A Theoretical Details

A.1 Hydraulic Principles in WDSs

A WDS can be modeled as a graph, consisting of nodes $V = \{v_1, \dots, v_{n_\mathrm{n}}\}$ (such as consumer junctions, reservoirs and tanks) and edges $E = \{e_1, \dots, e_{n_\mathrm{e}}\} = \{e_{vu} \mid \forall v \in V, u \in \mathcal{N}(v)\}$ (such as pipes, pumps and valves). Among all nodes, we consider the set of reservoir nodes V_r, the consumer nodes $V_\mathrm{c} = V \setminus V_\mathrm{r}$ and the sensor nodes $V_\mathrm{s} \subset V_\mathrm{c}$, which are equipped with a sensor that measures the pressure head $h_v \in \mathbb{R}_+$ at the corresponding node $v \in V_\mathrm{s}$.

More precisely, the state of a WDS is characterized by pressure heads $\mathbf{h} = (h_v)_{v \in V} \in \mathbb{R}_+^{n_\mathrm{n}}$ at every node and water demands $\mathbf{d} = (d_v)_{v \in V_\mathrm{c}} \in \mathbb{R}_+^{n_\mathrm{c}}$ at every consumer node along with the water flows $\mathbf{q} = (q_e)_{e \in E} \in \mathbb{R}^{n_\mathrm{e}}$ through every pipe. The relationships between these variables are governed by the following hydraulic principles [17].

Flow direction For two neighboring nodes $v, u \in V$, the flow $q_{vu} := q_{e_{vu}}$ from node v to u is equal to the negative of the flow $q_{uv} := q_{e_{uv}}$ from u to v:

$$q_{vu} = -q_{uv}. \tag{11}$$

We use the convention that an *outflow* q_{vu} from $v \in V$ (to $u \in \mathcal{N}(v)$) has a positive sign $\operatorname{sgn}(q_{vu}) = 1$ and that an *inflow* q_{vu} to $v \in V$ (from $u \in \mathcal{N}(v)$) has a negative sign $\operatorname{sgn}(q_{vu}) = -1$. By that, we can separate the neighborhood $\mathcal{N}(v)$ of v into the neighborhood of out- and inflows, respectively: $\mathcal{N}_\pm(v) := \{u \in \mathcal{N}(v) \mid \operatorname{sgn}(q_{vu}) = \pm 1\}$.

Conservation of mass The sum of inflows to a node $v \in V$ equals – up to the sign – the sum of the outflows from that node and its demand $d_v \in \mathbb{R}_+$, or, equivalently, the sum of flows at a node v is equal to the negative of the water demand d_v at that node:

$$-\sum_{u \in \mathcal{N}_-(v)} q_{vu} = d_v + \sum_{u \in \mathcal{N}_+(v)} q_{vu} \quad \Longleftrightarrow \quad \sum_{u \in \mathcal{N}(v)} q_{vu} = -d_v. \tag{12}$$

Conservation of energy The change in between pressure heads h_v and h_u between two neighboring nodes $v \in V$ and $u \in \mathcal{N}(v)$ is linked to the flow q_{vu} between them by

$$h_v - h_u = r_{vu} \operatorname{sgn}(q_{vu}) |q_{vu}|^x, \tag{13}$$

where $x = 1.852$ and $r_{vu} = 10.667\ l_{vu}\ \delta_{vu}^{-4.871}\ c_{vu}^{-1.852} > 0$ is a pipe-dependent parameter with length l_{vu}, diameter δ_{vu} and roughness coefficient c_{vu}.

Definition 4 (Physically correct states). *A triple* $(\mathbf{h}, \mathbf{d}, \mathbf{q}) \in \mathbb{R}^{n_n \times n_c \times n_r}$ *that satisfies the properties* (11) *to* (13) *is called physically correct.*

A.2 PI-GCN Architecture

The PI-GCN model from [2] solves the task of hydraulic state estimation in WDS by estimating the pressure head $\mathbf{h} = (h_v)_{v\in V}$ at every node and water flow $\mathbf{q} = (q_e)_{e\in E}$ through every link given *true* demands $\mathbf{d}^* = (d_v^*)_{v\in V\setminus V_r}$ at every consumer node, *true* pressure heads $\mathbf{h}_{V_r}^* = (h_v^*)_{v\in V_r}$ at the reservoirs, and pipe attributes $(r_e)_{e\in E}$ given by:

$$r_{e_{vu}} = 10.667\ l_{e_{vu}} \delta_{e_{vu}}^{-4.871} c_{e_{vu}}^{-1.852} > 0, \tag{14}$$

where $x = 1.852$ and $l_{e_{vu}}$, $\delta_{e_{vu}}$ and $c_{e_{vu}}$ are length, diameter and roughness coefficient of the pipe [17].

The model combines a local learnable GCN f_1 with a global physics-informed algorithm f_2 and solves the task at hand iteratively as depicted in Figure 5. Their methodology employs hydraulic principles (cf. subsection A.1) thus ensuring that the estimated hydraulic state is physically correct. This enables us to use physics-informed gradients of a trained PI-GCN to solve real-world problems in the WDS domain through backpropagation or sensitivity analysis.

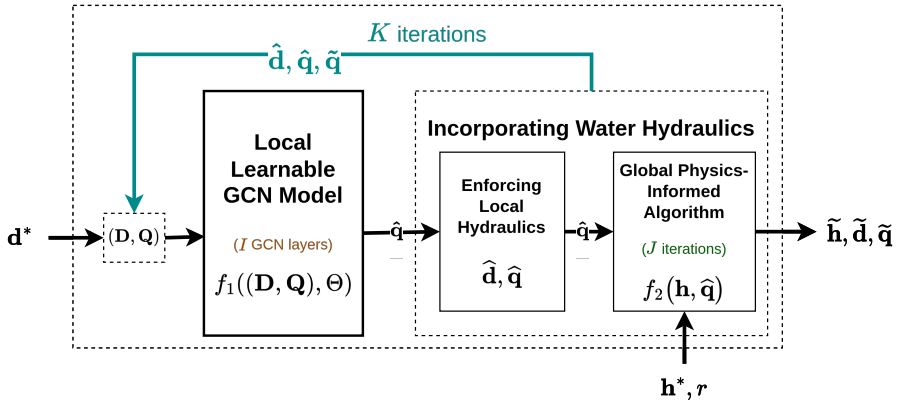

Fig. 5: PI-GCN model architecture: The local GCN f_1 learns from the global physics-informed algorithm f_2 through multiple iterations.

A.3 Approximation of the Mutual Information and Entropy of Continuously Distributed Random Variables

Since for continuously distributed random variables, their entropy and mutual information is defined by their density,[3] the first part of this subsection explains

[3] This automatically implies that entropy and mutual information are not defined for continuously distributed random variables which do not have a density function.

how to approximate a usually unknown density function by an elementary function leveraging the concept of histograms.

Definition 5 (Density approximation). *Let*

- $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space,*
- $\mathrm{X} : \Omega \to \mathbb{R}$ *a random variable with bounded image* $\mathrm{X}(\Omega) = [x_{\min}, x_{\max}]$ *and unknown, continuous density function* $\varphi : \mathbb{R} \longrightarrow [0, \infty)$ *w.r.t. the Lebesgue-measure* λ*.*

We approximate the unknown density φ *as follows:*

- *For* $n \in \mathbb{N}$*, let* $x_1, ..., x_n \in \mathbb{R}$ *be drawn from independent and identically distributed (iid) random variables* $\mathrm{X}_1, ..., \mathrm{X}_n$ *with distribution* $\mathbb{P}\mathrm{X}_i^{-1} = \mathbb{P}\mathrm{X}^{-1}$ *for all* $i = 1, \ldots, n$*.*
- *For* $n_b \in \mathbb{N}$*, divide the image* $\mathrm{X}(\Omega) = [x_{\min}, x_{\max}]$ *into* n_b *bins. To do so, choose a fixed bin width* $\triangle_{n_b} := \frac{x_{\max} - x_{\min}}{n_b}$ *and bin boundaries* $a_j := x_{\min} + j \cdot \triangle_{n_b} \in [x_{\min}, x_{\max}]$ *for all* $j = 0, ..., n_b$ *in order to define the bins as* $A_j := [a_j, a_{j+1})$ *for* $j = 0, ..., n_b - 2$ *and* $A_j = A_{n_b - 1} = [a_{n_b - 1}, a_{n_b}]$ *for* $j = n_b - 1$*.*
- *Based on both* $n \in \mathbb{N}$ *and* $n_b \in \mathbb{N}$*, compute normalized bin heights* $z_j := \frac{1}{n \cdot \triangle_{n_b}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(x_i)$ *for all* $j = 0, ..., n_b - 1$*.*
- *Based on all magnitudes above, approximate the unknown density function* $\varphi : \mathbb{R} \longrightarrow [0, \infty)$ *by the non-negative function* $\varphi_{n,n_b} : \mathbb{R} \longrightarrow [0, \infty)$*, given by*

$$\hat{\varphi}(x) := \varphi_{n,n_b}(x) := \sum_{j=0}^{n_b - 1} z_j \cdot \mathbb{1}_{A_j}(x) \quad \forall x \in \mathbb{R}.$$

Note that A_j *and* z_j *depend on* n_b *and on* n, n_b*, respectively, which we omit for readability.*

Remark 2. Definition 5 and the following results easily extend to random vectors $(\mathrm{X}, \mathrm{Y}) : \Omega \to \mathbb{R} \times \mathbb{R}$ with bounded image $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$ and continuous density function $\varphi : \mathbb{R}^2 \to [0, \infty)$ w.r.t. the Lebesgue-measure λ^2 by considering bins of the kind $A_{j_1} \times A_{j_2}$, where A_{j_1} and A_{j_2} are defined as in definition 5 for $j_1, j_2 = 0, ..., n_\mathrm{b} - 1$. We leave it as an exercise to the reader to verify the details along the following results and their proofs.

Lemma 1. *In the setting of definition 5, for all* $n, n_b \in \mathbb{N}$*,* φ_{n,n_b} *is a density function with respect to the Lebesgue measure* λ*.*

Proof. Let $n, n_\mathrm{b} \in \mathbb{N}$.

- **Lebesgue-integrability:** φ_{n,n_b} is Borel-measureable since the elementary functions $\mathbb{1}_{A_j}(\cdot)$ are. This is, in turn, because the bins A_j are Borel-measureable with $\lambda(A_j) = a_{j+1} - a_j = \triangle_{n_\mathrm{b}}$ for all $j = 0, ..., n_\mathrm{b} - 1$.
  Moreover, since φ_{n,n_b} is a non-negative function, showing that φ_{n,n_b} is Lebesgue-integrable ($\mathcal{L}^1(\lambda)$), i.e., that $\int_{\mathbb{R}} |\varphi_{n,n_\mathrm{b}}(x)| \, d\lambda(x) < \infty$ holds, follows from showing that $\int_{\mathbb{R}} \varphi_{n,n_\mathrm{b}}(x) \, d\lambda(x) = 1$ holds, which we do next.

– **Probability measure:** By (1) definition of φ_{n,n_b}, (2) linearity of the Lebesgue-integral, (3) basic properties from measure and probability theory, (4) definition of z_j and the fact that $\lambda(A_j) = a_{j+1} - a_j = \triangle_{n_\mathrm{b}}$ holds for all $j = 0, ..., n_\mathrm{b} - 1$, (5) basic transformations and (6) the fact that due to the disjointness of the bins A_j for all $j = 0, ..., n_\mathrm{b} - 1$, each sample x_i lies in exactly one bin A_j for all $i = 1, ..., n$, we obtain

$$\begin{aligned}
\int_{\mathbb{R}} \varphi_{n,n_\mathrm{b}}(x)\, d\lambda(x) &\overset{(1)}{=} \int_{\mathbb{R}} \sum_{j=0}^{n_\mathrm{b}-1} z_j \cdot \mathbb{1}_{A_j}(x)\, d\lambda(x) \\
&\overset{(2)}{=} \sum_{j=0}^{n_\mathrm{b}-1} z_j \int_{\mathbb{R}} \mathbb{1}_{A_j}(x)\, d\lambda(x) \\
&\overset{(3)}{=} \sum_{j=0}^{n_\mathrm{b}-1} z_j \cdot \lambda(A_j) \\
&\overset{(4)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \left(\frac{1}{n \cdot \triangle_{n_\mathrm{b}}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(x_i) \right) \triangle_{n_\mathrm{b}} \\
&\overset{(5)}{=} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{n_\mathrm{b}-1} \mathbb{1}_{A_j}(x_i) \\
&\overset{(6)}{=} \frac{1}{n} \sum_{i=1}^{n} 1 \\
&\overset{(5)}{=} 1.
\end{aligned}$$

□

Note that the densities φ_{n,n_b} have a probabilistic character introduced by the sampling of $x_1, \ldots, x_n \in \mathbb{R}$: In fact, for each $n, n_\mathrm{b} \in \mathbb{N}$ and $x \in X$, it can be considered as a random variable $\varphi_{n,n_\mathrm{b}}(x) : \Omega \to \mathbb{R}$, given by a composition of the random variables $\mathrm{X}_1, \ldots, \mathrm{X}_n$:

$$\varphi_{n,n_\mathrm{b}}(x) = \sum_{j=0}^{n_\mathrm{b}-1} \mathrm{Z}_j \cdot \mathbb{1}_{A_j}(x) = \sum_{j=0}^{n_\mathrm{b}} \left(\frac{1}{n \cdot \triangle_{n_\mathrm{b}}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(\mathrm{X}_i) \right) \mathbb{1}_{A_j}(x).$$

Therefore, in summary, $(\varphi_{n,n_\mathrm{b}})_{n,n_\mathrm{b} \in \mathbb{N}}$ is a sequence of functions $\varphi_{n,n_\mathrm{b}} : \Omega \times \mathbb{R} \to \mathbb{R}$ *with probabilistic character*. In this realm, we want to show $\mathbb{P}$-almost sure uniform convergence of the probabilistic sequence of functions towards the true, but unknown density φ. As a first step, we investigate on the expectation and the variance of the approximated densities, considered as a random variable.

Lemma 2 (Density approximation as a random variable). *In the setting of definition 5, for each $n_b \in \mathbb{N}$ and each $j \in \{0, \ldots, n_b - 1\}$, there exists a $\xi_j \in \overline{A}_j$, such that for all $n \in \mathbb{N}$, $i \in \{1, ..., n\}$ and $x \in [x_{\min}, x_{\max}]$, we obtain*

$$\mathbb{E}\left(\mathbb{1}_{A_j}(\mathrm{X}_i)\right) = \triangle_{n_b} \cdot \varphi(\xi_j), \tag{15}$$

$$\mathbb{E}\left(\mathrm{Z}_j\right) = \varphi(\xi_j), \tag{16}$$

$$\mathrm{Var}\left(\mathrm{Z}_j\right) = \frac{\varphi(\xi_j)\left(1 - \triangle_{n_b} \cdot \varphi(\xi_j)\right)}{n \cdot \triangle_{n_b}}, \tag{17}$$

$$\mathbb{E}\left(\varphi_{n,n_b}(x)\right) = \sum_{j=0}^{n_b-1} \varphi(\xi_j) \cdot \mathbb{1}_{A_j}(x), \tag{18}$$

$$\mathrm{Var}\left(\varphi_{n,n_b}(x)\right) = \sum_{j=0}^{n_b-1} \frac{\varphi(\xi_j)\left(1 - \triangle_{n_b} \cdot \varphi(\xi_j)\right)}{n \cdot \triangle_{n_b}} \cdot \mathbb{1}_{A_j}(x). \tag{19}$$

Proof.

- $\mathbb{P}$**-Integrability:** By definition of a random variable, X_i is $\mathcal{F}$-$\mathcal{B}$-measureable for all $i = 1, .., n$, and as discussed in the proof of lemma 1, $\mathbb{1}_{A_j}(\cdot)$ is $\mathcal{B}$-$\mathcal{B}$-measureable. Therefore, the composition $\mathbb{1}_{A_j}(\mathrm{X}_i)$ is $\mathcal{F}$-$\mathcal{B}$-measureable for all $i = 1, .., n$.
  Additionally, any compositions of the $\mathcal{F}$-$\mathcal{B}$-measureable functions $\mathbb{1}_{A_j}(\mathrm{X}_i)$ and any other $\mathcal{B}$-$\mathcal{B}$-measureable function (e.g., continuous functions from $\mathbb{R} \to \mathbb{R}$, and especially, any linear combinations) will again be $\mathcal{F}$-$\mathcal{B}$-measureable. Therefore, we can test whether expectations (and variances) of such functions with respect to the measure $\mathbb{P}$ exist, i.e., wether the appearing functions are $\mathbb{P}$-integrable ($\mathcal{L}^1(\mathbb{P})$). Since all appearing functions are positive, we can test this along the same lines, and if the appearing expressions are finite, the appearing functions are automatically integrable.
- $\mathbb{P}$**-Independence:** Since the random variables $\mathrm{X}_1, ..., \mathrm{X}_n$ are iid w.r.t. $\mathbb{P}$, so are $\mathbb{1}_{A_j}(\mathrm{X}_1), \ldots, \mathbb{1}_{A_j}(\mathrm{X}_n)$ due to the measurability of the elementary function $\mathbb{1}_{A_j}(\cdot)$.

Now we can prove the equations (15) to (19) step by step:

1. By (1) basic properties from measure and probability theory and (2) the usage of the density φ of X (and X_i), for each $n_\mathrm{b}, n \in \mathbb{N}$, $j \in \{0, \ldots, n_\mathrm{b} - 1\}$ and $i \in \{1, ..., n\}$, we obtain

$$\begin{aligned}
\mathbb{E}\left(\mathbb{1}_{A_j}(\mathrm{X}_i)\right) &\overset{(1)}{=} \mathbb{P}(\mathrm{X}_i \in A_j) \\
&\overset{(1)}{=} \int_{A_j} d\mathbb{P}\mathrm{X}_i^{-1}(x) \\
&\overset{(2)}{=} \int_{a_j}^{a_{j+1}} \varphi(x)\ d\lambda(x).
\end{aligned}$$

   Since φ is continuous, the Lebesgue-integral equals the Riemann-integral, and we can apply the mean value theorem for integration, according to which there exists a $\xi_j \in \overline{A}_j = [a_j, a_{j+1}]$, such that

$$\int_{a_j}^{a_{j+1}} \varphi(x)\ d\lambda(x) = \int_{a_j}^{a_{j+1}} \varphi(x)\ dx = (a_{j+1} - a_j) \cdot \varphi(\xi_j) = \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)$$

holds and therefore, in summary, we obtain

$$\mathbb{E}\left(\mathbb{1}_{A_j}(\mathrm{X}_i)\right) = \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j).$$

2. Consequently, by (1) definition of Z_j and (2) linearity of expectation, for each $n_\mathrm{b}, n \in N$ and $j \in \{0, \ldots, n_\mathrm{b} - 1\}$, we obtain

$$\mathbb{E}\left(n \triangle_{n_\mathrm{b}} \mathrm{Z}_j\right) \overset{(1)}{=} \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}_{A_j}(\mathrm{X}_i)\right) \overset{(2)}{=} \sum_{i=1}^{n} \mathbb{E}\left(\mathbb{1}_{A_j}(\mathrm{X}_i)\right) = n \cdot \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)$$

and

$$\mathbb{E}\left(\mathrm{Z}_j\right) \overset{(1)}{=} \mathbb{E}\left(\frac{1}{n \cdot \triangle_{n_\mathrm{b}}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(\mathrm{X}_i)\right) \overset{(2)}{=} \varphi(\xi_j).$$

3. Consequently, by (1) definition of Z_j, (2) basic transformations, (3) linearity of expectation, (4) independence of X_{i_1} and X_{i_2} for all $i_1, i_2 \in \{1, \ldots, n\}$ with $i_1 \neq i_2$ and (5) the above results, for each $n_\mathrm{b}, n \in \mathbb{N}$ and $j \in \{0, \ldots, n_\mathrm{b} - 1\}$, we obtain

$$\begin{aligned}
\mathbb{E}\left(\left(n \triangle_{n_\mathrm{b}} \mathrm{Z}_j\right)^2\right) &\overset{(1)}{=} \mathbb{E}\left(\left(\sum_{i=1}^{n} \mathbb{1}_{A_j}(\mathrm{X}_i)\right)^2\right) \\
&\overset{(2)}{=} \mathbb{E}\left(\left(\sum_{i=1}^{n} \mathbb{1}^2_{A_j}(\mathrm{X}_i)\right) + \sum_{i_1=1}^{n} \sum_{\substack{i_2=1 \\ i_1 \neq i_1}}^{n} \mathbb{1}_{A_j}(\mathrm{X}_{i_1}) \cdot \mathbb{1}_{A_j}(\mathrm{X}_{i_2})\right) \\
&\overset{(3)}{=} \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}^2_{A_j}(\mathrm{X}_i)\right) + \sum_{i_1=1}^{n} \sum_{\substack{i_2=1 \\ i_1 \neq i_1}}^{n} \mathbb{E}\left(\mathbb{1}_{A_j}(\mathrm{X}_{i_1}) \cdot \mathbb{1}_{A_j}(\mathrm{X}_{i_2})\right) \\
&\overset{(4)}{=} \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}_{A_j}(\mathrm{X}_i)\right) + \sum_{i_1=1}^{n} \sum_{\substack{i_2=1 \\ i_1 \neq i_1}}^{n} \mathbb{E}\left(\mathbb{1}_{A_j}(\mathrm{X}_{i_1})\right) \cdot \mathbb{E}\left(\mathbb{1}_{A_j}(\mathrm{X}_{i_2})\right) \\
&\overset{(5)}{=} n \cdot \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j) + \sum_{i_1=1}^{n} \sum_{\substack{i_2=1 \\ i_1 \neq i_1}}^{n} \left(\triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)\right)^2 \\
&\overset{(2)}{=} n \cdot \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j) + n \cdot (n-1) \cdot \left(\triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)\right)^2
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left(\mathrm{Z}_j^2\right) &\overset{(1)}{=} \mathbb{E}\left(\left(\frac{1}{n \cdot \triangle_{n_\mathrm{b}}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(\mathrm{X}_i)\right)^2\right) \\
&\overset{(3)}{=} \frac{1}{(n \cdot \triangle_{n_\mathrm{b}})^2} \mathbb{E}\left(\left(\sum_{i=1}^{n} \mathbb{1}_{A_j}(\mathrm{X}_i)\right)^2\right)
\end{aligned}$$

$$\overset{(5)}{=} \frac{n \cdot \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j) + n \cdot (n-1) \cdot (\triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j))^2}{(n \cdot \triangle_{n_\mathrm{b}})^2}$$
$$\overset{(2)}{=} \frac{\varphi(\xi_j) + (n-1) \cdot \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)^2}{n \cdot \triangle_{n_\mathrm{b}}}.$$

4. Consequently, by (1) definition of the variance, (2) basic tranformations, (3) linearity of expectation and (4) the above results, for each $n_\mathrm{b}, n \in \mathbb{N}$ and $j \in \{0, \ldots, n_\mathrm{b} - 1\}$, we obtain

$$\begin{aligned}
\mathrm{Var}\left(\mathrm{Z}_j\right) &\overset{(1,2,3)}{=} \mathbb{E}\left(\mathrm{Z}_j^2\right) - \mathbb{E}\left(\mathrm{Z}_j\right)^2 \\
&\overset{(4)}{=} \frac{\varphi(\xi_j) + (n-1) \cdot \varphi(\xi_j)^2}{n \cdot \triangle_{n_\mathrm{b}}} - \varphi(\xi_j)^2 \\
&\overset{(2)}{=} \frac{\varphi(\xi_j) - \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)^2 + n \cdot \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)^2 - n \cdot \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)^2}{n \cdot \triangle_{n_\mathrm{b}}} \\
&\overset{(2)}{=} \frac{\varphi(\xi_j)\left(1 - \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)\right)}{n \cdot \triangle_{n_\mathrm{b}}}.
\end{aligned}$$

5. Consequently, by (1) definition of φ_{n,n_b}, (2) linearity of expectation and (3) the above results, for each $n_\mathrm{b}, n \in \mathbb{N}$ and $x \in [x_{\min}, x_{\max}]$, we obtain,

$$\begin{aligned}
\mathbb{E}\left(\varphi_{n,n_\mathrm{b}}(x)\right) &\overset{(1)}{=} \mathbb{E}\left(\sum_{j=0}^{n_\mathrm{b}-1} \mathrm{Z}_j \cdot \mathbb{1}_{A_j}(x)\right) \\
&\overset{(2)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \mathbb{E}\left(\mathrm{Z}_j\right) \cdot \mathbb{1}_{A_j}(x) \\
&\overset{(3)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \varphi(\xi_j) \cdot \mathbb{1}_{A_j}(x).
\end{aligned}$$

6. Consequently, by (1) definition of φ_{n,n_b}, (2) basic transformations, (3) the fact that the bins A_{j_1} and A_{j_2} are disjoint for each $j_1, j_2 \in \{0, \ldots, n_\mathrm{b} - 1\}$ with $j_1 \neq j_2$, (4) linearity of expectation and (5) the above results, for each $n_\mathrm{b}, n \in \mathbb{N}$ and $x \in [x_{\min}, x_{\max}]$, we obtain,

$$\begin{aligned}
\mathbb{E}\left(\varphi_{n,n_\mathrm{b}}^2(x)\right) &\overset{(1)}{=} \mathbb{E}\left(\left(\sum_{j=0}^{n_\mathrm{b}-1} \mathrm{Z}_j \cdot \mathbb{1}_{A_j}(x)\right)^2\right) \\
&\overset{(2)}{=} \mathbb{E}\left(\left(\sum_{j=0}^{n_\mathrm{b}-1} \mathrm{Z}_j^2 \cdot \mathbb{1}_{A_j}^2(x)\right) + \left(\sum_{j_1=0}^{n_\mathrm{b}-1} \sum_{\substack{j_2=0 \\ j_1 \neq j_2}}^{n_\mathrm{b}-1} \mathrm{Z}_{j_1} \cdot \mathrm{Z}_{j_2} \cdot \underbrace{\mathbb{1}_{A_{j_1}}(x) \cdot \mathbb{1}_{A_{j_2}}(x)}_{\overset{(3)}{=} 0}\right)\right) \\
&\overset{(4)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \mathbb{E}\left(\mathrm{Z}_j^2\right) \cdot \mathbb{1}_{A_j}(x)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left(\varphi_{n,n_\mathrm{b}}(x)\right)^2 &\overset{(5)}{=} \left(\sum_{j=0}^{n_\mathrm{b}-1} \varphi(\xi_j)\cdot \mathbb{1}_{A_j}(x)\right)^2 \\
&\overset{(2)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \varphi(\xi_j)^2 \cdot \mathbb{1}_{A_j}^2(x) + \sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{\substack{j_2=0\\ j_1\neq j_2}}^{n_\mathrm{b}-1} \varphi(\xi_{j_1})\cdot\varphi(\xi_{j_2})\cdot \underbrace{\mathbb{1}_{A_{j_1}}(x)\cdot\mathbb{1}_{A_{j_2}}(x)}_{\overset{(3)}{=}0} \\
&\overset{(2)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \varphi(\xi_j)^2 \cdot \mathbb{1}_{A_j}(x) \\
&\overset{(5)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \mathbb{E}\left(\mathrm{Z}_j\right)^2 \cdot \mathbb{1}_{A_j}(x).
\end{aligned}$$

7. Consequently, by (1) definition of the variance, (2) basic tranformations, (3) linearity of expectation and (4) the above results, for each $n_\mathrm{b}, n \in \mathbb{N}$ and $x \in [x_{\min}, x_{\max}]$, we obtain,

$$\begin{aligned}
\mathrm{Var}\left(\varphi_{n,n_\mathrm{b}}(x)\right) &\overset{(1,2,3)}{=} \mathbb{E}\left(\varphi_{n,n_\mathrm{b}}^2(x)\right) - \mathbb{E}\left(\varphi_{n,n_\mathrm{b}}(x)\right)^2 \\
&\overset{(4)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \mathbb{E}\left(\mathrm{Z}_j^2\right)\cdot\mathbb{1}_{A_j}(x) - \sum_{j=0}^{n_\mathrm{b}-1} \mathbb{E}\left(\mathrm{Z}_j\right)^2 \mathbb{1}_{A_j}(x) \\
&\overset{(2)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \left(\mathbb{E}\left(\mathrm{Z}_j^2\right) - \mathbb{E}\left(\mathrm{Z}_j\right)^2\right)\cdot\mathbb{1}_{A_j}(x) \\
&\overset{(1,2,3)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \mathrm{Var}\left(\mathrm{Z}_j\right)\cdot\mathbb{1}_{A_j}(x) \\
&\overset{(4)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \frac{\varphi(\xi_j)\left(1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j)\right)}{n\cdot\triangle_{n_\mathrm{b}}}\cdot\mathbb{1}_{A_j}(x).
\end{aligned}$$

□

Remark 3. [19] shows equation (18) and (19) more efficiently by leveraging that the counts $\sum_{i=1}^{n}\mathbb{1}_{A_j}(X_i)$ are Bernoulli distributed with $p_j := \int_{A_j}\varphi(x)\,dx = \triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j)$ (which is equal to $\mathbb{E}(\mathbb{1}_{A_j}(X_i))$, as is can be seen in our proof) and then using that for such a distribution, expectation and variance are known. This allows to conclude step 4 directly without the computations in step 3 (we though think this can still be helpful in preparation of step 6 and 7). However, as it can be seen in our proof, given knowledge of $\mathbb{E}(\sum_{i=1}^{n}\mathbb{1}_{A_j}(X_i)) = \mathbb{E}(n\cdot\triangle_{n_\mathrm{b}}\cdot\mathrm{Z}_j)$ and $\mathrm{Var}(\sum_{i=1}^{n}\mathbb{1}_{A_j}(X_i)) = \mathrm{Var}(n\ \triangle_{n_\mathrm{b}}\ \mathrm{Z}_j)$, it still requires some argumentation to conclude equation (18) and (19) rigorously (step 5 to 7). We conduct this argumentation.

Lemma 3 (Convergence of normalized histogram counts). *In the setting of definition 5, for each $n_b \in \mathbb{N}$ and each $j \in \{0, \ldots, n_b - 1\}$, there exists a $\xi_j \in \overline{A}_j$, such that $\mathrm{Z}_j \to \varphi(\xi_j)$ as $n \to \infty$ $\mathbb{P}$-almost surely.*

Proof. Let $n_\mathrm{b} \in \mathbb{N}$ and $j \in \{0, ..., n_\mathrm{b} - 1\}$.

- **$\mathbb{P}$-Integrability:** In the proof of lemma 2, we already showed that $\mathbb{1}_{A_j}(\mathrm{X}_i)$ is $\mathcal{F}$-$\mathcal{B}$-measureable and $\mathbb{P}$-integrable ($\mathcal{L}^1(\mathbb{P})$) with $\mathbb{E}(|\mathbb{1}_{A_j}(\mathrm{X}_i)|) = \mathbb{E}(\mathbb{1}_{A_j}(\mathrm{X}_i)) = \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j) < \infty$ for all $i = 1, .., n$ (the latter follows from the fact that φ is continuous and has a compact support, and thus, is bounded).
- **$\mathbb{P}$-Independence:** Also in the proof of lemma 2, we already showed that the random variables are $\mathbb{1}_{A_j}(\mathrm{X}_1), \ldots, \mathbb{1}_{A_j}(\mathrm{X}_n)$ iid w.r.t. $\mathbb{P}$.
- **Convergence:** Using the results from above, we can apply the law of large numbers, yielding

$$\triangle_{n_\mathrm{b}} \cdot \mathrm{Z}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{A_j}(\mathrm{X}_i) \overset{n \to \infty}{\longrightarrow} \mathbb{E}\left(\mathbb{1}_{A_j}(\mathrm{X}_1)\right) = \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j)$$

  $\mathbb{P}$-almost surely.

□

Theorem 2 (Uniform convergence of approximated density). *For each $n_b \in \mathbb{N}$, choose $n = {n_b}^2$. Then the sequence of approximated densities with probabilistic character $(\varphi_{n,n_b})_{n,n_b \in \mathbb{N}}$ converges uniformly towards the true density φ $\mathbb{P}$-almost surely as $n, n_b \to \infty$. That is, $\varphi_{n,n_b}(x) \to \varphi(x)$ $\mathbb{P}$-almost surely as $n, n_b \to \infty$ for all $x \in \mathbb{R}$ (and independently on x).*

Proof.

- **The variance of the approximated density converges towards zero as $n, n_\mathbf{b} \to \infty$:** We want to show that the variance $\mathrm{Var}(\varphi_{n,n_\mathrm{b}})$ converges uniformly towards zero as $n_\mathrm{b} \to \infty$, i.e., that for each $\epsilon > 0$, there exists an $\eta(\epsilon) \in \mathbb{N}$, such that for all $n_\mathrm{b} \geq \eta(\epsilon)$ and all $x \in \mathbb{R}$,

$$|\operatorname{Var}(\varphi_{n,n_\mathrm{b}}(x))| = \operatorname{Var}(\varphi_{n,n_\mathrm{b}}(x)) < \epsilon$$

  holds.
  Let $\epsilon > 0$. Since the density φ is continuous and has compact support, it is bounded. Consequently, $\varphi_{\max} := \max_{x \in \mathbb{R}} \varphi(x)$ exists. Therefore, we can choose $\eta(\epsilon) := \lceil \frac{\varphi_{\max}}{(x_{\max} - x_{\min})\epsilon} \rceil + 1$ and let $n_\mathrm{b} \geq \eta(\epsilon)$ and $x \in \mathbb{R}$.
  By (1) definition of the variance, (2) theorem 2 and (3) the choice of $n = {n_\mathrm{b}}^2$ with $n_\mathrm{b} = \frac{x_{\max} - x_{\min}}{\triangle_{n_\mathrm{b}}}$, we obtain

$$\begin{aligned} |\operatorname{Var}(\varphi_{n,n_\mathrm{b}}(x))| &\overset{(1)}{=} \operatorname{Var}(\varphi_{n,n_\mathrm{b}}(x)) \\ &\overset{(2)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \frac{\varphi(\xi_j)\,(1 - \triangle_{n_\mathrm{b}} \cdot \varphi(\xi_j))}{n \cdot \triangle_{n_\mathrm{b}}} \cdot \mathbb{1}_{A_j}(x) \end{aligned}$$

$$\overset{(3)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \triangle_{n_\mathrm{b}} \cdot \frac{\varphi(\xi_j)\,(1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j))}{(x_{\max}-x_{\min})^2} \cdot \mathbb{1}_{A_j}(x)$$
$$\overset{(2)}{=} \sum_{j=0}^{n_\mathrm{b}-1} \frac{1}{n_\mathrm{b}} \cdot \frac{\varphi(\xi_j)\,(1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j))}{x_{\max}-x_{\min}} \cdot \mathbb{1}_{A_j}(x).$$

Again by definition of the variance and lemma 2,

$$0 \leq \mathrm{Var}\,(\mathrm{Z}_j) = \frac{\varphi(\xi_j)\,(1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j))}{n\cdot\triangle_{n_\mathrm{b}}}$$

holds, which is equivalent to

$$0 \leq \varphi(\xi_j)\,(1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j))\,.$$

Even more, since the density φ is non-negative, $1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j) \geq 0$ needs to hold too. Additionally, $\triangle_{n_\mathrm{b}} = \frac{x_{\max}-x_{\min}}{n_\mathrm{b}}$ is non-negative by construction and choice of $n_\mathrm{b} \in \mathbb{N}$. Therefore, in summary, we observe that $1 \geq 1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j) \geq 0$ and thus,

$$\varphi(\xi_j)\,(1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j)) \leq \varphi(\xi_j) \leq \varphi_{\max}$$

holds. Consequently, we can estimate the sum, which has only positive summands, by

$$\sum_{j=0}^{n_\mathrm{b}-1} \frac{1}{n_\mathrm{b}} \cdot \frac{\varphi(\xi_j)\,(1-\triangle_{n_\mathrm{b}}\cdot\varphi(\xi_j))}{x_{\max}-x_{\min}} \cdot \mathbb{1}_{A_j}(x) \leq \sum_{j=0}^{n_\mathrm{b}-1} \frac{1}{n_\mathrm{b}} \cdot \frac{\varphi_{\max}}{x_{\max}-x_{\min}} \cdot \mathbb{1}_{A_j}(x).$$

Finally, by (1) the above investigations, (2) the fact that due to the disjointness of the bins A_j for all $j = 0, ..., n_\mathrm{b}-1$, each $x \in \mathbb{R}$ lies in exactly one bin A_j, (3) the choice of $n_\mathrm{b} \geq \eta(\epsilon)$ and (4) the choice of $\eta(\epsilon) := \lceil \frac{\varphi_{\max}}{(x_{\max}-x_{\min})\epsilon} \rceil + 1$, we obtain

$$\begin{aligned}
|\,\mathrm{Var}\,(\varphi_{n,n_\mathrm{b}}(x))\,| &\overset{(1)}{\leq} \sum_{j=0}^{n_\mathrm{b}-1} \frac{1}{n_\mathrm{b}} \cdot \frac{\varphi_{\max}}{x_{\max}-x_{\min}} \cdot \mathbb{1}_{A_j}(x) \\
&\overset{(2)}{\leq} \frac{1}{n_\mathrm{b}} \cdot \frac{\varphi_{\max}}{x_{\max}-x_{\min}} \\
&\overset{(3)}{\leq} \frac{1}{\eta(\epsilon)} \cdot \frac{\varphi_{\max}}{x_{\max}-x_{\min}} \\
&\overset{(4)}{<} \frac{(x_{\max}-x_{\min})\,\epsilon}{\varphi_{\max}} \cdot \frac{\varphi_{\max}}{x_{\max}-x_{\min}} \\
&= \epsilon,
\end{aligned}$$

which was to show.

Since the variance $\mathrm{Var}(\varphi_{n,n_\mathrm{b}})$ converges uniformly towards zero as $n_\mathrm{b} \to \infty$, for each $x \in \mathbb{R}$, the density $\varphi_{n,n_\mathrm{b}}(x)$ converges uniformly towards a constant value as $n = {n_\mathrm{b}}^2, n_\mathrm{b} \to \infty$. In the next steps, we show that this limit is equal to $\varphi(x)$.

– **The approximated density converges towards its expectation as $n \to \infty$:** We want to show that the density φ_{n,n_b} converges uniformly towards $\mathbb{E}(\varphi_{n,n_\mathrm{b}})$ $\mathbb{P}$-almost surely for each $n_\mathrm{b} \in \mathbb{N}$ as $n \to \infty$, i.e., that for each $n_\mathrm{b} \in \mathbb{N}$ and for each $\epsilon > 0$, there exists an $\eta(\epsilon, n_\mathrm{b}) \in \mathbb{N}$, such that for all $n \geq \eta(\epsilon, n_\mathrm{b})$ and all $x \in \mathbb{R}$,

$$|\varphi_{n,n_\mathrm{b}}(x) - \mathbb{E}\left(\varphi_{n,n_\mathrm{b}}(x)\right)| < \epsilon$$

holds $\mathbb{P}$-almost surely.
Let $\epsilon > 0$. By lemma 3, for each $n_\mathrm{b} \in \mathbb{N}$ and $j \in \{0, \ldots, n_\mathrm{b} - 1\}$, there exists a $\xi_j \in \overline{A_j}$, such that for each $\epsilon > 0$, there exists an $\eta(\epsilon, n_\mathrm{b}, j) \in \mathbb{N}$, such that for all $n \geq \eta(\epsilon, n_\mathrm{b}, j)$, $|\mathrm{Z}_j - \varphi(\xi_j)| < \epsilon$ holds $\mathbb{P}$-almost surely. Therefore, we can choose $\eta(\epsilon, n_\mathrm{b}) = \max_{j \in \{0,\ldots,n_\mathrm{b}-1\}} \eta(\epsilon, n_\mathrm{b}, j)$ and let $n \geq \eta(\epsilon, n_\mathrm{b})$ and $x \in \mathbb{R}$.
By (1) by definition of the density φ_{n,n_b}, (2) lemma 2, (3) basic transformations, (4) the choice of $n \geq \eta(\epsilon, n_\mathrm{b})$ and (5) the fact that due to the disjointness of the bins A_j for all $j = 0, ..., n_\mathrm{b} - 1$, each $x \in \mathbb{R}$ lies in exactly one bin A_j, we obtain

$$\begin{aligned}
|\varphi_{n,n_\mathrm{b}}(x) - \mathbb{E}\left(\varphi_{n,n_\mathrm{b}}(x)\right)| &\overset{(1,2)}{=} |\sum_{j=0}^{n_\mathrm{b}-1} \mathrm{Z}_j \cdot \mathbb{1}_{A_j}(x) - \sum_{j=0}^{n_\mathrm{b}-1} \varphi(\xi_j) \cdot \mathbb{1}_{A_j}(x)| \\
&\overset{(3)}{\leq} \sum_{j=0}^{n_\mathrm{b}-1} |\mathrm{Z}_j - \varphi(\xi_j)| \cdot \mathbb{1}_{A_j}(x) \\
&\overset{(4)}{<} \sum_{j=0}^{n_\mathrm{b}-1} \epsilon \cdot \mathbb{1}_{A_j}(x) \\
&\overset{(5)}{=} \epsilon,
\end{aligned}$$

which was to show.

– **The expectation of the approximated density converges towards the density as $n_\mathbf{b} \to \infty$:** We want to show that the expectation $\mathbb{E}(\varphi_{n,n_\mathrm{b}})$ converges uniformly towards φ as $n_\mathrm{b} \to \infty$, i.e., that for each $\epsilon > 0$, there exists an $\eta(\epsilon) \in \mathbb{N}$, such that for all $n_\mathrm{b} \geq \eta(\epsilon)$ and all $x \in \mathbb{R}$,

$$|\mathbb{E}\left(\varphi_{n,n_\mathrm{b}}(x)\right) - \varphi(x)| < \epsilon$$

holds.
Let $\epsilon > 0$. Since the density φ is continuous and has compact support, it is uniformly continuous. Consequently, for each $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that for any $x, \xi \in \mathbb{R}$ such that $|x - \xi| < \delta$ holds implies $|\varphi(x) - \varphi(\xi)| < \epsilon$. Therefore, we can choose $\eta(\epsilon) := \lceil \frac{x_\mathrm{max} - x_\mathrm{min}}{\delta(\epsilon)} \rceil + 1$ and let $n_\mathrm{b} \geq \eta(\epsilon)$ and $x \in \mathbb{R}$.

By (1) lemma 2 and (2) the fact that due to the disjointness of the bins A_j for all $j = 0, ..., n_\mathrm{b} - 1$, each $x \in \mathbb{R}$ lies in exactly one bin A_j, we obtain

$$|\mathbb{E}\left(\varphi_{n,n_\mathrm{b}}(x)\right) - \varphi(x)| \overset{(1)}{=} |\sum_{j=0}^{n_\mathrm{b}-1} \varphi(\xi_j) \cdot \mathbb{1}_{A_j}(x) - \varphi(x)| \overset{(2)}{=} |\varphi(\xi_j) - \varphi(x)|$$

for some $j \in \{0, \ldots, n_\mathrm{b}\}$, such that $x, \xi_j \in \overline{A}_j$ holds (cf. lemma 3). Therefore, by (1) definition of A_j (cf. definition 5), (2) definition of $\triangle a$ (cf. definition 5), (3) the choice of $n_\mathrm{b} \geq \eta(\epsilon)$ and (4) the choice of $\eta(\epsilon) := \lceil \frac{x_\mathrm{max} - x_\mathrm{min}}{\delta(\epsilon)} \rceil + 1$, we obtain

$$|\xi_j - x| \overset{(1,2)}{\leq} \triangle a \overset{(2)}{=} \frac{x_\mathrm{max} - x_\mathrm{min}}{n_\mathrm{b}} \overset{(3)}{\leq} \frac{x_\mathrm{max} - x_\mathrm{min}}{\eta(\epsilon)} \overset{(4)}{<} \delta(\epsilon)$$

and consequently, by the uniform continuity of φ, we obtain

$$|\mathbb{E}\left(\varphi_{n,n_\mathrm{b}}(x)\right) - \varphi(x)| = |\varphi(\xi_j) - \varphi(x)| < \epsilon,$$

which was to show.

In summary, φ_{n,n_b} converges uniformly towards φ $\mathbb{P}$-almost surely as $n, n_\mathrm{b} \to \infty$. Note that the $\mathbb{P}$-almost sure uniform convergence of φ_{n,n_b} towards $\mathbb{E}(\varphi_{n,n_\mathrm{b}})$ as $n \to \infty$ is a convergence in dependence of the binning determined by n_b. This justifies why n needs to be chosen in dependence of n_b. The analysis of the variance shows in more detail that the number of samples n has to grow sufficiently faster than the number of bins n_b, e.g, guaranteed by $n = {n_\mathrm{b}}^2$.

□

In the rest of this subsection, we want to leverage the possibility to approximate an unknown density function φ by an elementary density function $\hat{\varphi} := \varphi_{n,n_\mathrm{b}}$ to define an approximated version of the mutual information and the entropy for continuously distributed random variables.

Definition 6 (Mutual information). *Let H_v and H_u be two real-valued random variables. Let $\varphi_{vu} : \mathbb{R} \times \mathbb{R} \longrightarrow [0, \infty)$ and $\varphi_v, \varphi_u : \mathbb{R} \longrightarrow [0, \infty)$ denote the densities according to the joint distribution $(\mathrm{H}_v, \mathrm{H}_u)$, and the marginal distribution of H_v and H_u, respectively. Let λ the Lebesgue-measure.*
The mutual information is defined by

$$\begin{aligned} I(\mathrm{H}_v, \mathrm{H}_u) =& I(\varphi_v, \varphi_u, \varphi_{vu}) \\ :=& \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi_{vu}(h_v, h_u) \cdot \log\left(\frac{\varphi_{vu}(h_v, h_u)}{\varphi_v(h_v) \cdot \varphi_u(h_u)}\right) \, d\lambda(h_v) \, d\lambda(h_u). \end{aligned}$$

Definition 7 (Approximated mutual information). *Let H_v and H_u be two real-valued random variables. Let $\varphi_{vu} : \mathbb{R} \times \mathbb{R} \longrightarrow [0, \infty)$ and $\varphi_v, \varphi_u : \mathbb{R} \longrightarrow [0, \infty)$ denote the unknown densities according to the joint distribution $(\mathrm{H}_v, \mathrm{H}_u)$, and the marginal distribution of H_v and H_u, respectively.*
The approximated mutual information $\hat{I}(\mathrm{H}_v, \mathrm{H}_u) = \hat{I}(\varphi_v, \varphi_u, \varphi_{vu})$ is defined as the mutual information $I(\hat{\varphi}_v, \hat{\varphi}_u, \hat{\varphi}_{vu})$ with approximated densities according to definition 5.

Theorem 3 (Convergence of approximated mutual information). *For each $n_b \in \mathbb{N}$, choose $n = {n_b}^2$. In the setting of definition 6 and 7, the sequence of approximated mutual information with probabilistic character converges towards the true mutual information $\mathbb{P}$-almost surely as $n, n_b \to \infty$. That is, $\hat{I}(\mathrm{H}_v, \mathrm{H}_u) \to I(\mathrm{H}_v, \mathrm{H}_u)$ $\mathbb{P}$-almost surely as $n, n_b \to \infty$.*

Proof. By theorem 2, the sequences of approximated densities with probabilistic character $(\hat{\varphi}_v)_{n,n_\mathrm{b}\in\mathbb{N}}$, $(\hat{\varphi}_u)_{n,n_\mathrm{b}\in\mathbb{N}}$ and $(\hat{\varphi}_{vu})_{n,n_\mathrm{b}\in\mathbb{N}}$ converge uniformly towards the true densities φ_v, φ_u and φ_{vu} $\mathbb{P}$-almost surely as $n, n_\mathrm{b} \to \infty$, respectively.

Consequently, since the function $\mathbb{R}^3_{\geq 0} \to \mathbb{R},\ (x, y, z) \mapsto x \cdot \log(\frac{x}{yz})$ is clearly continuous, by common results from analysis, the sequence of approximated integrands with probabilistic character

$$\hat{\varphi}_{vu} \cdot \log\left(\frac{\hat{\varphi}_{vu}}{\hat{\varphi}_v \cdot \hat{\varphi}_u}\right)$$

converges uniformly towards the true integrand

$$\varphi_{vu} \cdot \log\left(\frac{\varphi_{vu}}{\varphi_v \cdot \varphi_u}\right)$$

$\mathbb{P}$-almost surely as $n, n_\mathrm{b} \to \infty$.

Finally, by (1) the definition of the approximated mutual information (definition 7), (2) the definition of the mutual information (definition 6), (3) the fact that all appearing densities have a compact support $[h_{v,\min}, h_{v,\max}]$, $[h_{u,\min}, h_{u,\max}]$ and $[h_{v,\min}, h_{v,\max}] \times [h_{u,\min}, h_{u,\max}]$, respectively, and (4) the common result from analysis that the uniform convergence of a sequence of functions on a compact (i.e., a closed and bounded) interval implies the convergence of the integrals,

$$\begin{aligned}
&\hat{I}(\mathrm{H}_v, \mathrm{H}_u) \\
&\overset{(1)}{=} I(\hat{\varphi}_v, \hat{\varphi}_u, \hat{\varphi}_{vu}) \\
&\overset{(2)}{=} \int_{\mathbb{R}}\int_{\mathbb{R}} \hat{\varphi}_{vu}(h_v, h_u) \cdot \log\left(\frac{\hat{\varphi}_{vu}(h_v, h_u)}{\hat{\varphi}_v(h_v) \cdot \hat{\varphi}_u(h_u)}\right)\, d\lambda(h_v)\, d\lambda(h_u) \\
&\overset{(3)}{=} \int_{h_{u,\min}}^{h_{u,\max}}\int_{h_{v,\min}}^{h_{v,\max}} \hat{\varphi}_{vu}(h_v, h_u) \cdot \log\left(\frac{\hat{\varphi}_{vu}(h_v, h_u)}{\hat{\varphi}_v(h_v) \cdot \hat{\varphi}_u(h_u)}\right)\, d\lambda(h_v)\, d\lambda(h_u) \\
&\overset{(4)}{\longrightarrow} \int_{h_{u,\min}}^{h_{u,\max}}\int_{h_{v,\min}}^{h_{v,\max}} \varphi_{vu}(h_v, h_u) \cdot \log\left(\frac{\varphi_{vu}(h_v, h_u)}{\varphi_v(h_v) \cdot \varphi_u(h_u)}\right)\, d\lambda(h_v)\, d\lambda(h_u) \\
&\overset{(3)}{=} \int_{\mathbb{R}}\int_{\mathbb{R}} \varphi_{vu}(h_v, h_u) \cdot \log\left(\frac{\varphi_{vu}(h_v, h_u)}{\varphi_v(h_v) \cdot \varphi_u(h_u)}\right)\, d\lambda(h_v)\, d\lambda(h_u) \\
&\overset{(2)}{=} I(\varphi_v, \varphi_u, \varphi_{vu}) \\
&\overset{(2)}{=} I(\mathrm{H}_v, \mathrm{H}_u)
\end{aligned}$$

holds $\mathbb{P}$-almost surely as $n, n_\mathrm{b} \to \infty$. □

Fortunately, the approximated mutual information, which converges towards the true mutual information as the number of samples n and the number of bins n_b converges towards infinity, is much easier to compute:

Theorem 4 (Approximated mutual information). *In the setting of definition 5 and 7, we denote the samples drawn iid from* H_v *and* H_u *by* $\tilde{h}_{1v}, ..., \tilde{h}_{nv}$ *and* $\tilde{h}_{1u}, ..., \tilde{h}_{nu}$*, respectively. Then*

$$\hat{I}(\mathrm{H}_v, \mathrm{H}_u) = \sum_{j_1=0}^{n_b-1} \sum_{j_2=0}^{n_b-1} p_{j_1,j_2}(v,u) \cdot \log\Big(\frac{p_{j_1,j_2}(v,u)}{p_{j_1}(v) \cdot p_{j_2}(u)}\Big)$$

with

$$p_{j_1,j_2}(v,u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{A_{j_1}\times A_{j_2}}(\tilde{h}_{iv}, \tilde{h}_{iu}), \quad p_j(v) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{A_j}(\tilde{h}_{iv})$$

holds.

Proof. By definition 5, we approximate the density φ_v by

$$\hat{\varphi}_v(h_v) = \sum_{j=0}^{n_\mathrm{b}-1} z_j \cdot \mathbb{1}_{A_j}(h_v) \quad \text{with} \quad z_j = \frac{1}{n \cdot \triangle_{n_\mathrm{b}}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(\tilde{h}_{iv})$$

for all $j = 0, ..., n_\mathrm{b} - 1$ and all $h_v \in \mathbb{R}$. Analogously, we approximate the density φ_u by $\hat{\varphi}_u$. Similarly, by definition 5 and remark 2, we approximate the density φ_{vu} by

$$\hat{\varphi}_{vu}(h_v, h_u) = \sum_{j_1=0}^{n_\mathrm{b}-1} \sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_1 j_2} \cdot \mathbb{1}_{A_{j_1}\times A_{j_2}}(h_v, h_u) \quad \text{with}$$

$$z_{j_1 j_2} = \frac{1}{n \cdot (\triangle_{n_\mathrm{b}})^2} \sum_{i=1}^{n} \mathbb{1}_{A_{j_1}\times A_{j_2}}(\tilde{h}_{iv}, \tilde{h}_{iu})$$

for all $j_1, j_2 = 0, ..., n_\mathrm{b} - 1$ and all $h_v, h_u \in \mathbb{R}$.

Consequently, by (1) the definition of the approximated mutual information (definition 7), (2) the definition of the mutual information (definition 6), (3) the definition of approximating densities (definition 5), (4) linearity of the Lebesgue-integral, (5) the definition of the elementary functions $\mathbb{1}_{A_j}$ and $\mathbb{1}_{A_{j_1}\times A_{j_2}}$ for all $j, j_1, j_2 = 0, ..., n_\mathrm{b} - 1$, (6) basic properties from measure and probability theory, (7) the fact that $\lambda(A_j) = a_{j+1} - a_j = \triangle_{n_\mathrm{b}}$ holds for all $j = 0, ..., n_\mathrm{b} - 1$ and (8) the definition of p_j and p_{j_1,j_2} for all $j, j_1, j_2 = 0, ..., n_\mathrm{b} - 1$, we obtain

$$\begin{aligned}
&\hat{I}(\mathrm{H}_v, \mathrm{H}_u) \\
&\overset{(1)}{=} I(\hat{\varphi}_v, \hat{\varphi}_u, \hat{\varphi}_{vu}) \\
&\overset{(2)}{=} \int_{\mathbb{R}}\int_{\mathbb{R}} \hat{\varphi}_{vu}(h_v, h_u) \cdot \log\left(\frac{\hat{\varphi}_{vu}(h_v, h_u)}{\hat{\varphi}_v(h_v) \cdot \hat{\varphi}_u(h_u)}\right) \, d\lambda(h_v) \, d\lambda(h_u)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(3)}{=} \int_{\mathbb{R}}\int_{\mathbb{R}} \left(\sum_{j_1=0}^{n_\mathrm{b}-1} \sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_1 j_2} \cdot \mathbb{1}_{A_{j_1}\times A_{j_2}}(h_v, h_u) \right) \cdot \log\left(\frac{\sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_1 j_2}\cdot \mathbb{1}_{A_{j_1}\times A_{j_2}}(h_v,h_u)}{\left(\sum_{j_1=0}^{n_\mathrm{b}-1} z_{j_1}\cdot \mathbb{1}_{A_{j_1}}(h_v)\right)\left(\sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_2}\cdot \mathbb{1}_{A_{j_2}}(h_u)\right)} \right) \, d\lambda(h_v)\, d\lambda(h_u) \\
&\overset{(4)}{=} \sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_1 j_2} \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{A_{j_1}\times A_{j_2}}(h_v,h_u) \cdot \log\left(\frac{\sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_1 j_2}\cdot \mathbb{1}_{A_{j_1}\times A_{j_2}}(h_v,h_u)}{\left(\sum_{j_1=0}^{n_\mathrm{b}-1} z_{j_1}\cdot \mathbb{1}_{A_{j_1}}(h_v)\right)\left(\sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_2}\cdot \mathbb{1}_{A_{j_2}}(h_u)\right)} \right) \, d\lambda(h_v)\, d\lambda(h_u) \\
&\overset{(5)}{=} \sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_1 j_2} \int_{A_{j_2}}\int_{A_{j_1}} \log\left(\frac{z_{j_1 j_2}}{z_{j_1} z_{j_2}}\right) \, d\lambda(h_v)\, d\lambda(h_u) \\
&\overset{(6)}{=} \sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{j_2=0}^{n_\mathrm{b}-1} z_{j_1 j_2} \cdot \lambda(A_{j_1}) \cdot \lambda(A_{j_2}) \cdot \log\left(\frac{z_{j_1 j_2}}{z_{j_1} z_{j_2}}\right) \\
&\overset{(3,7)}{=} \sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{j_2=0}^{n_\mathrm{b}-1} (\triangle_{n_\mathrm{b}})^2 \cdot \left(\frac{1}{n\cdot(\triangle_{n_\mathrm{b}})^2} \sum_{i=1}^{n} \mathbb{1}_{A_{j_1}\times A_{j_2}}(\tilde{h}_{iv}, \tilde{h}_{iu}) \right) \cdot \log\left(\frac{\frac{1}{n\cdot(\triangle_{n_\mathrm{b}})^2}\sum_{i=1}^{n} \mathbb{1}_{A_{j_1}\times A_{j_2}}(\tilde{h}_{iv},\tilde{h}_{iu})}{\left(\frac{1}{n\cdot\triangle_{n_\mathrm{b}}}\sum_{i=1}^{n}\mathbb{1}_{A_{j_1}}(\tilde{h}_{iv})\right)\left(\frac{1}{n\cdot\triangle_{n_\mathrm{b}}}\sum_{i=1}^{n}\mathbb{1}_{A_{j_2}}(\tilde{h}_{iu})\right)} \right) \\
&\overset{(8)}{=} \sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{j_2=0}^{n_\mathrm{b}-1} p_{j_1,j_2}(v,u) \cdot \log\left(\frac{p_{j_1,j_2}(v,u)}{p_{j_1}(v)\cdot p_{j_2}(u)}\right).
\end{aligned}
$$

Note that we tacitly assumed that the integrand in the Lebesgue-integral above is indeed Lebesgue-integrable. Of course, we do catch up on this now: Using that the density is non-negative, along similar lines, we can show that

$$
\begin{aligned}
&\int_{\mathbb{R}}\int_{\mathbb{R}} \left| \hat{\varphi}_{vu}(h_v,h_u) \cdot \log\left(\frac{\hat{\varphi}_{vu}(h_v,h_u)}{\hat{\varphi}_v(h_v)\cdot\hat{\varphi}_u(h_u)}\right) \right| \, d\lambda(h_v)\, d\lambda(h_u) \\
&= \sum_{j_1=0}^{n_\mathrm{b}-1}\sum_{j_2=0}^{n_\mathrm{b}-1} p_{j_1,j_2}(v,u) \cdot \left| \log\left(\frac{p_{j_1,j_2}(v,u)}{p_{j_1}(v)\cdot p_{j_2}(u)}\right) \right|
\end{aligned}
$$

holds. This term is finite as long as $p_{j_1}(v) = 0$ or $p_{j_2}(u) = 0$ implies $p_{j_1,j_2}(v,u) = 0$ for any two $j_1, j_2 \in \{0,\dots,n_\mathrm{b}-1\}$, since in this case, by convention, $0\cdot\log(\frac{0}{0}) := 0$ holds. However, by definition of $p_{j_1,j_2}(v,u)$, $p_{j_1}(v)$ and $p_{j_2}(u)$, this requirement is satisfied: If without loss of generality (w.l.o.g.), $p_{j_1}(v) = 0$ holds for some $j_1, j_2 \in \{0,\dots,n_\mathrm{b}-1\}$, there does not exist any $i \in \{1,\dots,n\}$ such that $\tilde{h}_{iv} \in A_{j_1}$ holds. As a consequence, there does also not exist any $i \in \{1,\dots,n\}$ such that $(\tilde{h}_{iv}, \tilde{h}_{iu}) \in A_{j_1} \times A_{j_2}$ holds. Consequently, $p_{j_1,j_2}(v,u) = 0$ follows. □

Theorem 4 states that the approximated mutual information $\hat{I}(\mathrm{H}_v, \mathrm{H}_u)$ of two real-valued random variables with unknown densities equals the mutual information of the discrete probability distribution given by the relative amount of sampled observations $\tilde{h}_{iv}$ and $\tilde{h}_{iu}$ for $i = 1, ..., n$ in the discretized bins A_j and $A_{j_1} \times A_{j_2}$ for $j, j_1, j_2 = 0, ..., n_\mathrm{b} - 1$.

Now, we leverage the same concept of density approximation in the case of entropy approximation.

Definition 8 (Entropy). *Let* H_v *be a real-valued random variable. Let* $\varphi_v : \mathbb{R} \longrightarrow [0,\infty)$ *denote the density according to the distribution of* H_v*. Let* λ *the*

Lebesgue-measure.
The entropy is defined by

$$E(\mathrm{H}_v) = E(\varphi_v) := -\int_{\mathbb{R}} \varphi_v(h_v) \cdot \log\Big(\varphi_v(h_v)\Big)\, d\lambda(h_v).$$

Definition 9 (Approximated entropy). *Let* H_v *be a real-valued random variable. Let* $\varphi_v : \mathbb{R} \longrightarrow [0, \infty)$ *denote the unknown density according to the distribution of* H_v*. Let* λ *the Lebesgue-measure.*
The approximated entropy $\hat{E}(\mathrm{H}_v) = \hat{E}(\varphi_v)$ *is defined as the entropy* $E(\hat{\varphi}_v)$ *with approximated density according to definition 5.*

Theorem 5 (Approximated entropy). *In the setting of definition 5 and 9, we denote the samples drawn iid from* H_v *by* $\tilde{h}_{1v}, ..., \tilde{h}_{nv}$*. Then*

$$\hat{E}(\mathrm{H}_v) = -\sum_{j=0}^{n_b-1} p_j(v) \cdot \log\left(p_j(v)\right) + \log(\triangle_{n_b}) \quad \textit{with} \quad p_j(v) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{A_j}(\tilde{h}_{iv})$$

holds.

Proof. The proof is very similar to the one of theorem 4: By definition 5, we approximate the density φ_v by

$$\hat{\varphi}_v(h_v) = \sum_{j=0}^{n_\mathrm{b}-1} z_j \cdot \mathbb{1}_{A_j}(h_v) \quad \text{with} \quad z_j = \frac{1}{n \cdot \triangle_{n_\mathrm{b}}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(\tilde{h}_{iv})$$

for all $j = 0, ..., n_\mathrm{b} - 1$ and all $h_v \in \mathbb{R}$.

Consequently, by (1) the definition of the approximated entropy (definition 9), (2) the definition of the entropy (definition 8), (3) the definition of approximating densities (definition 5), (4) linearity of the Lebesgue-integral, (5) the definition of the elementary functions $\mathbb{1}_{A_j}$ for all $j = 0, ..., n_\mathrm{b}-1$, (6) basic properties from measure and probability theory, (7) the fact that $\lambda(A_j) = a_{j+1} - a_j = \triangle_{n_\mathrm{b}}$ holds for all $j = 0, ..., n_\mathrm{b} - 1$, (8) the definition of p_j for all $j = 0, ..., n_\mathrm{b} - 1$, (9) basic transformations and (10) the fact that due to the disjointness of the bins A_j for all $j = 0, ..., n_\mathrm{b} - 1$, each sample x_i lies in exactly one bin A_j for all $i = 1, ..., n$, we obtain

$$\begin{aligned}
&\hat{E}(\mathrm{H}_v) \\
&\overset{(1)}{=} E(\hat{\varphi}_v) \\
&\overset{(2)}{=} -\int_{\mathbb{R}} \hat{\varphi}_v(h_v) \cdot \log\left(\hat{\varphi}_v(h_v)\right)\, d\lambda(h_v) \\
&\overset{(3)}{=} -\int_{\mathbb{R}} \left(\sum_{j=0}^{n_\mathrm{b}-1} z_j \cdot \mathbb{1}_{A_j}(h_v)\right) \cdot \log\left(\sum_{j=0}^{n_\mathrm{b}-1} z_j \cdot \mathbb{1}_{A_j}(h_v)\right)\, d\lambda(h_v) \\
&\overset{(4)}{=} -\sum_{j=0}^{n_\mathrm{b}-1} z_j \int_{\mathbb{R}} \mathbb{1}_{A_j}(h_v) \cdot \log\left(\sum_{j=0}^{n_\mathrm{b}-1} z_j \cdot \mathbb{1}_{A_j}(h_v)\right)\, d\lambda(h_v)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(5)}{=} -\sum_{j=0}^{n_\mathrm{b}-1} z_j \int_{A_j} \log\left(z_j\right)\ d\lambda(h_v) \\
&\overset{(6)}{=} -\sum_{j=0}^{n_\mathrm{b}-1} z_j \cdot \lambda(A_j) \cdot \log\left(z_j\right) \\
&\overset{(3,7)}{=} -\sum_{j_1=0}^{n_\mathrm{b}-1} \sum_{j_2=0}^{n_\mathrm{b}-1} \triangle_{n_\mathrm{b}} \cdot \left(\frac{1}{n \cdot \triangle_{n_\mathrm{b}}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(\tilde{h}_{iv})\right) \cdot \log\left(\frac{1}{n \cdot \triangle_{n_\mathrm{b}}} \sum_{i=1}^{n} \mathbb{1}_{A_j}(\tilde{h}_{iv})\right) \\
&\overset{(8)}{=} -\sum_{j=0}^{n_\mathrm{b}-1} p_j(v) \cdot \log\left(\frac{1}{\triangle_{n_\mathrm{b}}} \cdot p_j(v)\right) \\
&\overset{(9)}{=} -\sum_{j=0}^{n_\mathrm{b}-1} p_j(v) \cdot \log\left(p_j(v)\right) + \sum_{j=0}^{n_\mathrm{b}-1} p_j(v) \cdot \log\left(\triangle_{n_\mathrm{b}}\right) \\
&\overset{(8,9)}{=} -\sum_{j=0}^{n_\mathrm{b}-1} p_j(v) \cdot \log\left(p_j(v)\right) + \log\left(\triangle_{n_\mathrm{b}}\right) \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{n_\mathrm{b}-1} \mathbb{1}_{A_j}(\tilde{h}_{iv}) \\
&\overset{(10)}{=} -\sum_{j=0}^{n_\mathrm{b}-1} p_j(v) \cdot \log\left(p_j(v)\right) + \log\left(\triangle_{n_\mathrm{b}}\right).
\end{aligned}
$$

□

B Methodological Details

B.1 Sparse to Dense Pressure Estimation

Initial parameters In order to optimize the loss $\mathcal{L}_{\mathrm{S2D}}(\mathbf{d})$ w.r.t. the demand $\mathbf{d}$ via back-propagation, we need to initialize a starting point $\mathbf{d}_0$ and choose a reservoir and sensor head observation $\mathbf{h}_{V_\mathrm{r} \cup V_\mathrm{s}}$ for a fixed set of sensors V_s. However, as the sparse-to-dense pressure estimation task is ill-posed, different initializations $\mathbf{d}_0$ can lead to different solutions $\tilde{\mathbf{h}}$ still satisfying $\tilde{\mathbf{h}}_{V_\mathrm{r} \cup V_\mathrm{s}} = \mathbf{h}_{V_\mathrm{r} \cup V_\mathrm{s}}$ for given reservoir and sensor heads $\mathbf{h}_{V_\mathrm{r} \cup V_\mathrm{s}}$. At the same time, different heads $\mathbf{h}_{V_\mathrm{r} \cup V_\mathrm{s}}$ can also cause different solutions $\tilde{\mathbf{h}}$.

Therefore, in order to obtain statistically significant results, we will consider different statistics over multiple solutions $\tilde{\mathbf{h}}_{li}$, via Monte Carlo sampling. The multiple solutions are obtained by different initializations $\mathbf{d}_{li}$ based on multiple sensor readings $\mathbf{h}_{lV_\mathrm{r} \cup V_\mathrm{s}}$ for $l = 1, ..., n_0$ and a suitably chosen subset $\{\mathbf{d}_i \mid i \in I\}$ of the demands $\mathbf{d}_i$ for $i = 1, ..., n$ (with I_l a subset of $\{1, ..., n\}$).

Intuitively, we denoise realistic demands such that they are close to prior demands (step 1). Consequently, for each sensor configuration $l \in \{1, ..., n_0\}$, we subsample only those demands $\mathbf{d}_i$, whose corresponding model prediction $\tilde{\mathbf{h}}_i$ lies close to the sensor observations $\mathbf{h}_{lV_\mathrm{r} \cup V_\mathrm{s}}$ (step 2-5). Formally, we construct the multiple solutions as follows:

1. If the demands $(\mathbf{d}_i)_{i=1,...,n}$ are *realistic*, we Fourier-transform them node-wise, cut off frequencies higher than some $\xi \in \mathbb{R}$ and back-transform the

cut-off samples to denoise them. We denote the resulting prior demands by $\mathbf{d}_{li} := f_{\text{Four.}}^{-1}(\min\{f_{\text{Four.}}(\mathbf{d}_i), \xi\})$ for all $l = 1, ..., n_0$ and $i = 1, ..., n$. Note that up to here, we simply create n_0 duplicates of each denoised version of the demand $\mathbf{d}_i$.
If the demands $(\mathbf{d}_i)_{i=1,...,n}$ are *priors*, we directly create duplicates $\mathbf{d}_{li}$ for all $l = 1, ..., n_1$ and $i = 1, ..., n$ without denoising them.

2. Afterwards, for each $l \in \{1, ..., n_0\}$, we keep only those prior demands $\mathbf{d}_{li}$ whose corresponding heads $\tilde{\mathbf{h}}_{li}$ of the model's predictions $(\tilde{\mathbf{h}}_{li}, \tilde{\mathbf{q}}_{li}, \tilde{\mathbf{d}}_{li})$ are already close to the reservoir and sensor readings $\mathbf{h}_{lV_\text{r} \cup V_\text{s}}$, i.e., whose corresponding predictions $(\tilde{\mathbf{h}}_{li}, \tilde{\mathbf{q}}_{li}, \tilde{\mathbf{d}}_{li})$ are an element of the set

$$S_l := \Big\{(\tilde{\mathbf{h}}, \tilde{\mathbf{q}}, \tilde{\mathbf{d}}) \mid \sum_{v \in V_\text{r} \cup V_\text{s}} |\tilde{h}_v - h_{lv}|^2 < \alpha_l \Big\}.$$

3. For each $l \in \{1, ..., n_0\}$, we choose the hyperparameter $\alpha_l > 0$ such that the set of prior demands $\{\mathbf{d}_{li} \mid (\tilde{\mathbf{h}}_{li}, \tilde{\mathbf{q}}_{li}, \tilde{\mathbf{d}}_{li}) \in S_l\}$ is of some size $n_1 \in \mathbb{N}$.
4. For each $l \in \{1, ..., n_0\}$, we choose the set $I_l \subset \{1, ..., n\}$ as the subset of indices for which $\{\mathbf{d}_{li} \mid i \in I_l\} = \{\mathbf{d}_{li} \mid (\tilde{\mathbf{h}}_{li}, \tilde{\mathbf{q}}_{li}, \tilde{\mathbf{d}}_{li}) \in S_l\}$ holds. Note that by construction, $|I_l| = n_1$ is satisfied.
5. In total, we get $n_0 \cdot n_1$ demands $\mathbf{d}_{li}$ that serve as an initial parameter $\mathbf{d}_0$ and $n_0 \cdot n_1$ solutions $\tilde{\mathbf{h}}_{li}$ of the sparse-to-dense pressure estimation task for $l = 1, ..., n_0$ and $i \in I_l$.

Remark 4. While realistic demands $\mathbf{d}_i$ with noise might not be known in some scenarios, engineers usually have knowledge of prior demands without noise.

The denoising of the realistic demands (step 1) guarantees that the denoised demands $\mathbf{d}_{li}$ are rough estimations of prior demands. Consequently, step 2 and 3 assure that the initial prior demands $\mathbf{d}_{li}$ are somewhat close to the desired solution in order to decrease the training time.

B.2 Sensor Placement

Entropy Approximation We can leverage the solutions $\tilde{\mathbf{h}}_{li}$ from the sparse-to-dense pressure estimation task for $l = 1, ..., n_0$ and $i \in I_l$ (cf. subsection 5.2 and B.1) to evaluate different sensor placements.

More precisely, for a fixed observation of reservoir and sensor heads $\mathbf{h}_{lV_\text{r} \cup V_\text{s}}$ for $l \in \{1, ..., n_0\}$ and for a fixed node $v \in V$, similar to the mutual information in subsection 5.2, we want to leverage the distribution of the random variable H_{lv} determined by these solutions. Again, we want to approximate its corresponding density using elementary functions in order to compute approximated metrics from information theory. However, this time, instead of using the samples $\tilde{\mathbf{h}}_{iV_\text{c}} = (\tilde{h}_{iv})_{v \in V_\text{c}}$ for $i = 1, ..., n$ to do so, we use the samples $\tilde{\mathbf{h}}_{li} = (\tilde{h}_{liv})_{v \in V}$ for $i \in I_l$.

We assume that a better set of sensors V_s will lead to solutions $(\tilde{h}_{liv})_{i \in I_l}$ more similar to each other, or in other words, to a less surprising distribution of H_{lv}. In this case, we aim to minimize its entropy. We compute the entropy

similar to as how we approximate the mutual information in subsection 5.2. Up to a constant, the resulting discrete entropy is given by

$$\hat{E}(\mathrm{H}_{lv}) = -\sum_{j=0}^{n_\mathrm{b}} p_j(v) \cdot \log\Big(p_j(v)\Big)$$

with $p_j(v)$ as defined in theorem 1 (or theorem 5 in appendix A.3), but using the new heads $\tilde{\mathbf{h}}_{li}$ for $i \in I_l \subset \{1, ..., n\}$ instead of the heads $\tilde{\mathbf{h}}_i$ for $i = 1, ..., n$.

Finally, to summarize the entropy, we report the mean $\hat{E}_{\mathrm{SP1}}$ and $\hat{E}_{\mathrm{SP2}}$ over all reservoir and sensor observations $l = 1, ..., n_0$ as well as reservoir and sensor nodes $V_\mathrm{r} \cup V_\mathrm{s}$, and all nodes V, respectively:

$$\hat{E}_{\mathrm{SP1}} = \frac{1}{n_0 n_\mathrm{r}} \sum_{l=1}^{n_0} \sum_{v \in V_\mathrm{r}} \hat{E}(\mathrm{H}_{lv}), \quad \hat{E}_{\mathrm{SP2}} = \frac{1}{n_0 n_\mathrm{n}} \sum_{l=1}^{n_0} \sum_{v \in V} \hat{E}(\mathrm{H}_{lv}).$$

C Experimental Details

C.1 Water Distribution Systems

We conduct experiments on the same five WDS datasets as [2], as displayed in table 5.

WDS	Hanoi	Fossolo	Pescara	Area-C	Zhi Jiang
No. of junctions	32	37	71	93	114
No. of links	68	116	198	218	328
No. of reservoirs	1	1	3	1	1
Graph diameter	13	8	20	20	24
Node degree (min, mean, max)	(2, 4.25, 8)	(2, 6.27, 8)	(2, 5.52, 10)	(2, 4.69, 8)	(2, 5.75, 8)

Table 5: Attributes of WDSs used for experiments (cf. [2]).

C.2 Training of PI-GCNs

We solve all the water-related tasks presented in section 3 utilizing trained models from [2]. However, we train them in a slightly different way.

For training, we create 500 different scenarios, each one day long, where the sampling frequency is twice an hour. This gives us a total of 24,000 samples, which we divide into a train-val-test split using the ratio 60-20-20. More precisely, for each WDS, the scenarios are based on demands sampled from $\mathcal{N}(1, 0.15)$ with an offset sampled from $\mathcal{N}(0, 0.15)$ (*toy* demands). Moreover, we add variations to the diameters by adding noise sampled from $\mathcal{N}(0, 1/30)$ (cf. table 6). The rest of the training setup is similar to [2] (cf. table 7).

For evaluation, we generate realistic demands using the algorithm published by [22]. Since the original algorithm is tuned for Hanoi WDS, we have to optimize

	Hanoi	Fossolo	Pescara	Area-C	Zhi Jiang
			toy		
Demand pattern sampling distribution	$\mathcal{N}(1, 0.15)$	$\mathcal{N}(1, 0.15)$	$\mathcal{N}(1, 0.15)$	$\mathcal{N}(1, 0.15)$	$\mathcal{N}(1, 0.15)$
Demand offset sampling distribution	$\mathcal{N}(0, 0.15)$	$\mathcal{N}(0, 0.15)$	$\mathcal{N}(0, 0.15)$	$\mathcal{N}(0, 0.15)$	$\mathcal{N}(0, 0.15)$
Demand multiplier	1.0	3.0	0.5	5.0	0.1
Diameter noise sampling distribution	$\mathcal{N}(0, 1/30)$	$\mathcal{N}(0, 1/30)$	$\mathcal{N}(0, 1/30)$	$\mathcal{N}(0, 1/30)$	$\mathcal{N}(0, 1/30)$
Diameter multiplier	1.0	2.0	1.0	1.0	1.0
			realistic		
Demand multiplier	-	3.0	0.5	5.0	0.1
Demand pattern multiplier	-	0.5	0.5	0.5	0.5
Offsets sampling range	-	[0,0.125,0.025]	[0,0.125,0.025]	[0,0.125,0.025]	[0,0.125,0.025]

Table 6: Hyperparameters for dataset generation and evaluation of PI-GCN models.

Hyperparameters	Values
No. of GCN layers I	5
No. of MLP layers	2
Latent dimension	128
No. of training epochs	300
Learning Rate (LR)	0.0001
LR scheduler step size	30
LR scheduler decay rate	0.75
No. of training scenarios	500
No. of training samples	24,000
No. of training iterations K	[10, 15]
ρ and δ	0.1
No. of evaluation iterations	20
No. of evaluation scenarios	10
No. of evaluation samples	6,720

Table 7: Hyperparameters used for training PI-GCN models for different WDSs

the hyperparameters for other WDS (cf. table 6). Using this method, we generate 10 different demand patterns for each WDS, each 14 days long. This results in 6,720 samples.

For training and test results, we report the same metric (MRAE) between the model demand predictions and the ground truth demand (table 8) as well as between the model head and flow predictions and the corresponding EPANET predictions as [2] do, for both *toy* (test split) and *realistic* demands (table 9).

MRAE (%)	Hanoi	Fossolo	Pescara	Area-C	Zhi Jiang
	toy				
EPANET	0.0008 ± 0.0001	0.037 ± 0.004	0.083 ± 0.042	0.316 ± 0.054	0.007 ± 0.001
PI-GCN Model	0.059 ± 0.016	0.383 ± 0.089	1.467 ± 0.738	0.596 ± 0.140	0.568 ± 0.048
	realistic				
EPANET	1.793 ± 5.446	0.041 ± 0.008	0.091 ± 0.052	0.337 ± 0.072	0.007 ± 0.001
PI-GCN Model	0.320 ± 0.399	0.583 ± 0.337	1.285 ± 0.692	0.623 ± 0.190	1.851 ± 2.206

Table 8: MRAE on estimated vs true demands.

MRAE (%)	Hanoi	Fossolo	Pescara	Area-C	Zhi Jiang
	toy				
Flows	0.200 ± 0.656	5.149 ± 2.489	4.532 ± 7.284	0.252 ± 0.261	1.155 ± 0.594
Heads	0.003 ± 0.003	0.0007 ± 0.0002	0.044 ± 0.012	0.001 ± 0.001	0.196 ± 0.052
	realistic				
Flows	0.519 ± 3.085	5.357 ± 2.563	3.355 ± 4.518	0.199 ± 0.087	1.467 ± 1.560
Heads	0.003 ± 0.004	0.0005 ± 0.0003	0.036 ± 0.010	0.001 ± 0.001	0.282 ± 0.294

Table 9: MRAE for PI-GCN Model vs EPANET on flows and heads.

C.3 Sensor Placement

Table 10 shows the downstream task-dependent mean entropy $\hat{E}_{\text{SP1}}$ and $\hat{E}_{\text{SP2}}$ (cf. eq. (8)) per WDS and per hyperparameter λ.

C.4 Network Rehabilitation

For the task of network rehabilitation, table 11 shows the hyperparameters used for dataset generation, optimizing PI-GCN models using gradient method and the NSGA-II baselines.

Method	Hanoi	Fossolo	Pescara	Area-C	Zhi Jiang
	$\hat{E}_{\text{SP1}}$ (x 10^2)				
Random SP	14.02 ± 14.09	**3.68 ± 9.75**	8.14 ± 17.39	11.27 ± 20.84	5.15 ± 15.06
Clustering SP	15.24 ± 15.20	5.07 ± 14.10	9.61 ± 19.20	11.18 ± 20.91	4.55 ± 14.11
NSGA-II [9]	17.15 ± 13.55	11.02 ± 20.66	13.84 ± 21.59	12.71 ± 23.00	4.52 ± 14.72
Ours $\lambda = 0.0$	**7.39 ± 13.80**	**4.81 ± 14.47**	**5.43 ± 15.44**	**9.78 ± 20.69**	**3.53 ± 12.45**
Ours $\lambda = 0.125$	**7.39 ± 13.80**	6.14 ± 14.88	**5.43 ± 15.44**	**9.82 ± 19.41**	**3.53 ± 12.45**
Ours $\lambda = 0.25$	**7.39 ± 13.80**	6.14 ± 14.88	**5.43 ± 15.44**	**9.82 ± 19.41**	**3.47 ± 12.69**
Ours $\lambda = 0.5$	**13.65 ± 15.06**	11.02 ± 20.66	**7.48 ± 17.17**	11.33 ± 20.27	4.72 ± 14.85
Ours $\lambda = 0.75$	**13.65 ± 15.06**	11.02 ± 20.66	15.23 ± 22.54	11.69 ± 21.46	6.01 ± 16.28
Ours $\lambda = 1.0$	**13.65 ± 15.06**	12.37 ± 19.14	12.00 ± 19.22	12.57 ± 21.32	4.48 ± 14.48
	$\hat{E}_{\text{SP2}}$ (x 10^2)				
Random SP	**17.08 ± 16.35**	23.28 ± 31.38	23.21 ± 27.50	**11.30 ± 21.06**	5.36 ± 15.66
Clustering SP	**17.74 ± 17.07**	24.13 ± 30.80	**18.15 ± 24.40**	11.50 ± 21.28	4.97 ± 15.09
NSGA-II [9]	21.70 ± 17.98	**20.72 ± 25.91**	**16.99 ± 23.60**	12.77 ± 22.41	4.67 ± 15.06
Ours $\lambda = 0.0$	26.51 ± 18.88	29.66 ± 35.99	26.76 ± 25.27	**11.36 ± 21.24**	4.29 ± 14.00
Ours $\lambda = 0.125$	26.51 ± 18.88	29.60 ± 36.16	26.76 ± 25.27	11.90 ± 21.48	4.29 ± 14.00
Ours $\lambda = 0.25$	26.51 ± 18.88	29.60 ± 36.16	26.76 ± 25.27	11.90 ± 21.48	**3.93 ± 13.94**
Ours $\lambda = 0.5$	25.14 ± 19.52	**20.72 ± 25.91**	38.61 ± 33.15	11.70 ± 21.23	5.11 ± 15.29
Ours $\lambda = 0.75$	25.14 ± 19.52	**20.72 ± 25.91**	45.71 ± 37.35	12.16 ± 21.67	4.72 ± 14.34
Ours $\lambda = 1.0$	25.14 ± 19.52	**20.74 ± 25.42**	46.20 ± 37.36	12.26 ± 21.87	**3.66 ± 12.70**

Table 10: Mean entropy $\hat{E}_{\text{SP1}}$ and $\hat{E}_{\text{SP2}}$ (cf. eq. (8)) for the sensor placement task for different sensor configurations in dependence of different hyperparameter λ and on different WDSs. Lowest values are highlighted in **bold** and second lowest in **gray**.

	Hanoi	Fossolo	Pescara	Area-C	Zhi
WDS rehabilitation dataset					
Demand multiplier	1.5	6.0	0.85	8.5	0.11
Demand pattern multiplier	0.5	0.5	0.5	0.5	0.5
Offsets sampling range	[0,0.125,0.025]	[0,0.125,0.025]	[0,0.125,0.025]	[0,0.125,0.025]	[0,0.125,0.025]
WDS rehabilitation experiments					
PI-GCN					
α	10	100	100	100	10
β	0.1	0.1	1.0	0.01	1.0
γ	1.0	1.0	1.0	1.0	1.0
Minimum pressure (m)	40.0	52.5	20.0	115.0	30.0
NSGA-II					
Initial population size (m)	100	100	100	100	100
Number of generations (m)	100	100	100	100	100
Number of parents meeting	50	50	50	50	50
Random mutation min value	-0.01	-0.01	-0.01	-0.01	-0.01
Random mutation max value	0.10	0.10	0.15	0.10	0.10

Table 11: Hyperparameters for dataset generation and experiments for WDS rehabilitation task.