Supplementary Material:

Topologically Densified Distributions

In this supplementary material, we provide (1) all proofs (in §S1 to §S5) which were omitted in the main document, as well as (2) full architectural details, hyper-parameters and optimization settings (in §S6). Results which are restated from the main manuscript have the same numbering, while Definitions, Lemmas, etc., which are only present in the supplementary material have their labels suffixed by an "S". Additionally, restatements are given in purple.

S1. Generalization – Proof of Lemma 1

Recall that for $h : \mathcal{X} \to \mathcal{Y}$ and $X \sim P$, and a given labeling function $c : \mathrm{supp}(P) \to \mathcal{Y}$, we define the *generalization error* as

$$\mathop{\mathbb{E}}_{X\sim P}[\mathbb{1}_{h,c}(X)] \ .$$

where

$$\mathbb{1}_{h,c}(x) = \begin{cases} 0, & h(x) = c(x), \\ 1, & \text{else} \ . \end{cases}$$

We will now prove the following result of the main manuscript.

Lemma 1. *For any class $k \in [K]$, let $C_k = \varphi\left(c^{-1}(\{k\})\right)$ be its internal representation and $D_k = \gamma^{-1}(\{k\})$ be its decision region in $\mathcal{Z}$ w.r.t. γ. If, for $\varepsilon > 0$,*

$$\forall k : 1 - Q_k(D_k) \leq \varepsilon \ , \tag{3}$$

then

$$\mathop{\mathbb{E}}_{X\sim P}[\mathbb{1}_{\gamma\circ\varphi,c}(X)] \leq K\varepsilon \ .$$

To prove this lemma, we first introduce an auxiliary indicator function in order to deal with possible label overlaps the mapping φ can impose in $\mathcal{Z}$.

Definition S1. Let $h' : \mathcal{Z} \to \mathcal{Y}$ and $c^{\varphi}(z) = c\big(\varphi^{-1}(\{z\}) \cap \mathrm{supp}(P)\big) \subseteq \mathcal{Y}$. Then, we define

$$\mathbb{1}^{\varphi}_{h',c}(z) = \begin{cases} 0, & |c^{\varphi}(z)| = 1 \text{ and } h'(z) \in c^{\varphi}(z), \\ 1, & \text{else} \ . \end{cases}$$

Setting $h' = \gamma$, this auxiliary indicator function, $\mathbb{1}^{\varphi}_{\gamma,c}$, vanishes if and only if all $x \in \mathcal{X}$ which are mapped to an internal representation $z \in \mathcal{Z}$ *have the same label* $c(x)$. In other words, we *pessimistically* assume that internal representation where this is not the case are falsely classified. Thus, the auxiliary indicator function composed with φ, $\mathbb{1}^{\varphi}_{\gamma,c} \circ \varphi$, has to be greater or equal than the original $\mathbb{1}_{\gamma\circ\varphi,c}$. We formalize this insight next.

Lemma S1. *It holds that*

$$\mathbb{1}_{\gamma\circ\varphi,c} \leq \mathbb{1}^{\varphi}_{\gamma,c} \circ \varphi$$

Proof. Let $x \in \mathcal{X}$. It is sufficient to show that

$$\mathbb{1}^{\varphi}_{\gamma,c} \circ \varphi(x) = 0 \Rightarrow \mathbb{1}_{\gamma\circ\varphi,c}(x) = 0 \ .$$

Thus, let $\mathbb{1}^{\varphi}_{\gamma,c} \circ \varphi(x) = 0$. Then, by definition

(i) $|c\big(\varphi^{-1}\big(\{\varphi(x)\}\big)\big)| = 1$ and,

(ii) $\gamma(\varphi(x)) \in c\big(\varphi^{-1}\big(\{\varphi(x)\}\big)\big)$.

By (i) there is some $y \in \mathcal{Y}$ such that $c\big(\varphi^{-1}\big(\{\varphi(x)\}\big)\big) = \{y\}$ and thus $c(x) = y$. With this, and (ii), we get $\gamma \circ \varphi(x) = y$ and thus $\gamma \circ \varphi(x) = c(x)$. Therefore, $\mathbb{1}_{\gamma\circ\varphi,c}(x) = 0$ which concludes the proof. □

We now have all necessary tools to prove a slightly more general version of Lemma 1 from the main manuscript.

Lemma S2. *Let for any class $k \in \{1, \ldots, K\}$, $C_k = \varphi\left(c^{-1}\big(\{k\}\big)\right)$ be its internal representation and $D_k = \gamma^{-1}\big(\{k\}\big)$ its decision region in $\mathcal{Z}$. If*

$$\forall k : 1 - Q_k(D_k) \leq \varepsilon_k \ ,$$

then

$$\mathop{\mathbb{E}}_{X\sim P}[\mathbb{1}_{\gamma\circ\varphi,c}(X)] \leq \sum_{k=1}^{K} \varepsilon_k \ .$$

Proof. For brevity, let $\widehat{C}_k = c^{-1}\big(\{k\}\big)$ and write $\mathbb{1}_{\gamma\circ\varphi}$ instead of $\mathbb{1}_{\gamma\circ\varphi,c}$. Then, we get

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim P}[\mathbb{1}_{\gamma\circ\varphi}(x)] &= \int_{\mathcal{X}} \mathbb{1}_{\gamma\circ\varphi}(x) dP(x) \\
&= \int_{\mathrm{supp}(P)} \mathbb{1}_{\gamma\circ\varphi}(x) dP(x) \\
&\leq \int_{\mathrm{supp}(P)} \mathbb{1}^{\varphi}_{\gamma} \circ\varphi(x) dP(x) && \text{(by Lemma S1)} \\
&= \int_{\varphi(\mathrm{supp}(P))} \mathbb{1}^{\varphi}_{\gamma}(z) dQ(z) && \text{(change of variables)} \\
&= \sum_{k=1}^{K} \int_{\underbrace{\varphi\left(\mathrm{supp}(P)\right) \cap D_k}_{D_k^{\cap}}} \mathbb{1}^{\varphi}_{\gamma}(z) dQ(z) \ . && \text{(as } D_1, \ldots, D_K \text{ partition } \mathcal{Z})
\end{aligned}$$

For $1 \leq k \leq K$, let $D_k^{\cap} = \varphi\left(\mathrm{supp}(P)\right) \cap D_k$ and consider each summation term separately. First, we can re-write $D_k^{\cap}$ as

$$D_k^{\cap} = \left(D_k^{\cap} \setminus \bigcup_{\substack{i=1 \\ i\neq k}}^{K} C_i \right) \cup \left(D_k^{\cap} \cap \bigcup_{\substack{i=1 \\ i\neq k}}^{K} C_i \right) \ .$$

Second, consider

$$z \in D_k^{\cap} \setminus \bigcup_{\substack{i=1 \\ i\neq k}}^{K} C_i \ .$$

Then, $z \in D_k$ and thus $\gamma(z) = k$ (by definition of D_k). Further, we have $c\big(\varphi^{-1}\big(\{z\}\big) \cap \mathrm{supp}(P)\big) = \{k\}$ and thus, by Definition S1,

$$\mathbb{1}^{\varphi}_{\gamma}(z) = 0 \ . \tag{S1}$$

With this, considering each summation term yields

$$\begin{aligned}
\int_{D_k^{\cap}} \mathbb{1}_\gamma^\varphi(z)dQ(z) &= \int_{D_k^{\cap}\setminus \bigcup_{\substack{i=1\\ i\neq k}}^{K} C_i} \mathbb{1}_\gamma^\varphi(z)dQ(z) + \int_{D_k^{\cap}\cap \bigcup_{\substack{i=1\\ i\neq k}}^{K} C_i} \mathbb{1}_\gamma^\varphi(z)dQ(z) \\
&= \int_{D_k^{\cap}\cap \bigcup_{\substack{i=1\\ i\neq k}}^{K} C_i} \mathbb{1}_\gamma^\varphi(z)dQ(z) && \text{(by Eq. (S1))} \\
&= \int_{\bigcup_{\substack{i=1\\ i\neq k}}^{K} (C_i\cap D_k^{\cap})} \mathbb{1}_\gamma^\varphi(z)dQ(z) \\
&\leq Q\left(\bigcup_{\substack{i=1\\ i\neq k}}^{K}(C_i\cap D_k^{\cap})\right) && \text{(as } \mathbb{1}_\gamma^\varphi \leq 1\text{)} \\
&\leq Q\left(\bigcup_{\substack{i=1\\ i\neq k}}^{K}(C_i\cap D_k)\right) && \text{(as } C_i \subset \varphi(\mathrm{supp}(P))\text{)} \\
&\leq \sum_{\substack{i=1\\ i\neq k}}^{K} Q\left(C_i\cap D_k\right)
\end{aligned}$$

In order to obtain the final result, we use the fact that the decision regions D_k are disjoint and cover the representation space, i.e. $\mathcal{Z} = \bigsqcup_{k=1}^{K} D_k$. Thus for any $1 \leq i \leq K$,

$$Q(C_i) = \sum_{k=1}^{K} Q\left(C_i \cap D_k\right) \quad . \tag{S2}$$

Consequently, changing the summation order allows to simplify the bound from above to

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim P}[\mathbb{1}_{\gamma\circ\varphi}(x)] &\leq \sum_{k=1}^{K}\sum_{\substack{i=1\\ i\neq k}}^{K} Q\left(C_i\cap D_k\right) = \sum_{i=1}^{K}\sum_{\substack{k=1\\ k\neq i}}^{K} Q\left(C_i\cap D_k\right) \\
&= \sum_{i=1}^{K}\left(Q(C_i) - Q(C_i\cap D_i)\right) && \text{(by Eq. (S2))} \\
&= \sum_{i=1}^{K} Q(C_i)\left(1 - \frac{Q(C_i\cap D_i)}{Q(C_i)}\right) \\
&= \sum_{i=1}^{K} Q(C_i)\left(1 - Q_i(D_i)\right) && (*) \\
&\leq \sum_{i=1}^{K} Q(C_i)\varepsilon_i && (**) \\
&\leq \sum_{i=1}^{K} \varepsilon_i \quad ,
\end{aligned}$$

where $(*)$ follows from the definition of the class-specific probability mass in Eq. (2) and $(**)$ holds by the assumption of the lemma. □

Proof of Lemma 1. By setting $\varepsilon_k = \varepsilon > 0$ in Lemma S2 we get the desired result. □

S2. A sufficient condition on *not*-β-connectivity

Definition S2. Let $(\mathcal{Z}, \mathbb{d})$ be a metric space and $\emptyset \neq A, B \subset \mathcal{Z}$. We define the *set margin* between A and B as

$$\mathbb{m}(A, B) = \inf_{a \in A, b \in B} \mathbb{d}(a, b) \ .$$

The following lemma formalizes the intuition that $z_1, \ldots, z_b$ cannot be β-connected if the z_i are distributed among two sets which are separated by a sufficiently large set margin.

Lemma S3. *Let $(\mathcal{Z}, \mathbb{d})$ be a metric space and $\mathbf{z} = (z_1, \ldots, z_b) \in \mathcal{Z}^b$. Define, for $l \in \mathbb{N}$ and $A, B \subset \mathcal{Z}$, the index sets*

$$\begin{aligned} I_{A,\mathbf{z}} &= \{i \in [b] : z_i \in A\}, \\ I_{B,\mathbf{z}} &= \{i \in [b] : z_i \in B\}, \text{and} \\ I_{C,\mathbf{z}} &= \{i \in [b] : z_i \in (A \cup B)^{\complement}\} \ , \end{aligned}$$

where $[b] = \{1, \ldots, b\}$ and $(A \cup B)^{\complement}$ denotes the set complement of $(A \cup B)$.

If

$$\mathbb{m}(A, B) \geq l \cdot \beta \text{ and } |I_{A,\mathbf{z}}|, \ |I_{B,\mathbf{z}}| \geq 1 \text{ and } |I_{C,\mathbf{z}}| \leq l - 1$$

then

$$\mathfrak{c}_b^{\beta}(\mathbf{z}) = 0 \ ,$$

i.e., $\mathbf{z}$ is not β-connected.

Proof. We prove this by way of contradiction. For brevity, let $\mathfrak{c}_b^{\beta} = \mathfrak{c}$. Consider $\mathbf{z} = (z_1, \ldots, z_b)$ as above and assume $\mathfrak{c}(\mathbf{z}) = 1$. Let, w.l.o.g., $z_1 \in A$ and $z_b \in B$. Then, there is a path of distinct nodes

$$z_1 \leftrightarrow \cdots \leftrightarrow z_b$$

connecting z_1 and z_b with line segments of length $< \beta$. However, by assumption $\mathbb{m}(A, B) \geq \beta$, and thus there is a sub-path

$$z_{i_1} \leftrightarrow z_{i_2} \leftrightarrow \cdots \leftrightarrow z_{i_m}$$

such that $z_{i_1} \in A$, $z_{i_m} \in B$ and $z_{i_2}, \ldots, z_{i_{m-1}} \in (A \cup B)^{\complement}$. Thus, we get

$$\mathbb{d}(z_{i_1}, z_{i_m}) \leq \mathbb{d}(z_{i_1}, z_{i_2}) + \cdots + \mathbb{d}(z_{i_{m-1}}, z_{i_m}) < (m-1) \cdot \beta \ .$$

By construction, $z_{i_2}, \ldots, z_{i_{m-1}} \in C$ and thus $\{i_2, \ldots, i_{m-1}\} \subseteq I_{C,\mathbf{z}}$. Hence, $m - 2 \leq |I_{C,\mathbf{z}}| \leq l - 1$. Therefore, $(m-1) \cdot \beta \leq l \cdot \beta$, leading to

$$\mathbb{d}(z_{i_1}, z_{i_m}) < l \cdot \beta \ .$$

This directly contradicts $\mathbb{m}(A, B) \geq l \cdot \beta$. □

S3. Concentration results

Lemma S4. *Let $(\mathcal{Z}, \mathbb{d})$ be a metric space and Σ the corresponding Borel σ-algebra. Further, let $b \in \mathbb{N}$, $\beta > 0$ and Q be a (b, c_β)-connected probability measure (cf. Definition 2) on $(\mathcal{Z}, \Sigma)$. For $l \in \mathbb{N}$ and $A, B \in \Sigma$, such that $\mathbb{m}(A, B) \geq l \cdot \beta$, the following inequality holds*

$$1 - c_\beta \geq \sum_{(n_1, n_2, n_3) \in I(b,l)} \frac{b!}{n_1! n_2! n_3!} \cdot Q(A)^{n_1} Q(B)^{n_2} (1 - Q(A) - Q(B))^{n_3} \ ,$$

where the index set $I(b, l)$ is given as

$$I(b, l) = \{ \ (n_1, n_2, n_3) \in \{0, \ldots, b\}^3 : n_1 + n_2 + n_3 = b, \ 1 \leq n_1, \ 1 \leq n_2, \ n_3 \leq l - 1 \ \} \ . \quad \text{(S3)}$$

Proof. Let $C = (A \cup B)^{\complement}$. The proof is structured in *three parts*: First, we construct an auxiliary random variable that captures the scattering of b-sized samples across A, B and C. This allows us to express probabilities for different scattering configurations. Second, we describe scattering configurations where β-connectivity cannot be satisfied. Finally, by combining both previous parts, we derive the claimed inequality.

Part I. First, define the categorical function

$$f : \mathcal{Z} \to \{1, 2, 3\}, \quad f(z) = \begin{cases} 1, & z \in A, \\ 2, & z \in B, \\ 3, & z \in C = (A \cup B)^{\complement} \end{cases} .$$

For a random variable $Z \sim Q$, we now consider the random variable $f \circ Z$ with

$$\begin{aligned} Q(\{f \circ Z = 1\}) &= Q(A) \\ Q(\{f \circ Z = 2\}) &= Q(B) \\ Q(\{f \circ Z = 3\}) &= Q(C) \ . \end{aligned}$$

By drawing b-times i.i.d. from $f \circ Z$ and counting the occurrences of $1, 2, 3$, we get a multinomially distributed random variable, K. This means that, for

$$\{K = (n_1, n_2, n_3)\}$$

where $n_1 + n_2 + n_3 = b$, it holds that

$$Q^b\left(\{K = (n_1, n_2, n_3)\}\right) = \frac{b!}{n_1! n_2! n_3!} \cdot Q(A)^{n_1} Q(B)^{n_2} (1 - Q(A) - Q(B))^{n_3} \ .$$

Part II. Similar to Lemma S3, we define

$$\begin{aligned} I_{A,\mathbf{z}} &= \{i \in I : z_i \in A\}, \\ I_{B,\mathbf{z}} &= \{i \in I : z_i \in B\}, \text{and} \\ I_{C,\mathbf{z}} &= \{i \in I : z_i \in (A \cup B)^{\complement}\} \ . \end{aligned}$$

Then, by construction, it holds that

$$\left\{K = (n_1, n_2, n_3)\right\} = \left\{\mathbf{z} \in \mathcal{Z}^b : \left(|I_{A,\mathbf{z}}|, |I_{B,\mathbf{z}}|, |I_{C,\mathbf{z}}|\right) = (n_1, n_2, n_3)\right\} \ .$$

Now let $(n_1, n_2, n_3) \in I(b, l)$ and consider $\mathbf{z} \in \{K = (n_1, n_2, n_3)\}$. By definition of $I(b, l)$, we get

$$1 \leq |I_{A,\mathbf{z}}|, \quad 1 \leq |I_{B,\mathbf{z}}|, \quad \text{and} \quad |I_{C,\mathbf{z}}| \leq l - 1 \ .$$

Remember that, by assumption, $\mathbb{m}(A, B) \geq l \cdot \beta$ and thus, by Lemma S3, $\mathbf{z} \in \{\mathfrak{c} = 0\}$, i.e., the points $z_1, \ldots, z_b$ are not β-connected. This yields the following implication:

$$(n_1, n_2, n_3) \in I(b, l) \Rightarrow \{K = (n_1, n_2, n_3)\} \subseteq \{\mathfrak{c} = 0\} \ .$$

Part III. Combining the results of Part I and II, we obtain the following inequality:

$$\begin{aligned} 1 - c_\beta &= Q^b(\{\mathfrak{c} = 0\}) \\ &\geq Q^b\left(\bigcup_{\substack{(n_1,n_2,n_3) \\ \in I(b,l)}} \{K = (n_1, n_2, n_3)\}\right) \\ &= \sum_{\substack{(n_1,n_2,n_3) \\ \in I(b,l)}} Q^b\left(\{K = (n_1, n_2, n_3)\}\right) \\ &= \sum_{\substack{(n_1,n_2,n_3) \\ \in I(b,l)}} \frac{b!}{n_1! n_2! n_3!} \cdot Q(A)^{n_1} Q(B)^{n_2} (1 - Q(A) - Q(B))^{n_3} \ . \end{aligned}$$

□

With Lemma S4 in mind, we can restate the definition of the polynomial Ψ from the main manuscript.

Definition 3. Let $b, l \in \mathbb{N}$ and $p, q \in [0,1]$. For $p \leq q$, we define the polynomial

$$\Psi(p, q\,;\, b, l) = \sum_{\substack{(u,v,w) \\ \in I(b,l)}} \frac{b!}{u!\; v!\; w!} p^u (1-q)^v (q-p)^w \quad ,$$

where the index set $I(b,l)$ is given by

$$I(b,l) = \{\, (u,v,w) \in \mathbb{N}_0^3 : \\ u+v+w = b \;\wedge\; u,v \geq 1 \;\wedge\; w \leq l-1 \,\} \quad .$$

While all previous results in this supplementary material are stated for a probability measure Q on $\mathcal{Z}$, the results equally transfer to Q_k, i.e., the restriction of probability measure Q to a particular class k, which is the specific setting considered in the main part of the manuscript.

Theorem 1. *Let $b, l \in \mathbb{N}$ and let Q_k be (b, c_β)-connected. Then, for all reference sets $M \in \Sigma$ and*

$$\mathfrak{p} = Q_k(M), \;\; \mathfrak{q} = Q_k(M_{l\cdot\beta})$$

it holds that

$$1 - c_\beta \geq \Psi\big(\mathfrak{p}, \mathfrak{q}\,;\, b, l\big) \quad . \tag{5}$$

Proof. The proof relies on Lemma S4 with

$$A = M \text{ and } B = (M_{l\cdot\beta})^{\complement} \quad .$$

For Q_k as Q, we get

$$\begin{aligned}
1 - c_\beta &\geq \sum_{\substack{(n_1,n_2,n_3) \\ \in I(b,l)}} \frac{b!}{n_1! n_2! n_3!} \cdot Q(A)^{n_1} Q(B)^{n_2} (1 - Q(A) - Q(B))^{n_3} \\
&= \sum_{\substack{(n_1,n_2,n_3) \\ \in I(b,l)}} \frac{b!}{n_1! n_2! n_3!} \cdot Q(M)^{n_1} \big(1 - Q(M_{l\cdot\beta})\big)^{n_2} \big(1 - Q(M) - (1 - Q(M_{l\cdot\beta}))\big)^{n_3} \\
&= \sum_{\substack{(n_1,n_2,n_3) \\ \in I(b,l)}} \frac{b!}{n_1! n_2! n_3!} \cdot Q(M)^{n_1} \big(1 - Q(M_{l\cdot\beta})\big)^{n_2} \big(Q(M_{l\cdot\beta}) - Q(M)\big)^{n_3} \\
&= \sum_{\substack{(u,v,w) \\ \in I(b,l)}} \frac{b!}{u!v!w!} \cdot p^u (1-q)^v (q-p)^w \quad ,
\end{aligned}$$

where, in the last equality, we have set $p = Q(M)$, $q = Q(M_{l\cdot\beta})$ and renamed the indices. □

S4. Properties of Ψ

In the main manuscript, we list three important properties of Ψ. These are:

(1) Ψ is *monotonically increasing* in p,
(2) Ψ is *monotonically decreasing* in q, and
(3) Ψ is *monotonically increasing* in l and Ψ vanishes for $q = 1$.

While the latter trivially follows from Definition 3, as $(1-q) = (1-1) = 0$ is always present with non-zero exponent, properties (1) and (2) need more careful (tedious) consideration. The monotonicity in l results from the fact that increasing l increases the size of $I(b,l)$ and thus more non-negative terms are present in the summation in Ψ (see Definition 3) . We start by providing two beneficial ways of re-writing the index set, $I(\cdot,\cdot)$, which is used to define Ψ.

Lemma S5. *Let $b, l \in \mathbb{N}$ and define*

$$g(x) = \max\{1, b - x - l + 1\} \ .$$

This yields the following re-write of the index set as follows:

$$\begin{aligned} I(b,l) &= \left\{(n_1,n_2,n_3) \in \{0,\dots,b\}^3 : n_1+n_2+n_3 = b, 1 \le n_1, 1 \le n_2, n_3 \le l-1\right\} \\ &= \\ &\bigcup_{n_1=1}^{b-1} \{(n_1, n_2, b-n_1-n_2) : g(n_1) \le n_2 \le b - n_1\} \\ &= \\ &\bigcup_{n_2=1}^{b-1} \{(n_1, n_2, b-n_1-n_2) : g(n_2) \le n_1 \le b - n_2\} \end{aligned} \tag{S4}$$

Proof. We only show the first equality as the second is analogous with switched roles of n_1 and n_2.

Part I ($\subseteq$): Let $(k,i,j) \in I(b,l)$, i.e., $(k,i,j) \in \{0,\dots,b\}^3$ such that

(1) $k + i + j = b$,
(2) $1 \le k$,
(3) $1 \le i$, and
(4) $j \le l - 1$.

From (1) it follows that $j = b - k - i$.
From (4) we get

$$b - k - i \overset{1)}{=} j \overset{(4)}{\le} l - 1 \Leftrightarrow b - k - l + 1 \le i \ .$$

Combining this with (3), we conclude

$$g(k) = \max\{1, b-k-l+1\} \le i \ .$$

From (1), we see that $i \le b - k$, as $i + j = b - k$ and $j > 0$. This means

$$(k,i,j) \in \{(k, n_2, b-k-n_2) : g(k) \le n_2 \le b-k\} \ . \tag{S5}$$

Finally, (1), (2), and (3) yield $1 \le k \le b-1$ and therefore

$$(k,i,j) \in \bigcup_{n_1=1}^{b-1} \{(n_1, n_2, b-n_1-n_2) : g(n_1) \le n_2 \le b-k\} \tag{S6}$$

which concludes the "$\subseteq$" part.

Part II ($\supseteq$): Let $1 \leq n_1 \leq b-1$ and consider

$$(k,i,j) \in \{(n_1, n_2, b-n_1-n_2) : g(n_1) \leq n_2 \leq b-k\} \ . \tag{S7}$$

Then, $k+i+j=b$ and $1 \leq k, 1 \leq i$.
For the last condition, i.e., $j \leq l-1$, consider

$$j = b-k-i \leq b-k-g(k) \ .$$

We next distinguish the two possible outcomes of $g(k)$:

Case 1: $g(k) = b-k-l+1$: Then, we get

$$\begin{aligned} j &\leq b-k-g(k) \\ &= b-k-(b-k-l+1) \\ &= l-1 \ . \end{aligned}$$

Case 2: $g(k)=1$: Then, by definition of $g(k)$, we get

$$\begin{aligned} b-k-l+1 &\leq 1 \\ \Leftrightarrow b-k &\leq l \\ \Leftrightarrow b-k-1 &\leq l-1 \end{aligned}$$

and therefore

$$\begin{aligned} j &\leq b-k-g(k) \\ &= b-k-1 \\ &\leq l-1 \ . \end{aligned}$$

□

We now use the results of Lemma S5 to re-arrange the sum in the definition of Ψ.

Corollary S1. *For $g(x) = \max\{1, b-x-l+1\}$, it holds that*

$$\Psi(p,q\,;\,b,l) = \sum_{n_1=1}^{b-1} \sum_{n_2=g(n_1)}^{b-n_1} \frac{b!}{n_1!n_2!(b-n_1-n_2)!} p^{n_1}(1-q)^{n_2}(q-p)^{b-n_1-n_2} \tag{S8}$$

and

$$\Psi(p,q\,;\,b,l) = \sum_{n_2=1}^{b-1} \sum_{n_1=g(n_2)}^{b-n_2} \frac{b!}{n_1!n_2!(b-n_1-n_2)!} p^{n_1}(1-q)^{n_2}(q-p)^{b-n_1-n_2} \ . \tag{S9}$$

In the following lemma we will prove the claimed monotonicity properties of Ψ by (i) using the previously derived rearrangements of the summation and (ii) considering the corresponding derivatives.

Lemma S6. *Let $b, l \in \mathbb{N}$ and $p_0, q_0 \in (0,1)$ arbitrary but fixed. Then, it holds that*

(1) $\Psi(\cdot, q_0\,;\,b,l)$ is monotonically increasing on $[0, q_0]$

and

(2) $\Psi(p_0, \cdot\,;\,b,l)$ is monotonically decreasing on $[p_0, 1]$.

Proof. Ad (1). For brevity, we write q instead of q_0. First, we leverage Corollary S1, Eq. (S9), and re-arrange the sum:

$$\begin{aligned}\Psi(p,q\,;\,b,l) &= \sum_{n_2=1}^{b-1}\sum_{n_1=g(n_2)}^{b-n_2}\frac{b!}{n_1!n_2!(b-n_1-n_2)!}p^{n_1}(1-q)^{n_2}(q-p)^{b-n_1-n_2}\\ &= \sum_{n_2=1}^{b-1}(1-q)^{n_2}\underbrace{\sum_{n_1=g(n_2)}^{b-n_2}\frac{b!}{n_1!n_2!(b-n_1-n_2)!}p^{n_1}(q-p)^{b-n_1-n_2}}_{=A_{n_2}(p)}\quad.\end{aligned}$$

For studying the monotonicity properties of $\Psi(\cdot,q\,;\,b,l)$, it is sufficient to consider A_{n_2}, for $1\le n_2\le b-1$.

We define two auxiliary functions

$$\begin{aligned}a_{n_1}(p) &= \frac{b!}{n_1!n_2!(b-n_1-n_2)!}p^{n_1}(q-p)^{b-n_1-n_2},\\ c_{n_1}(p) &= \frac{b!}{n_1!n_2!(b-n_1-n_2)!}p^{n_1}(b-n_1-n_2)(q-p)^{b-n_1-n_2-1}\quad.\end{aligned}$$

Note that $A_{n_2}(p)=\sum_{n_1=g(n_2)}^{b-n_2}a_{n_1}(p)$ and that

$$c_{n_1}(p)=\begin{cases}\frac{b!}{n_1!n_2!(b-n_1-n_2-1)!}p^{n_1}(q-p)^{b-n_1-n_2-1}\ge 0, & 0\le n_1<b-n_2\\ 0, & n_1=b-n_2\end{cases},$$

because by assumption $0\le p\le q\le 1$. As we will show below,

$$\frac{\partial a_{n_1}(p)}{\partial p}=c_{n_1-1}(p)-c_{n_1}(p).$$

Hence,

$$\frac{\partial A_{n_2}(p)}{\partial p}=\sum_{n_1=g(n_2)}^{b-n_2}(c_{n_1-1}(p)-c_{n_1}(p))=c_{g(n_2)-1}(p)-c_{b-n_2}(p)=c_{g(n_2)-1}(p)\ge 0.$$

Consequently, A_{n_2} is monotonically increasing and thus, so is $\Psi(\cdot,q\,;\,b,l)$.

It remains to calculate the derivative $\frac{\partial a_{n_1}(p)}{\partial p}$:

$$\begin{aligned}\frac{\partial a_{n_1}(p)}{\partial p} &= \frac{b!}{n_1!n_2!(b-n_1-n_2)!}\cdot\left[n_1p^{n_1-1}(q-p)^{b-n_1-n_2}-p^{n_1}(b-n_1-n_2)(q-p)^{b-n_1-n_2-1}\right]\\ &= \frac{b!}{n_1!n_2!(b-n_1-n_2)!}\cdot n_1p^{n_1-1}(q-p)^{b-n_1-n_2}\\ &\quad-\frac{b!}{n_1!n_2!(b-n_1-n_2)!}\cdot p^{n_1}(b-n_1-n_2)(q-p)^{b-n_1-n_2-1}\\ &= \frac{b!}{(n_1-1)!n_2!(b-n_1-n_2)!}\cdot p^{n_1-1}(q-p)^{b-n_1-n_2}\\ &\quad-\frac{b!}{n_1!n_2!(b-n_1-n_2)!}\cdot p^{n_1}(b-n_1-n_2)(q-p)^{b-n_1-n_2-1}\\ &= c_{n_1-1}(p)-c_{n_1}(p)\end{aligned}$$

Ad 2). The proof is rather similar to the first part. Nevertheless, we will exercise it for completeness. For brevity, we will write p instead of p_0. Again we start by leveraging Corollary S1, but, this time,

use Eq. (S8), and re-arrange the sum,

$$\begin{aligned}\Psi(p,q\,;\,b,l) &= \sum_{n_1=1}^{b-1}\sum_{n_2=g(n_1)}^{b-n_1}\frac{b}{n_1!n_2!(b-n_1-n_2)!}p^{n_1}(1-q)^{n_2}(q-p)^{b-n_1-n_2}\\ &= \sum_{n_1=1}^{b-1}p^{n_1}\underbrace{\sum_{n_2=g(n_1)}^{b-n_1}\frac{b}{n_1!n_2!(b-n_1-n_2)!}(1-q)^{n_2}(q-p)^{b-n_1-n_2}}_{A_{n_1}(q)}\end{aligned}$$

For studying the monotonicity properties of $\Psi(p,\cdot\,;\,b,l)$ it is sufficient to consider A_{n_1}, for $1\le n_1\le b-1$.

Again, we define two auxiliary functions

$$\begin{aligned}a_{n_2}(q) &= \frac{b!}{n_1!n_2!(b-n_1-n_2)!}(1-q)^{n_2}(q-p)^{b-n_1-n_2},\\ c_{n_2}(q) &= \frac{b!}{n_1!n_2!(b-n_1-n_2)!}(1-q)^{n_2}(b-n_1-n_2)(q-p)^{b-n_1-n_2-1}.\end{aligned}$$

Note that $A_{n_1}(q)=\sum_{n_2=g(n_1)}^{b-n_1}a_{n_2}(q)$ and that

$$c_{n_1}(q)=\begin{cases}\frac{b!}{n_1!n_2!(b-n_1-n_2-1)!}(1-q)^{n_2}(q-p)^{b-n_1-n_2-1}\ge 0, & 0\le n_2<b-n_1\\ 0, & n_2=b-n_1\end{cases},$$

because by assumption $0\le p\le q\le 1$.

As we will show below,

$$\frac{\partial a_{n_2}(q)}{\partial q}=c_{n_2}(q)-c_{n_2-1}(q).$$

Hence,

$$\frac{\partial A_{n_1}(q)}{\partial q}=\sum_{n_2=g(n_1)}^{b-n_1}(c_{n_2}(q)-c_{n_2-1}(q))=c_{b-n_1}(q)-c_{g(n_1)-1}(q)=-c_{g(n_1)-1}(q)\le 0.$$

Consequently, A_{n_1} is monotonically decreasing and thus, so is $\Psi(p,\cdot,\,;\,b,l)$.

It remains to calculate the derivative $\frac{\partial a_{n_2}(q)}{\partial q}$:

$$\begin{aligned}\frac{\partial a_{n_2}}{\partial q}(q) &= -\frac{b!}{n_1!n_2!(b-n_1-n_2)!}n_2(1-q)^{n_2-1}(q-p)^{b-n_1-n_2}\\ &\quad+\frac{b!}{n_1!n_2!(b-n_1-n_2)!}(1-q)^{n_2}(b-n_1-n_2)(q-p)^{b-n_1-n_2-1}\\ &= -\frac{b!}{n_1!(n_2-1)!(b-n_1-n_2)!}(1-q)^{n_2-1}(q-p)^{b-n_1-n_2}\\ &\quad+\frac{b!}{n_1!n_2!(b-n_1-n_2)!}(1-q)^{n_2}(b-n_1-n_2)(q-p)^{b-n_1-n_2-1}\\ &= -c_{n_2-1}+c_{n_2}\end{aligned}$$

□

S5. Monotonicity properties of $\mathcal{R}$

Definition S3.

$$\mathcal{R}_{b,c_\beta}(p,l)=\min_q\Big\{q\in[p,1]:1-c_\beta\ge\Psi(p,q\,;\,b,l)\Big\}$$

Lemma S7. *$\mathcal{R}_{b,c_\beta}(p,l)$ is*

(1) monotonically increasing in p, and
(2) monotonically increasing in l .

Proof. Ad (1). Let $p, \hat{p} \in [0,1]$ with $p < \hat{p}$.
First, assume $\mathcal{R}_{b,c_\beta}(p,l) \in [p,\hat{p}]$. Then, by definition, $\mathcal{R}_{b,c_\beta}(\hat{p},l) \in [\hat{p},1]$ and thus

$$\mathcal{R}_{b,c_\beta}(p,l) \leq \mathcal{R}_{b,c_\beta}(\hat{p},l) \ .$$

Second, assume $\mathcal{R}_{b,c_\beta}(p,l) \notin [p,\hat{p}]$, i.e., $\mathcal{R}_{b,c_\beta}(p,l) \in [\hat{p},1]$. Let

$$A = \{q \in [\hat{p},1] : 1 - c_\beta \geq \Psi(p,q\,;\,b,l)\}$$

and

$$B = \{q \in [\hat{p},1] : 1 - c_\beta \geq \Psi(\hat{p},q\,;\,b,l)\} \ .$$

By Lemma S6, Ψ is monotonically increasing in p and thus

$$\Psi(\hat{p},q\,;\,b,l) \geq \Psi(p,q\,;\,b,l) \text{ for } q \in [\hat{p},1] \ .$$

This implies $B \subseteq A$ and therefore

$$\mathcal{R}_{b,c_\beta}(p,l) = \min A \leq \min B = \mathcal{R}_{b,c_\beta}(\hat{p},l) \ .$$

Ad 2). Let $l < \hat{l}$. It follows from the definition of I, see Definition 3, that $I(b,l) \subseteq I(b,\hat{l})$. As all addends in the sum defining Ψ are positive, the claim follows. □

S6. Experimental details

For reproducibility, we provide full architectural details, optimization settings and hyper-parameters.

S6.1. Architecture

On SVHN and CIFAR10, we use the CNN-13 architecture of (Laine & Aila, 2017, Table 5), without the Gaussian noise input layer. The configuration is provided in Table S1 below (essentially reproduced from the original paper). `BN` denotes 2D batch normalization (Ioffe & Szegedy, 2015), `LReLU` denotes leaky ReLU activation with $\alpha = 0.1$.

		BN	LReLU
Input	32×32 RGB image		
`Conv (2D)`	`Filters: 128; Kernel: 3x3; Pad: 1`	✓	✓
`Conv (2D)`	`Filters: 128; Kernel: 3x3; Pad: 1`	✓	✓
`Conv (2D)`	`Filters: 128; Kernel: 3x3; Pad: 1`	✓	✓
`MaxPool (2D)`	`Window: 2x2, Stride: 2, Pad: 0`	-	-
`Dropout (0.5)`			
`Conv (2D)`	`Filters: 256; Kernel: 3x3; Pad: 1`	✓	✓
`Conv (2D)`	`Filters: 256; Kernel: 3x3; Pad: 1`	✓	✓
`Conv (2D)`	`Filters: 256; Kernel: 3x3; Pad: 1`	✓	✓
`MaxPool (2D)`	`Window: 2x2, Stride: 2, Pad: 0`	-	-
`Dropout (0.5)`			
`Conv (2D)`	`Filters: 512; Kernel: 3x3; Pad: 0`	✓	✓
`Conv (2D)`	`Filters: 256; Kernel: 1x1; Pad: 0`	✓	✓
`Conv (2D)`	`Filters: 128; Kernel: 1x1; Pad: 0`	✓	✓
`AvgPool (2D)`	`Window: 6x6, Stride: 2, Pad: 0`	-	-
Vectorize to $z \in \mathbb{R}^{128}$ (this is where the topological regularizer operates)			
`FullyConn.`	`128 -> 10`	-	-

Table S1. CNN-13 architecture. The network part up to the vectorization operation constitutes φ.

On MNIST, we use a simpler CNN architecture, listed in Table S2.

		BN	LReLU
Input	28×28 grayscale image		
Conv (2D)	Filters: 8; Kernel: 3x3; Pad: 1	✓	✓
MaxPool (2D)	Window: 2x2, Stride: 2, Pad: 0	-	-
Conv (2D)	Filters: 32; Kernel: 3x3; Pad: 1	✓	✓
MaxPool (2D)	Window: 2x2, Stride: 2, Pad: 0	-	-
Conv (2D)	Filters: 64; Kernel: 3x3; Pad: 1	✓	✓
MaxPool (2D)	Window: 2x2, Stride: 2, Pad: 0	-	-
Conv (2D)	Filters: 128; Kernel: 3x3; Pad: 1	✓	✓
MaxPool (2D)	Window: 2x2, Stride: 2, Pad: 0	-	-
Vectorize to $z \in \mathbb{R}^{128}$ (this is where the topological regularizer operates)			
FullyConn.	128 -> 10	-	-

Table S2. MNIST CNN architecture.

S6.2. Optimization & Augmentation

For optimization, we use SGD with momentum (set to 0.9). As customary in the literature (see e.g., Verma et al., 2019b), training images for CIFAR10 and SVHN are augmented by (1) zero-padding images by 2 pixel on each side, followed by random cropping of a 32×32 region and (2) random horizontal flipping (with probability 0.5). On MNIST, no augmentation is applied. All images are further normalized by subtracting the mean and dividing by the standard deviation. Importantly, these statistics are computed for each cross-validation split separately (from the training instances), as this is the only practical choice in a small sample-size regime.

S6.3. Hyper-parameter settings

Except for the last row, Table 1 lists the *best achievable error* for different regularization approaches over a hyper-parameter grid to establish a *lower bound* on the obtainable error.

Across all experiments, weight decay on φ is fixed to $1\text{e}{-3}$. Due to the use of batch normalization, this primarily affects the effective learning rate (see van Laarhoven, 2017; Zhang et al., 2019). On MNIST, we fix the initial learning rate to 0.1. On SVHN and CIFAR10, we additionally experimented with an initial learning rate of 0.3 and 0.5 and include these in our hyper-parameter grid. The learning rate is annealed following the cosine learning rate annealing proposed in (Loshchilov & Hutter, 2017).

Regularization. The hyper-parameter grid is constructed as follows: *weight decay* on γ is varied in $\{1\text{e}{-3}, 5\text{e}{-4}, 1\text{e}{-4}\}$. For *Jacobian regularization* (Hoffman et al., 2019), the weighting of the regularization term is varied in $\{1\text{e}{-3}, 0.05, 0.01, 0.1\}$. For *DeCov* (Cogswell et al., 2016), *VR* and *cw-CR/VR* (Choi & Rhee, 2019), weighting of the regularization term is varied in $\{1\text{e}{-4}, 1\text{e}{-3}, 0.01, 0.1\}$. All these different choices are evaluated over 10 cross-validation runs with exactly the same training/testing split configuration.

Topological regularization. To evaluate the sub-batch construction in combination with our proposed topological regularizer, the initial learning rate on MNIST is fixed to 0.1 and 0.5 for SVHN and CIFAR10. With topological regularization enabled, this always produced stable results. Importantly, weight decay for γ is fixed to to 0.001, except for the CIFAR10-1k experiment, where we set it to $5\text{e}{-4}$. *The lowest achievable error is thus only selected by varying* β in $[0.1, 1.9]$.

To obtain the last row of Table 1, we no longer sweep over β, but select β via cross-validation over held-out validation sets of size 250 on SVHN and MNIST, and 500/1,000 on CIFAR10, respectively.

The full, `PyTorch`-compatible, source code will be made publicly available at `https://github.com/c-hofer/topologically_densified_distributions`.