\mathcal{E}$ only contains elements of $\mathcal{F}$ with uniform marginals (i.e. copula distributions). The I-projection P_{AM} of U on $\mathcal{F}$, which exists because $\mathcal{F}$ is convex, non-empty, and variation cloosed, is the minimizer of the problem (A-MIND).

By Theorem 2.3 in Csiszár (1975), P_{M} is the I-projection of P_{AM} on $\mathcal{E}$.

Applying Theorem 3.1 (Case A) in Csiszár (1975), we get that P_{AM} has density with respect to U, which is also its pdf, of the form $p_{\mathrm{AM}} = e^{\boldsymbol{\theta}^T \phi_m(\boldsymbol{u})}$. Moreover, any distribution in $\mathcal{F}$ with density with respect to U of this form is the I-projection of U on $\mathcal{F}$.

Applying Theorem 3.1 (Case B) in Csiszár (1975), we get that P_{M} has density with respect to P_{AM} of the form $\prod_{i=1}^d f_i(u_i)$, where f_i are non-negative and log-integrable. Moreover, any distribution in $\mathcal{E}$ with density with respect to P_{AM} of this form is the I-projection of P_{AM} on $\mathcal{E}$.

Hence, P_{M} has density with respect to U, which is also its pdf,

$$p_{\mathrm{M}} = e^{\boldsymbol{\theta}^T \phi_m(\boldsymbol{u})} \prod_{i=1}^d f_i(u_i),$$

and any distribution on $[0,1]^d$ whose pdf of this form is the minimizer of (MIND).

Pythagoras' Identity: Theorem 3.1 in Csiszár (1975) guarantees that identity (3.1) in Csiszár (1975) holds and

$$-h(P) = -h(P_{\mathrm{AM}}) + KL(P||P_{\mathrm{AM}}) \tag{14}$$

for any $P \in \mathcal{F}$ and

$$KL(Q||P_{\mathrm{AM}}) = KL(P_{\mathrm{M}}||P_{\mathrm{AM}}) + KL(Q||P_{\mathrm{M}}) \tag{15}$$

for any $Q \in \mathcal{E} \subset \mathcal{F}$.

As $Q \in \mathcal{F}$,

$$KL(Q||P_{\mathrm{AM}}) = h(P_{\mathrm{AM}}) - h(Q).$$

Thus,

$$h(P_{\mathrm{AM}}) - h(Q) = KL(P_{\mathrm{M}}||P_{\mathrm{AM}}) + KL(Q||P_{\mathrm{M}})$$

and

$$-h(Q) = -h(P_{\mathrm{AM}}) + KL(P_{\mathrm{M}}||P_{\mathrm{AM}}) + KL(Q||P_{\mathrm{M}}).$$

As $P_{\mathrm{M}} \in \mathcal{F}$,

$$-h(P_{\mathrm{M}}) = -h(P_{\mathrm{AM}}) + KL(P_{\mathrm{M}}||P_{\mathrm{AM}}),$$

and we get

$$-h(Q) = -h(P_{\mathrm{M}}) + KL(Q||P_{\mathrm{M}}).$$

B.3 Proof of Theorem 3.2

To prove Theorem 3.1 we had to prove Theorem 3.2. See Section B.2.

B.4 Proof of Theorem 3.3

Let

$$g(\boldsymbol{u}; \phi_m, \boldsymbol{\alpha}_m) := \prod_{i=1}^q p_{\mathrm{M}}(w, \boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i).$$

We want to prove that

$$g(\boldsymbol{u}; \phi_m, \boldsymbol{\alpha}_m) = p_{\mathrm{M}}(\boldsymbol{u}; \phi_m, \boldsymbol{\alpha}_m).$$

We know from Theorem 3.1 that $p_{\mathrm{M}}(\boldsymbol{u}; \phi_m, \boldsymbol{\alpha}_m)$ is the only copula density of the form

$$e^{\boldsymbol{\theta}^T \phi_m(\boldsymbol{u})} = e^{\sum_{i=1}^q \boldsymbol{\theta}_i^T \eta_i(w, \boldsymbol{v}_i)} = \prod_{i=1}^q e^{\boldsymbol{\theta}_i^T \eta_i(w, \boldsymbol{v}_i)}$$

that satisfies the constraints $E_P[\psi_m(\boldsymbol{u})] = \boldsymbol{\beta}_m$.

First we note that g is a copula entropy. Indeed, integrating g with respect to every variable but a coordinate of $\boldsymbol{v}_i$ is always 1 by virtue of the fact that $p_{\mathrm{M}}(w, \boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i)$ are copula densities. To see why, note that we may first integrate with respect to $\boldsymbol{v}_j$ for all $j \neq i$, and then with respect to ω and all other coordinates of $\boldsymbol{v}_i$. Additionally, if we integrate with respect to all variables but ω, we get

$$\begin{aligned}
&\int g(\boldsymbol{u}; \phi_m, \boldsymbol{\alpha}_m)\, d\boldsymbol{v}_1 \dots d\boldsymbol{v}_q \\
&= \int p_{\mathrm{M}}(w, \boldsymbol{v}_q; \eta_q, \boldsymbol{\beta}_q) \times \cdots \times \\
&\left(\underbrace{\int p_{\mathrm{M}}(w, \boldsymbol{v}_1; \eta_1, \boldsymbol{\beta}_1)\, d\boldsymbol{v}_1}_{=1} \right) d\boldsymbol{v}_2 \dots d\boldsymbol{v}_q \\
&= 1.
\end{aligned}$$

Second, g clearly has the form $\prod_{i=1}^q e^{\boldsymbol{\theta}_i^T \eta_i(w, \boldsymbol{v}_i)}$.

Finally, g satisfies the constraints $E_P[\psi_m(\boldsymbol{u})] = \boldsymbol{\beta}_m$

as

$$\begin{aligned}
&\int \eta_i\left(w,\boldsymbol{v}_i\right) g\left(\boldsymbol{u};\phi_m,\boldsymbol{\alpha}_m\right) d\boldsymbol{u}\\
=&\int \eta_i\left(w,\boldsymbol{v}_i\right) p_{\mathrm{M}}\left(w,\boldsymbol{v}_i;\eta_i,\boldsymbol{\beta}_i\right)\\
&\times\left(\underbrace{\int\prod_{j\neq i} p_{\mathrm{M}}\left(w,\boldsymbol{v}_j;\eta_j,\boldsymbol{\beta}_j\right) d\boldsymbol{v}_j}_{=1}\right) d\omega d\boldsymbol{v}_i\\
=&\int \eta_i\left(w,\boldsymbol{v}_i\right) p_{\mathrm{M}}\left(w,\boldsymbol{v}_i;\eta_i,\boldsymbol{\beta}_i\right) d\omega d\boldsymbol{v}_i\\
=&\boldsymbol{\beta}_i,
\end{aligned}$$

where we've used the fact that each $p_{\mathrm{M}}\left(w,\boldsymbol{v}_j;\eta_j,\boldsymbol{\beta}_j\right)$ has uniform marginals, and satisfies the constraint

$$E_P\left[\eta_j\left(\omega,\boldsymbol{v}_j\right)\right]=\boldsymbol{\beta}_j.$$

B.5 Proof of Theorem 3.4

The essence of the proof is in the transitivity property of I-projections. Indeed, if $\mathcal{P}\subset\mathcal{Q}$ are linear sets of probability distributions supported on $[0,1]^d$, Q the I-projection of the standard uniform U on $\mathcal{Q}$, and P the I-projection of U on $\mathcal{P}$, then P is also the I-projection of Q on $\mathcal{P}$ (Theorem 2.3 Csiszár (1975)).

In the case of Theorem 3.4, $\mathcal{Q}$ is the set of probability distributions satisfying all constraints, and $\mathcal{Q}$ is the set of probability distributions satisfying all but the between-blocks constraints. If we denote, $q\left(\boldsymbol{u};\psi_m,\boldsymbol{\beta}_m\right):=\prod_{i=1}^q p_{\mathrm{AM}}\left(\boldsymbol{v}_i;\eta_i,\boldsymbol{\beta}_i\right)$, a direct application of Theorem 4.1 shows that q is the density of the I-projection Q of U on $\mathcal{Q}$. The maximizer of the full (A-MIND) problem is therefore the I-projection of Q on $\mathcal{P}$, and we know from Theorem 3.1 in Csiszár (1975) that it has Radon-Nikodym derivative with respect to Q of the form $\frac{dP}{dQ}=e^{\boldsymbol{\theta}^T\phi_m^k(\boldsymbol{u})}$. Putting everything together, we get that the maximizer P of the full (A-MIND) problem has pdf of the form

$$p_{\mathrm{AM}}\left(\boldsymbol{u};\phi_m,\boldsymbol{\beta}_m\right)=e^{\boldsymbol{\theta}^T\phi_m^k(\boldsymbol{u})}\prod_{i=1}^q p_{\mathrm{AM}}\left(\boldsymbol{v}_i;\eta_i,\boldsymbol{\beta}_i\right).$$

The unicity of this representation is a direct consequence of Theorem 4.1. Taking the negative log of this expression and then the expectation, we get

$$\begin{aligned}
&h_{\mathrm{AM}}\left(\boldsymbol{u};\phi_m^k,\boldsymbol{\beta}_m\right)=-\boldsymbol{\theta}^T\boldsymbol{\alpha}_m^k\\
&-\sum_{i=1}^q E_{p_{\mathrm{AM}}\left(\boldsymbol{v}_i;\phi_m^k,\boldsymbol{\beta}_m\right)}\left[\log p_{\mathrm{AM}}\left(\boldsymbol{v}_i;\eta_i,\boldsymbol{\beta}_i\right)\right].
\end{aligned}$$

The final result stems from the identity

$$E_p\left(-\log q\right)=E_p\left(-\log p\right)-\mathrm{KL}\left[p||q\right].$$

The fact that

$$-\boldsymbol{\theta}^T\boldsymbol{\alpha}_m^k\leq\sum_{i=1}^q\mathrm{KL}\left[p_{\mathrm{AM}}\left(\boldsymbol{v}_i;\phi_m^k,\boldsymbol{\beta}_m\right)||p_{\mathrm{AM}}\left(\boldsymbol{v}_i;\eta_i,\boldsymbol{\beta}_i\right)\right]$$

is a direct consequence of $\mathcal{P}\subset\mathcal{Q}$, which implies that the entropy of P cannot be greater than that of Q, and the fact that $\sum_{i=1}^q h_{\mathrm{AM}}\left(\boldsymbol{v}_i;\eta_i,\boldsymbol{\beta}_i\right)$ is the entropy of Q. The two entropies are the same if and only if $P=Q$ or, equivalently, $\boldsymbol{\theta}=0$. When $\forall\boldsymbol{u},\ \gamma\left(\boldsymbol{u}\right)$ is constant and equal to 1, $P=Q$.

B.6 Proof of Lemma 4.1

The Hessian of the objective, namely $\int_{[0,1]^d}\left(\phi_m\left(\boldsymbol{u}\right)\phi_m\left(\boldsymbol{u}\right)^T\right)e^{\boldsymbol{\theta}^T\phi_m(\boldsymbol{u})}d\boldsymbol{u}$, is clearly strictly positive-definite as coordinates of 1 are not linearly related.

B.7 Proof of Lemma 4.2

The only critical point of the objective of (CVX-MIND) satisfies

$$\boldsymbol{\alpha}_m=\int_{[0,1]^d}\phi_m\left(\boldsymbol{u}\right)e^{\boldsymbol{\theta}^{*T}\phi_m(\boldsymbol{u})}d\boldsymbol{u}.$$

Given that the first coordinates of ϕ_m and $\boldsymbol{\alpha}_m$ are both 1, $\int_{[0,1]^d}e^{\boldsymbol{\theta}^{*T}\phi_m(\boldsymbol{u})}d\boldsymbol{u}=1$. By Theorem 3.2, the distribution with pdf $e^{\boldsymbol{\theta}^{*T}\phi_m(\boldsymbol{u})}$ maximizes (A-MIND).

B.8 Proof of Theorem 4.1

Theorem 4.1-A is a consequence of the uniqueness of the solution to the Hausdorff moment problem.

Indeed, the uniform distribution on $[0,1]$ is uniquely characterized by the sequence of moments $\forall j, E(u^j)=1/(1+j)$ (Shohat and Tamarkin (1943)). Thus, we may replace the uniform marginal constraints in (MIND) with the constraints $\forall i,j,\ E(u_i^j)=1/(1+j)$. The only difference between (A-MIND) and (MIND) is that the former has k moment constraints whereas the latter has all moment constraints. It follows that

$$\forall m>0,\quad h_{\mathrm{AM}}\left(\boldsymbol{u};\phi_m^k,\boldsymbol{\beta}_m\right)\underset{k\to\infty}{\longrightarrow}h_{\mathrm{M}}\left(\boldsymbol{u};\psi_m,\boldsymbol{\beta}_m\right).$$

Additionally, a direct application of the basic consistency theorem for extremum estimators (see Newey and McFadden (1994)) to (CVX-MIND) shows that for any consistent estimator $\hat{\boldsymbol{\beta}}_{m,n}$ of $\boldsymbol{\beta}_m$,

$$h_{\mathrm{AM}}\left(\boldsymbol{u};\phi_m^k,\hat{\boldsymbol{\beta}}_{m,n}\right)\underset{n\to\infty}{\longrightarrow}h_{\mathrm{AM}}\left(\boldsymbol{u};\phi_m^k,\boldsymbol{\beta}_m\right).$$

As for Theorem 4.1-B, we simply need to prove that:

$$\forall k>0,\quad h_{\mathrm{AM}}\left(\boldsymbol{u};\phi_m^k,\boldsymbol{\beta}_m\right)\underset{m\to\infty}{\longrightarrow}h\left(\boldsymbol{u}_z\right).$$

We recall that $h(\boldsymbol{u}_z) = -KL(P_{\boldsymbol{u}_z}||U)$. Using the NWJ characterization of the KL divergence (Nguyen et al. (2010)), we get:

$$h(\boldsymbol{u}_z) = \inf_{T\in\mathcal{T}} \; - E_{P_{\boldsymbol{u}_z}}[T(\boldsymbol{u})] + \int_{[0,1]^d} e^{T(\boldsymbol{u})} - 1 d\boldsymbol{u}, \tag{16}$$

where $\mathcal{T}$ is the space of continuous functions on $[0,1]^d$. Using the fact that $(\phi_m)_m$ is dense in $\mathcal{T}$ (property (P3)), we may rewrite Equation (16) as

$$h(\boldsymbol{u}_z) = \lim_{m\to\infty} \min_{\boldsymbol{\theta}_m} \; - 1 - \boldsymbol{\theta}_m^T \underbrace{E_{P_{\boldsymbol{u}_z}}[\phi_m(\boldsymbol{u})]}_{:=\boldsymbol{\alpha}_m} + \int_{[0,1]^d} e^{\boldsymbol{\theta}_m^T \phi_m(\boldsymbol{u})} d\boldsymbol{u}, \tag{17}$$

where $\boldsymbol{\theta}_m$ satisfies $\boldsymbol{\theta}_m^T \phi_m(\boldsymbol{u}) = T(\boldsymbol{u}) - 1$. Note that the inner optimization problem has the same minimizer as (CVX-MIND), and the minimum is $h_{\mathrm{AM}}\left(\boldsymbol{u}; \phi_m^k, \boldsymbol{\beta}_m\right)$, where we have used the fact that the first coordinate of ϕ_m is 1.

Putting everything together, we get

$$\forall k,\; h(\boldsymbol{u}_z) = \lim_{m\to\infty} h_{\mathrm{AM}}\left(\boldsymbol{u}; \phi_m^k, \boldsymbol{\beta}_m\right).$$