Appendices

This is the Supplementary Material for "Few-shot Domain Adaptation by Causal Mechanism Transfer." Table 3 summarizes the abbreviations and the symbols used in the paper.

Table 3: List of abbreviations and symbols used in the paper.

Abbreviation / Symbol	Meaning
DA	Domain adaptation
TA	Transfer assumption
SEM	Structural equation model
GCM	Graphical causal model
SCM	Structural causal model
IC	Independent component
ICA	Independent component analysis
GCL	Generalized contrastive learning
i.i.d.	Independent and identically distributed
$[N]$	$\{1, 2, \ldots, N\}$ where $N \in \mathbb{N}$
$\|\cdot\|_{W^{k,p}}$	The (k, p)-Sobolev norm
X	The predictor random vector ($\mathbb{R}^{D-1}$-valued)
Y	The predicted random variable ($\mathbb{R}$-valued)
$Z = (X, Y)$	The joint random variable ($\mathbb{R}^D$-valued)
S	The independent component vector ($\mathbb{R}^D$-valued)
$\mathcal{X} \subset \mathbb{R}^{D-1}$	The space of X
$\mathcal{Y} \subset \mathbb{R}$	The space of Y
$\mathcal{Z} \subset \mathbb{R}^D$	The space of $Z = (X, Y)$
$\mathcal{G} \subset \{g : \mathbb{R}^{D-1} \to \mathbb{R}\}$	Predictor hypothesis class
$\ell : \mathcal{G} \times \mathcal{Z} \to [0, B_\ell]$	Loss function
$R(g)$	Target domain risk $\mathbb{E}_{p_{\mathrm{Tar}}} \ell(g, Z)$
$g^* \in \mathcal{G}$	Minimizer of target domain risk
$\mathcal{Q}$	The set of independent distributions
f	Ground truth mixing function
p_{Tar}	The target joint distribution
p_k	The joint distribution of source domain k
$q_{\mathrm{Tar}} \in \mathcal{Q}$	The target independent component (IC) distribution
$q_k \in \mathcal{Q}$	The IC distribution of source domain k
D	The dimension of $\mathcal{Z}$
K	The number of source domains
n_{Tar}	The size of the target labeled sample
n_k	The size of the labeled sample from source domain k
$\mathcal{D}_{\mathrm{Tar}} = \{Z_i\}_{i=1}^{n_{\mathrm{Tar}}}$	Target labeled data set
$\mathcal{D}_k = \{Z_{k,i}^{\mathrm{Src}}\}_{i=1}^{n_k}$	Source labeled data set of source domain k
$\hat{R}(g)$	The ordinary empirical risk estimator
$\check{R}(g)$	The proposed risk estimator (Eq. (2))
$\hat{f}$	The estimator of f
$\{\psi_d\}_{d=1}^D$	The penultimate layer functions composed with f during GCL

A. Preliminary: Nonlinear ICA

Here, we use the same notation as the main text. The recently developed nonlinear ICA provides an algorithm to estimate the mixing function f. For the case of nonlinear f, the impossibility of identification (i.e., consistent estimation) of f in the one-sample i.i.d. case had been established more than two decades ago (Hyvärinen & Pajunen, 1999). However, recently,

various conditions have been proposed under which f can be identified with the help of auxiliary information (Hyvärinen & Morioka, 2016; 2017; Hyvärinen et al., 2019; Khemakhem et al., 2019).

The identification condition that is directly relevant to this paper is that of the generalized contrastive learning (GCL) proposed in Hyvärinen et al. (2019). Hyvärinen et al. (2019) assumes that an auxiliary variable u_i from some measurable set $\mathcal{U}$ is obtained for each data point as $\{(z_i, u_i)\}_{i=1}^n$ and that the ICs $S = (S^{(1)}, \ldots, S^{(D)})$ are conditionally independent given u:

$$q(s|u) = \prod_{d=1}^{D} q^{(d)}(s^{(d)}|u).$$

Under such conditions, GCL estimates f by training a classification function

$$r_{\hat{f},\boldsymbol{\psi}}(z, u) = \sum_{d=1}^{D} \psi_d(\hat{f}^{-1}(z)_d, u) \tag{4}$$

parametrized by $\hat{f}$ and $\{\psi_d\}_{d=1}^D$ with the logistic loss for classifying

$$(z, u) \text{ vs. } (z, \tilde{u}),$$

where $\tilde{u} \in \mathcal{U} \setminus \{u\}$. The key condition for the identification of f is the following.

Assumption 1 (Assumption of variability; Hyvärinen et al., 2019, Theorem 1)**.** *For any z, there exist $2D + 1$ distinct points in $\mathcal{U}$, denoted by $\{u_j\}_{j=0}^{2D}$, such that the set of $(2D)$-dimensional vectors $\{w(z|u_j) - w(z|u_0)\}_{j=1}^{2D}$ are linearly independent, where*

$$w(z|u) := \left(\frac{\partial \log q^{(1)}(z_1|u)}{\partial z_1}, \ldots, \frac{\partial \log q^{(D)}(z_D|u)}{\partial z_D}, \frac{\partial^2 \log q^{(1)}(z_1|u)}{\partial z_1^2}, \ldots, \frac{\partial^2 \log q^{(D)}(z_D|u)}{\partial z_D^2} \right).$$

Under Assumption 1 and some regularity conditions, Theorem 1 of Hyvärinen et al. (2019) states that the transformation $\hat{f}$ in Eq. (4) trained by GCL is a consistent estimator of f upto additional dimension-wise invertible transformations. Note that the assumption is intrinsically difficult to confirm based on data due to the unsupervised nature of the problem setting. In this paper, we use the source domain index as the auxiliary variable and employ GCL for domain adaptation. The present version of Assumption 1 requires that we have at least $2D + 1$ distinct source domains. Although this condition can be restrictive in high-dimensional data, we conjecture that there is a possibility for this assumption to be made less stringent in the future because the identification condition is only known to be a sufficient condition, not a necessary condition. However, pursuing a refinement of the identification condition is out of the scope of this paper. Among the various methods for nonlinear ICA, we chose to use GCL (Hyvärinen et al., 2019) because it can operate under a nonparametric assumption on the IC distributions whereas other nonlinear ICA methods (Hyvärinen & Morioka, 2016; 2017; Khemakhem et al., 2019) may require parametric assumptions.

B. Experiment Details

Here, we describe more implementation details of the experiment. Our experiment code can be found at `https://github.com/takeshi-teshima/few-shot-domain-adaptation-by-causal-mechanism-transfer`.

B.1. Dataset details

Gasoline consumption data. The data was downloaded from `http://bcs.wiley.com/he-bcs/Books?action=resource&bcsId=4338&itemId=1118672321&resourceId=13452`.

B.2. Model details: Invertible neural networks

Here, we describe the details of the Glow architecture (Kingma & Dhariwal, 2018) used in our experiments. Glow consists of three types of layers which are invertible *by design*, namely affine coupling layers, 1×1 convolution layers, and activation normalization (actnorm) layers. In our implementation, we use actnorm as the first layer, and each of the subsequent layers consists of a 1×1 convolution layer followed by an affine coupling layer.

Affine coupling layers. The coefficients s and t for affine coupling layers in the notation of Kingma & Dhariwal (2018) are parametrized by two one-hidden-layer neural networks whose number of hidden units is the same and the first layer parameter is shared. The activation functions of the first layer, the second layer of s, and the second layer of t are the rectified linear unit (ReLU) activation (LeCun et al., 2015), the hyperbolic tangent function, and the linear activation function, respectively. A standard practice of affine coupling layers is to compose the coefficient s with an exponential function $x \mapsto \exp(x)$ so as to simplify the computation of the log-determinant of the Jacobian (Kingma & Dhariwal, 2018). In our implementation, since we do not require the computation of the log-determinant, we omit this device and instead compose $x \mapsto (x + 1)$. The addition of 1 shifts the parameter space so that $(s, t) = (0, 0)$ corresponds to the the identity map, where 0 denotes the constant zero function. The split of the affine coupling layers is fixed at $(\lfloor \frac{D}{2} \rfloor, D - \lfloor \frac{D}{2} \rfloor)$.

1×1 **convolution layers.** We initialize the parameters of the neural networks by $\mathcal{N}(0, \frac{1}{m})$ where m is the number of parameters of each layer and $\mathcal{N}$ is the normal distribution.

B.3. Model details: Penultimate layer networks

We initialize the parameter for each layer of ψ_d by $\mathrm{U}(-\sqrt{\frac{1}{m}}, \sqrt{\frac{1}{m}})$, where m is the number of input features and U is the uniform distribution.

B.4. Training details

During the training of GCL, we fix the batch size at 32.

B.5. Compared methods details

Here, we detail the methods compared through the experiment. Note that the present paper focuses on regression problems as our approach is based on ICA, hence the methods for classification domain adaptation are not comparable.

TrAdaBoost. As suggested in Pardoe & Stone (2010), we use the linear loss function and set the maximum number of internal boosting iterations at 30.

GDM. We fix the number of sampling required for approximating the maximization in the generalization discrepancy at 200. This method presumes using hypothesis classes in a reproducing kernel Hilbert space (RKHS).

Copula. For this model, the probabilistic model of non-parametric R-vine copula of depth 1 is used following Lopez-paz et al. (2012). Kernel density estimators with RBF kernel are used for estimating the marginal distributions and the copulas. The bandwidths of the RBF kernels are determined using the rule-of-thumb implemented as "normal-reference" in the *np* package of *R* language (Hayfield & Racine, 2008). The predictions are made by numerically aggregating the estimated conditional distribution over the interval $[\min_i Y_i - 2\sigma, \max_i Y_i + 2\sigma]$ where σ denotes the square root of the unbiased variance of $\{Y_i\}_{i=1}^{n_{\mathrm{src}}}$. The aggregation is performed by discretizing the interval into a grid of 300 points. The level of the two-sample test is fixed at 0.05 for all combination of the two-sample tests following the experiment code of Lopez-paz et al. (2012). This method is a single-source domain adaptation method and we pool all source domain data for adaptation.

C. Details and Proofs of Theorem 2

Here, we detail the assumptions, the statement, and the proof of Theorem 2.

C.1. Notation

To make the proof self-contained, we first recall some general and problem-specific notation. In the notation here, we omit the domain identifiers from the distributions and the sample size, such as *Tar* or *Src*, because only the target domain data or their distributions appear in the proofs. The theorem holds regardless of how $\hat{f}$ is estimated as long as $\hat{f}$ is independent of the target domain data. In the proof, we extend the maximal discrepancy bound of U-statistics previously proved for the case of degree-2 in Rejchel (2012), to allow higher degrees.

General mathematical notation. We denote the set of natural numbers (resp. real numbers) by $\mathbb{N}$ (resp. $\mathbb{R}$). For any $N \in \mathbb{N}$, we define $[N] := \{1, 2, \dots, N\}$. We use $\binom{a}{b}$ to denote the number of b-combinations of a elements. For a finite set A, the notation $\overline{\sum_{a \in A}}$ denotes the operator to take an average over A, i.e., $\overline{\sum_{a \in A}} h = \frac{1}{|A|} \sum_{a \in A} h(a)$. For a D-dimensional function h, we denote its j-th dimension ($j \in [D]$) by suffixing h_j. For a vector s, we denote its j-th element by $s^{(j)}$. We denote the Jacobian determinant of a differentiable function ψ at a by $\mathrm{J}\psi(a) := \det \frac{\mathrm{d}\psi(a)}{\mathrm{d}a}$. We denote the identity matrix by I regardless of the size of the matrices when there is no ambiguity. For finite dimensional vectors, we denote the 2-norm by $\|\cdot\|_{\ell^2}$ and the 1-norm by $\|\cdot\|_{\ell^1}$. For square matrices, we denote the operator-2 norm by $\|\cdot\|_{\mathrm{op}}$ and the operator-1 norm by $\|\cdot\|_{\mathrm{op}(1)}$. We use $W^{k,p}$ to denote the Sobolev space (on $\mathbb{R}^D$) of order k and define its associated norm by $\|h\|_{W^{k,p}} := \left(\sum_{|\alpha| \leq k} \left\|h^{(\alpha)}\right\|_{L^p}^p\right)^{1/p}$ where α is a multi-index and $h^{(\alpha)}$ denotes the partial derivative $\frac{\partial^{|\alpha|} h}{\partial s_1^{\alpha_1} \dots \partial s_D^{\alpha_D}}$ (Adams & Fournier, 2003, Paragraph 3.1). We let $\mathfrak{S}_D$ be the degree-D symmetric group, $\mathfrak{G}_j^D := \{\tau : [D] \to [j] \mid \tau \text{ is surjective}\}$ be the set of j grouping of indices in $[D]$, and $\mathfrak{I}_j^n := \{\rho : [j] \to [n] \mid \rho \text{ is injective}\}$ be the set of all size-j combinations (without replacement) of indices in $[n]$.

Distributions and expectations. We denote by $\mathcal{Q}$ the set of all factorized distributions on $\mathbb{R}^D$ with absolutely continuous marginals. For a measure P, we denote its j-product measure by $P^j := P \otimes \dots \otimes P$ (repeated j times). We assume that all measures appearing in this proof are absolutely continuous with respect to the Lebesgue measure. The push-forward of a distribution p by a function h is denoted by $h_\sharp p$. The expectation of a function h with respect to measure P is denoted by Ph (if it exists) by abuse of notation. We also abuse the notation to use $\psi(s, P, \dots, P)$ as the shorthand for $P^{D-1}\psi(s, S'_2, \dots, S'_D)$ where $\{S'_d\}_{d=2}^D \overset{\text{i.i.d.}}{\sim} P$.

C.2. Problem setup

We denote the target domain distribution by p.

We fix a hypothesis class $\mathcal{G}(\subset \{g : \mathbb{R}^{D-1} \to \mathbb{R}\})$, and our goal is to find a $g \in \mathcal{G}$ such that the risk functional

$$R(g) := \int p(z)\ell(g, z)\mathrm{d}z$$

is small, where $\ell : \mathcal{G} \times \mathbb{R}^D \to \mathbb{R}_{\geq 0}$ is a loss function. We denote by g^* a minimizer of R (assuming it exists). To this end, we are given the training data $\mathcal{D} := \{Z_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p$. Throughout, we assume $n \geq D$. To complement the smallness of n, we assume the existence of a generative mechanism. Concretely, we assume that there exists a diffeomorphism $f : \mathbb{R}^D \to \mathbb{R}^D$ such that $q := (f^{-1})_\sharp p$ satisfies $q \in \mathcal{Q}$. With this transform, the original risk functional is also expressed as

$$R(g) = \int q(s)\ell(g, f(s))\mathrm{d}s.$$

As an estimator of f, we are given another diffeomorphism $\hat{f} : \mathbb{R}^D \to \mathbb{R}^D$ such that $\hat{f} \simeq f$. With this $\hat{f}$, the proposed method converts the dataset $\mathcal{D}$ by $S_i := \hat{f}(Z_i)$. We can regard $\check{\mathcal{D}} := \{S_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \check{q}$, where $\check{q} := (\hat{f}^{-1} \circ f)_\sharp q$. We use Q (resp. $\check{Q}$) to denote the probability measure corresponding to the density q (resp. $\check{q}$). This conversion results in the relation:

$$\check{q}(s) = q(f^{-1} \circ \hat{f}(s)) \left|(\mathrm{J}f^{-1} \circ \hat{f})(s)\right|.$$

As a candidate hypothesis $g \in \mathcal{G}$, the proposed method selects a minimizer $\check{g} \in \mathcal{G}$ of the proposed risk estimator $\check{R}$ defined as

$$\check{R}(g) := \frac{1}{n^D} \sum_{(i_1, \dots, i_D) \in [n]^D} \ell(g, \hat{f}(\hat{s}_{i_1}^{(1)}, \dots, \hat{s}_{i_D}^{(D)})). \tag{5}$$

In the proof, we evaluate its concentration around the expectation $\bar{\check{R}}(g) := \mathbb{E}_{\check{\mathcal{D}}}\check{R}(g)$. We use $\mathbb{E}_{\check{\mathcal{D}}}$ to denote the expectation with respect to $\check{\mathcal{D}}$. Let $\bar{\check{g}}$ denote a hypothesis which minimizes $\bar{\check{R}}(g)$ (assuming it exists).

In what follows, for notational simplicity, we define the D-variate symmetric function $\tilde{\ell}$ as

$$\tilde{\ell}(s_1, \dots, s_D) = \overline{\sum_{\pi \in \mathfrak{S}_D}} \ell(g, \hat{f}(s_{\pi(1)}^{(1)}, \dots, s_{\pi(D)}^{(D)})),$$

where $\overline{\sum_{\pi \in \mathfrak{S}_D}}$ indicates an averaging operation over all permutations (without replacement) of $[D]$. We use $\hat{\mathbb{E}}_n$ to denote the sample average operator with respect to $\mathcal{D}$ or $\check{\mathcal{D}}$, depending on the context.

C.3. Assumptions

Assumption 2 (The underlying density function is bounded and Lipschitz continuous)**.** *Assume*

$$B_q := \sup_{s \in \mathbb{R}^D} q(s) < \infty, \quad L_q := \sup_{s_1 \neq s_2} \frac{|q(s_1) - q(s_2)|}{\|s_1 - s_2\|} < \infty.$$

Assumption 3 (f^{-1} is Lipschitz continuous and Hölder continuous)**.** *We assume $f^{-1} \in C^{1,1}$ where $C^{1,1}$ is the $(1,1)$-Hölder space (Adams & Fournier, 2003, Paragraph 1.29) and*

$$L_{f^{-1}} := \sup_{z_1 \neq z_2} \frac{\|f^{-1}(z_1) - f^{-1}(z_2)\|}{\|z_1 - z_2\|} < \infty.$$

Assumption 4 (Bounded derivatives of f and f^{-1})**.** *Assume that*

$$B^{\infty}_{\partial f} := \sup_{s \in \mathbb{R}^D} \left\| \frac{\mathrm{d}f}{\mathrm{d}s}(s) \right\|_{\infty} < \infty, \quad B^{\infty}_{\partial f^{-1}} := \sup_{z \in \mathbb{R}^D} \left\| \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(z) \right\|_{\infty} < \infty.$$

where $\|\cdot\|_{\infty}$ denotes the maximum absolute value of the elements of a matrix.

Assumption 5 (Loss function is bounded and uniformly Lipschitz continuous in Z)**.** *The considered loss function takes values in a bounded interval:*

$$\ell : \mathcal{G} \times \mathcal{Z} \to [0, B_\ell],$$

where $0 < B_\ell < \infty$. Also assume

$$L_{\ell_{\mathcal{G}}} := \sup_{g \in \mathcal{G}} \sup_{z_1 \neq z_2} \frac{|\ell(g, z_1) - \ell(g, z_2)|}{\|z_1 - z_2\|} < \infty.$$

Assumption 6 (Estimated feature extractor)**.** *Assume $\hat{f}$ is independent of $\mathcal{D}$ and that $f_j - \hat{f}_j \in W^{1,d}$ for all $(j,d) \in [D] \times [D]$.*

Although $\hat{f}$ and f are assumed to be diffeomorphisms in the classical sense (implying that they are strongly differentiable), we introduce the Sobolev space because we want to measure their difference and their difference of derivatives in terms of integration.

Assumption 7 (Entropic condition: Euclidean class (Sherman, 1994))**.** *The function class $\Phi := \{\tilde{\ell} : g \in \mathcal{G}\}$ is Euclidean for the envelope F and constants A and V (Sherman, 1994), i.e., if μ is a measure for which $\mu F^2 < \infty$, then*

$$D(t, d_\mu, \Phi) \leq A t^{-V}, \quad 0 < t \leq 1,$$

where d_μ is the pseudo metric defined by

$$d_\mu(\phi_1, \phi_2) := \left[\mu |\phi_1 - \phi_2|^2 / \mu F^2\right]^{1/2}$$

for $\phi_1, \phi_2 \in \Phi$, and $D(t, d_\mu, \Phi)$ denotes the packing number of Φ with respect to the pseudometric d_μ and radius t. Without loss of generality, we take the envelope F such that $F(\cdot) \leq B_\ell$.

Assumption 8. *The hypothesis class $\mathcal{G}$ is expressive enough so that the model approximation error does not expand due to $\hat{f}$, i.e.,*

$$\inf_{g \in \mathcal{G}} \bar{\tilde{R}}(g) \leq \inf_{g \in \mathcal{G}} R(g)$$

The following complexity measure of $\mathcal{G}$, which is a version of Rademacher complexity for our problem setting, is used to state the theorem.

Definition 1 (Effective Rademacher complexity)**.** *Define*

$$\mathfrak{R}(\mathcal{G}) := \frac{1}{n}\mathbb{E}_{\check{\mathcal{D}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\sigma_i\mathbb{E}_{S'_2,\ldots,S'_D}[\tilde{\ell}(S_i, S'_2, \ldots, S'_D)]\right|\right]$$

where $\{\sigma_i\}_{i=1}^n$ *are independent uniform sign variables and* $S'_2, \ldots, S'_D \overset{i.i.d.}{\sim} \check{Q}$ *are independent of all other random variables.*

We provide the definition of the ordinary Rademacher complexity in Section C.8 and make a comparison of the two complexity measures in terms of how they depend on the input dimensionality.

C.4. Theorem statement

Our goal is to prove the following theorem. This is a detailed version of the theorem appearing in the main body of the paper.

Theorem 3 (Excess risk bound)**.** *Assume Assumptions 2, 3, 4, 5, 6, 7, and 8.*

Then for arbitrary $\delta, \delta' \in (0, 1)$*, we have with probability at least* $1 - (\delta + \delta')$*,*

$$\begin{aligned}
&R(\check{g}) - R(g^*)\\
&\leq \underbrace{C\sum_{j=1}^{D}\left\|f_j - \hat{f}_j\right\|_{W^{1,1}}}_{\text{Approximation error}} + \underbrace{4D\mathfrak{R}(\mathcal{G}) + 2DB_\ell\sqrt{\frac{\log 2/\delta}{2n}}}_{\text{Estimation error}} + \underbrace{\kappa_1(\delta', n) + DB_\ell B_q\kappa_2(f - \hat{f})}_{\text{Higher order terms}}.
\end{aligned}$$

where

$$\begin{aligned}
C &:= B_q L_{\ell_{\mathcal{G}}} + DB_\ell(L_q L_{f^{-1}} + B_q D C'_1),\\
C'_1 &:= (D+1)^{3/2}\left(B^{\infty}_{\partial f}\left(\sum_{k=1}^{D}\left\|f_k^{-1}\right\|_{C^{1,1}}\right) + B^{\infty}_{\partial f^{-1}}\right),\\
\kappa_1(\delta', n) &= \mathcal{O}(n^{-1})/\delta' + \mathcal{O}(n^{-1}),\\
\kappa_2(f - \hat{f}) &= \sum_{d=2}^{D}\binom{D}{d}C'_d\sum_{j=1}^{D}\left\|f_j - \hat{f}_j\right\|^d_{W^{1,d}}.
\end{aligned}$$

and $C'_d (d = 1, \ldots, D)$ *are constants determined in Lemma 11.*

Proof of Theorem 3. By adding and subtracting terms, we have

$$R(\check{g}) - R(g^*) = \underbrace{(R - \bar{\bar{R}})(\check{g})}_{\text{(A) Approximation error}} + \underbrace{\bar{\bar{R}}(\check{g}) - \bar{\bar{R}}(\bar{g})}_{\text{(B) Pseudo estimation error}} + \underbrace{\bar{\bar{R}}(\bar{g}) - R(g^*)}_{\text{(C) Additional model misspecification error}}.$$

Applying Lemma 1 to (A), Lemma 2 to (B), and Assumption 8 to (C), we obtain the assertion. □

As it can be seen from the proof above, Theorem 3 is proved in two parts, each corresponding to the two lemmas below. The first lemma evaluates the *approximation error* which reflects the fact that we are approximating f by $\hat{f}$.

Lemma 1 (Approximation error bound)**.** *Given Assumptions 2, 3, 4, 5, and 6. we have*

$$(R - \bar{\bar{R}})(\check{g}) \leq C\sum_{j=1}^{D}\left\|f_j - \hat{f}_j\right\|_{W^{1,1}} + DB_\ell B_q\kappa_2(f - \hat{f})$$

where C and $\kappa_2(f-\hat{f})$ are

$$C := B_q L_{\ell_{\mathcal{G}}} + DB_\ell(L_q L_{f^{-1}} + B_q D C_1'),$$
$$\kappa_2(f-\hat{f}) := \sum_{d=2}^{D} \binom{D}{d} C_d' \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,d}}^d .$$

and $C_d'(d=1,\ldots,D)$ are constants determined in Lemma 11.

The second lemma evaluates the *pseudo estimation error* which reflects the fact that we rely on a finite sample to approximate the underlying distribution.

Lemma 2 (Pseudo estimation error bound)**.** *Assume that Assumptions 2 and 7 hold. Let the Rademacher complexity be defined as Definition 1. Then for any $\delta, \delta' \in (0,1)$, we have with probability at least $1-(\delta+\delta')$ that*

$$\bar{\tilde{R}}(\check{g}) - \bar{\tilde{R}}(\bar{\tilde{g}}) \leq 4Dw_D \mathfrak{R}(\mathcal{G}) + 2DB_\ell w_D \sqrt{\frac{\log 2/\delta}{2n}} + \underbrace{2w_D(D-1)\sum_{j=2}^{D} \frac{C_j}{\delta'} n^{-j/2} + 4B_\ell \sum_{j=1}^{D-1} w_j}_{\mathcal{O}(n^{-1})}$$

where $\{w_j\}_{j=1}^D$ are universal constants determined in Lemma 3, and $\{C_j\}_{j=2}^D$ are constants determined in Lemma 6. Note that $w_j = \mathcal{O}(n^{-(D-j)})$ and $w_D = \frac{n(n-1)\cdots(n-D+1)}{n^D} < 1$.

In what follows, we first present some basic facts in Section C.5 and provide the proofs for the lemmas. We provide the proof of Lemma 1 in Section C.7, and that of Lemma 2 in Section C.6.

C.5. V-statistic and U-statistic

The theoretical analysis is performed by interpreting the proposed risk estimator Eq. (5) as a *V-statistic* (explained shortly). The proofs will be based on applying the following facts in order:

1. V-statistic can be represented as a weighted average of *U-statistics* with degrees from 1 to D, and only the degree-D) term is the leading term.
2. The degree-D term is again decomposed into a degree-1 U-statistic and a set of *degenerate* U-statistics.
3. The degree-1 *U-statistic* is an i.i.d. sum admitting a Rademacher complexity bound.
4. The degenerate terms concentrate around zero following an exponential inequality under appropriate entropy conditions.

To consolidate the strategy given above, we describe what are V- and U-statistics, and how they relate to each other. These estimators emerge when we allow re-using the same data point repeatedly from a single sample to estimate a function which takes multiple data points.

V-statistic. For a given regular statistical functional of degree D (Lee, 1990):

$$\check{Q}^D \tilde{\ell} := \int \tilde{\ell}(s_1, \cdots, s_D) \check{q}(s_1) \cdots \check{q}(s_D) \mathrm{d}s_1 \cdots \mathrm{d}s_D, \tag{6}$$

its associated von-Mises statistic (V-statistic) is the following quantity (Lee, 1990):

$$V_n^D \tilde{\ell} := \frac{1}{n^D} \sum_{i_1=1}^{n} \cdots \sum_{i_D=1}^{n} \tilde{\ell}(S_{i_1}, \ldots, S_{i_D}).$$

Note that Eq. (6) does not coincide with the expectation of $V_n^D \tilde{\ell}$ in general, i.e., the V-statistic is generally not an unbiased estimator. However, it is known to be a consistent estimator of Eq. (6) (Lee, 1990).

U-statistic. Similarly, for a j-variate symmetric and integrable function $h(x_1, \ldots, x_j)$, its corresponding U-statistic (Lee, 1990) of degree j is

$$U_n^j h := \overline{\sum_{\rho \in \mathfrak{I}_j^n}} h(s_{\rho(1)}, \ldots, s_{\rho(j)}).$$

The V- and U-statistics are generalizations of the sample mean (which is the U- and V-statistics of degree 1). The important difference from the sample mean in higher degrees is that the summands may not be independent. To deal with the dependence, the following standard decompositions have been developed (Lee, 1990).

Lemma 3 (Decomposition of a V-statistic (Lee, 1990))**.** *A V-statistic can be expressed as a sum of U-statistics of degrees from* 1 *to* D *(Lee, 1990, Section 4.2, Theorem 1):*

$$V_n^D \tilde{\ell} = \sum_{j=1}^{D} w_j U_n^j \tilde{\ell}^{(j)}$$

where the weights w_j *and* j*-variate functions* $\tilde{\ell}^{(j)}$ *are*

$$w_j := \frac{1}{n^D} |\mathfrak{G}_j^D| \binom{n}{j}, \quad \tilde{\ell}^{(j)}(s_1, \ldots, s_j) := \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \tilde{\ell}(s_{\tau(1)}, \ldots, s_{\tau(D)}).$$

Proof. See (Lee, 1990, Section 4.2, Theorem 1 (p.183)). □

Remark 1. *The weights* $\{w_j\}_{j=1}^D$ *satisfy* $\sum_j w_j = 1$ *(Lee, 1990, Section 4.2, Theorem 1 (p.183)). We can also find the order of* w_j *with respect to* n *as:*

$$w_D = \frac{1}{n^D} \underbrace{|\mathfrak{G}_j^D|}_{D!} \binom{n}{D} = \frac{n(n-1)\cdots(n-D+1)}{n^D} = \mathcal{O}(1), \quad w_j = \mathcal{O}(n^{-(D-j)}), \quad \tilde{\ell}^{(D)} = \tilde{\ell}.$$

Lemma 4 (Hoeffding decomposition of a U-statistic (Sherman, 1994, p.449))**.** *A U-statistic with a symmetric kernel* ψ *can be decomposed as a sum of U-statistics of degrees from* 1 *to* D *as*

$$\begin{aligned} U_n^D \psi - \mathbb{E}_{\check{\mathcal{D}}} U_n^D \psi &= \sum_{j=1}^{D} U_n^j \psi_j \\ &= \hat{\mathbb{E}}_n \psi_1 + \sum_{j=2}^{D} U_n^j \psi_j \end{aligned} \tag{7}$$

where $\{\psi_j\}_{j=1}^D$ *are* j*-variate, symmetric and degenerate functions. Note that* $\mathbb{E}_{\check{\mathcal{D}}} U_n^D \psi = \check{Q}^D \psi$*. Here, a* j*-variate symmetric function* ψ_j *is said* degenerate *when*

$$\forall s_2, \ldots, s_j, \quad \psi_j(\check{Q}, s_2, \ldots, s_j) = 0.$$

Specifically, ψ_1 *is*

$$\begin{aligned} \psi_1(s) &= \psi(s, \check{Q}, \ldots, \check{Q}) + \cdots + \psi(\check{Q}, \ldots, \check{Q}, s) - D\check{Q}^D \psi \\ &= D \cdot (\psi(s, \check{Q}, \ldots, \check{Q}) - \check{Q}^D \psi) \quad (\textit{by symmetry}). \end{aligned} \tag{8}$$

For further details, see (Sherman, 1994, p.449). Note that in (Sherman, 1994, p.449), Eq. (7) is written using $\check{Q}^D \psi$ in place of $\mathbb{E}_{\check{\mathcal{D}}} U_n^D \psi$. This is because

$$\mathbb{E}_{\check{\mathcal{D}}} U_n^D \psi = U_n^D \mathbb{E}_{\check{\mathcal{D}}} \psi = U_n^D \check{Q}^D \psi = \check{Q}^D \psi$$

holds by linearity and symmetry.

Remark 2 (Connecting the lemmas to Section C.6)**.** *It can be easily checked by definition that the proposed risk estimator Eq. (5) takes the form of a V-statistic: $\check{R}(g) = V_n^D \tilde{\ell}$ for each $g \in \mathcal{G}$. Let us denote $\tilde{\ell}^*(s) := \tilde{\ell}(s, \check{Q}, \ldots, \check{Q})$. Then $\mathbb{E}_{\check{\mathcal{D}}} \tilde{\ell}^* = \check{Q}^D \tilde{\ell}$ holds by definition. Substituting these into Eq. (8), we have that Eq. (7) applied to $\psi = \tilde{\ell}$ is equivalent to*

$$U_n^D \tilde{\ell} - \mathbb{E}_{\check{\mathcal{D}}} U_n^D \tilde{\ell} = D \cdot (\hat{\mathbb{E}}_n \tilde{\ell}^* - \mathbb{E}_{\check{\mathcal{D}}} \tilde{\ell}^*) + \sum_{j=2}^{D} U_n^j \tilde{\ell}_j.$$

where $\{\tilde{\ell}_j\}_{j=2}^D$ are symmetric degenerate functions. In Section C.6, we first decompose $\check{R}(g)$ into a sum of U-statistics. After such conversion, we take a closer look at the leading term, $\hat{\mathbb{E}}_n \tilde{\ell}^*$.

C.6. Proof of pseudo estimation error bound

(Proof of Lemma 2). First, we have

$$\begin{aligned}\bar{\check{R}}(\check{g}) - \bar{\check{R}}(\bar{\check{g}}) = \bar{\check{R}}(\check{g}) - \check{R}(\check{g}) + \check{R}(\check{g}) - \bar{\check{R}}(\bar{\check{g}}) &\leq \bar{\check{R}}(\check{g}) - \check{R}(\check{g}) + \check{R}(\bar{\check{g}}) - \bar{\check{R}}(\bar{\check{g}}) \\ &\leq 2 \sup_{g \in \mathcal{G}} \left| \check{R}(g) - \bar{\check{R}}(g) \right|.\end{aligned}$$

Now the right-most expression can be decomposed as

$$\begin{aligned}&\sup_{g \in \mathcal{G}} |\check{R}(g) - \bar{\check{R}}(g)| = \sup_{g \in \mathcal{G}} |V_n^D \tilde{\ell} - \mathbb{E}_{\check{\mathcal{D}}} V_n^D \tilde{\ell}| \\ &\leq w_D \sup_{g \in \mathcal{G}} \left| U_n^D \tilde{\ell} - \mathbb{E}_{\check{\mathcal{D}}} U_n^D \tilde{\ell} \right| + \sum_{j=1}^{D-1} w_j \sup_{g \in \mathcal{G}} \left| U_n^j \tilde{\ell}^{(j)} - \mathbb{E}_{\check{\mathcal{D}}} U_n^j \tilde{\ell}^{(j)} \right| \quad (\because \text{Lemma 3}) \\ &\leq w_D \sup_{g \in \mathcal{G}} |U_n^D \tilde{\ell} - \mathbb{E}_{\check{\mathcal{D}}} U_n^D \tilde{\ell}| + 2B_\ell \sum_{j=1}^{D-1} w_j \\ &\leq w_D \left(\sup_{g \in \mathcal{G}} |\hat{\mathbb{E}}_n \tilde{\ell}_1| + \sum_{j=2}^{D} \sup_{g \in \mathcal{G}} |U_n^j \tilde{\ell}_j| \right) + 2B_\ell \sum_{j=1}^{D-1} w_j \quad (\because \text{Lemma 4}) \\ &= w_D \left(\underbrace{\sup_{g \in \mathcal{G}} |\hat{\mathbb{E}}_n D(\tilde{\ell}^* - \mathbb{E}_{\check{\mathcal{D}}} \tilde{\ell}^*)|}_{\text{Addressed in Lemma 5}} + \underbrace{\sum_{j=2}^{D} \sup_{g \in \mathcal{G}} |U_n^j \tilde{\ell}_j|}_{\text{Addressed in Lemma 6}} \right) + 2B_\ell \sum_{j=1}^{D-1} w_j.\end{aligned}$$

where $\tilde{\ell}_j$ are symmetric degenerate functions and $\tilde{\ell}^*$ is defined as in Remark 2. Applying Lemma 5 to the first term and Lemma 6 to the second term with the union bound, we obtain the assertion. □

In the last part of the proof we used the following lemmas. Because the leading term is an i.i.d. sum, the following Rademacher complexity bound can be proved.

Lemma 5 (U-process bound: the leading term)**.** *Assume Assumption 2 holds. Then, we have with probability at least $1 - \delta$,*

$$\sup_{g \in \mathcal{G}} |\hat{\mathbb{E}}_n (\tilde{\ell}^* - \mathbb{E}_{\check{\mathcal{D}}} \tilde{\ell}^*)| \leq 2\mathfrak{R}(\mathcal{G}) + B_\ell \sqrt{\frac{\log(2/\delta)}{2n}},$$

where $\mathfrak{R}$ is defined in Definition 1.

Proof. Applying the standard one-sided Rademacher complexity bound based on McDiarmid's inequality (Mohri et al., 2012, Theorem 3.1) twice with the union bound, we obtain the lemma. □

The other terms than the leading term are degenerate U-statistics, hence the following holds under appropriate entropy assumptions.

Lemma 6 (U-process bound: degenerate terms (Sherman, 1994, Corollary 7))**.** *Assume Assumption 7. Then for each $j = 2, \dots, D$, there exist constants C_j such that for any $\delta \in (0,1)$, we have with probability at least $1 - \delta'/(D-1)$,*

$$\sup_{g \in \mathcal{G}} |U_n^j \tilde{\ell}_j| \leq \frac{(D-1)}{\delta'} C_j n^{-j/2}$$

where C_j depends only on A, V, and B_ℓ.

Proof. The proof follows a similar path as that of (Sherman, 1994, Corollary 7), but we provide more explicit expressions to inspect the order with respect to n. Let $\Phi^{(j)}_{\mathcal{G},\hat{f}} := \{\tilde{\ell}_j : g \in \mathcal{G}\}$. Then $\Phi^{(j)}_{\mathcal{G},\hat{f}}$ is Euclidean for an envelope F_j satisfying $\check{Q}^j F_j^2 < \infty$ by Lemma 6 in Sherman (1994) and Assumption 7. In addition, $\Phi^{(j)}_{\mathcal{G},\hat{f}}$ is a set of functions degenerate with respect to $\check{Q}$. Without loss of generality, we can take F_j such that $F_j \leq B_\ell$. Similarly to the proof of (Sherman, 1994, Corollary 4), we can apply (Sherman, 1994, Main Corollary) with $p = 1$ in their notation to obtain

$$\mathbb{E}_{\check{\mathcal{D}}} \sup_{g \in \mathcal{G}} |n^{j/2} U_n^j \tilde{\ell}_j| \leq \Gamma A^{1/2mp} (\check{Q}^j F_j^2)^{(\epsilon+\alpha)/2} \leq \underbrace{\Gamma A^{1/2mp} (B_\ell)^{\epsilon+\alpha}}_{=: C_j}$$

where Γ is a universal constant (Sherman, 1994, Main Corollary), $\epsilon \in (0,1)$ and m are chosen to satisfy $1 - V/2m > 1 - \epsilon$, and $\alpha = 1 - V/2m$. By applying Markov inequality, we have for arbitrary $u > 0$,

$$\mathbb{P}_{\check{\mathcal{D}}} \left(\sup_{g \in \mathcal{G}} |n^{j/2} U_n^j \tilde{\ell}_j| > u \right) \leq \frac{C_j}{u},$$

where $\mathbb{P}_{\check{\mathcal{D}}}(E)$ denotes the probability of the event E with respect to $\check{\mathcal{D}}$. Equating the right hand side with $\delta'/(D-1)$ and solving for u, we obtain the result. □

C.7. Proof of approximation error bound

(Proof of Lemma 1). Due to Lemma 3, we have

$$\begin{aligned}
\sup_{g \in \mathcal{G}} \left(R(g) - \bar{\bar{R}}(g) \right) &= \sup_{g \in \mathcal{G}} \left(R(g) - \mathbb{E}_{\check{\mathcal{D}}} V_n^D \tilde{\ell} \right) \\
&= \sup_{g \in \mathcal{G}} \left(\sum_{j=1}^{D} w_j (R(g) - \mathbb{E}_{\check{\mathcal{D}}} U_n^j \tilde{\ell}^{(j)}) \right) \\
&\leq w_D \sup_{g \in \mathcal{G}} \left(R(g) - \mathbb{E}_{\check{\mathcal{D}}} U_n^D \tilde{\ell}^{(D)} \right) + 2B_\ell \sum_{j=1}^{D-1} \underbrace{w_j}_{\mathcal{O}(n^{-(D-j)})} \\
&\leq w_D \sup_{g \in \mathcal{G}} \left(R(g) - \mathbb{E}_{\check{\mathcal{D}}} U_n^D \tilde{\ell}^{(D)} \right) + 2B_\ell \mathcal{O}(n^{-1})
\end{aligned}$$

By applying Lemmas 7 (with $j = D$), we obtain

$$\sup_{g \in \mathcal{G}} \left(R(g) - \mathbb{E}_{\check{\mathcal{D}}} U_n^D \tilde{\ell}^{(D)} \right) \leq \sup_{g \in \mathcal{G}} \left\| \ell(f(g, \cdot)) - \ell(g, \hat{f}(\cdot)) \right\|_{L^1(q)} + DB_\ell \left\| q - \check{q} \right\|_{L^1}.$$

The right-hand side can be further bounded by applying Lemmas 9 and 8 by

$$\begin{aligned}
&B_q L_{\ell_{\mathcal{G}}} \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,1}} + DB_\ell \left((L_q L_{f^{-1}} + B_q D C_1') \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,1}} + B_q \kappa_2 (f - \hat{f}) \right) \\
&\leq (B_q L_{\ell_{\mathcal{G}}} + DB_\ell (L_q L_{f^{-1}} + B_q D C_1')) \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,1}} + DB_\ell B_q \kappa_2 (f - \hat{f})
\end{aligned}$$

and hence the assertion of the lemma. □

The above proof combined three approximation bounds, which are shown in the following lemmas. The following lemma reduces the difference in the expectation of U-statistic into the differences in the loss function and the density function. Although we apply the following Lemma 7 only with $j = D$, we prove its general form for $j \in [D]$.

Lemma 7 (Approximation bound for U-statistic of degree-j)**.** *Fix $j \in [D]$. Assume Assumption 2. Then we have for any $g \in \mathcal{G}$,*

$$R(g) - \mathbb{E}_{\check{\mathcal{D}}} U_n^j \tilde{\ell}^{(j)} \leq \left\| \ell(g, f(\cdot)) - \ell(g, \hat{f}(\cdot)) \right\|_{L^1(q)} + j B_\ell \left\| q - \check{q} \right\|_{L^1}$$

Proof. Let us define a D-variate function $\ell^\dagger$ and a j-variate function $\ell^{\dagger(j)}$ (similarly to $\tilde{\ell}$ and $\tilde{\ell}^{(j)}$, respectively) by

$$\ell^\dagger(s_1, \ldots, s_D) := \overline{\sum_{\pi \in \mathfrak{S}_D}} \ell(g, f(s^{(1)}_{\pi(1)}, \ldots, s^{(D)}_{\pi(D)})),$$
$$\ell^{\dagger(j)}(s_1, \ldots, s_j) := \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \ell^\dagger(s_{\tau(1)}, \ldots, s_{\tau(D)}).$$

Then, recalling $Q \in \mathcal{Q}$, we can show $R(g) = Q^n(U_n^j \ell^{\dagger(j)})$ because

$$\begin{aligned}
Q^n(U_n^j \ell^{\dagger(j)}) &= Q^n(\overline{\sum_{\rho \in \mathfrak{I}_j^n}} \ell^{\dagger(j)}(S_{\rho(1)}, \ldots, S_{\rho(j)})) \\
&= Q^n(\overline{\sum_{\rho \in \mathfrak{I}_j^n}} \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \ell^\dagger(S_{\rho\circ\tau(1)}, \ldots, S_{\rho\circ\tau(D)})) \\
&= Q^n(\overline{\sum_{\rho \in \mathfrak{I}_j^n}} \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \overline{\sum_{\pi \in \mathfrak{S}_D}} \ell(g, f(S^{(1)}_{\rho\circ\tau\circ\pi(1)}, \ldots, S^{(D)}_{\rho\circ\tau\circ\pi(D)}))) \\
&= \overline{\sum_{\rho \in \mathfrak{I}_j^n}} \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \overline{\sum_{\pi \in \mathfrak{S}_D}} Q^n \ell(g, f(S^{(1)}_{\rho\circ\tau\circ\pi(1)}, \ldots, S^{(D)}_{\rho\circ\tau\circ\pi(D)})) \\
&= \overline{\sum_{\rho \in \mathfrak{I}_j^n}} \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \overline{\sum_{\pi \in \mathfrak{S}_D}} Q[\ell(g, f(S^{(1)}, \ldots, S^{(D)}))] \quad (\because Q \in \mathcal{Q}) \\
&= \overline{\sum_{\rho \in \mathfrak{I}_j^n}} \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \overline{\sum_{\pi \in \mathfrak{S}_D}} R(g) = R(g).
\end{aligned}$$

Combining this expression with Lemma 3,

$$\begin{aligned}
R(g) - \mathbb{E}_{\check{\mathcal{D}}} U_n^j \tilde{\ell}^{(j)} &= Q^n(U_n^j \ell^{\dagger(j)}) - \check{Q}^n(U_n^j \tilde{\ell}^{(j)}) \\
&= \underbrace{Q^n(U_n^j \ell^{\dagger(j)} - U_n^j \tilde{\ell}^{(j)})}_{A} + \underbrace{(Q^n - \check{Q}^n)(U_n^j \tilde{\ell}^{(j)})}_{B}
\end{aligned}$$

Now, A can be bounded from above as

$$\begin{aligned}
A &= Q^n(U_n^j \ell^{\dagger(j)} - U_n^j \tilde{\ell}^{(j)}) \\
&= \overline{\sum_{\rho \in \mathfrak{I}_j^n}} \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \overline{\sum_{\pi \in \mathfrak{S}_D}} Q^n(\ell(g, f(S^{(1)}_{\rho\circ\tau\circ\pi(1)}, \ldots, S^{(D)}_{\rho\circ\tau\circ\pi(D)})) - \ell(g, \hat{f}(S^{(1)}_{\rho\circ\tau\circ\pi(1)}, \ldots, S^{(D)}_{\rho\circ\tau\circ\pi(D)}))) \\
&= \overline{\sum_{\rho \in \mathfrak{I}_j^n}} \overline{\sum_{\tau \in \mathfrak{G}_j^D}} \overline{\sum_{\pi \in \mathfrak{S}_D}} Q(\ell(g, f(S^{(1)}, \ldots, S^{(D)})) - \ell(g, \hat{f}(S^{(1)}, \ldots, S^{(D)}))) \quad (\because Q \in \mathcal{Q}) \\
&\leq \left\| \ell(g, f(\cdot)) - \ell(g, \hat{f}(\cdot)) \right\|_{L^1(q)}
\end{aligned}$$

Then recalling Assumption 2, we can bound B from above as

$$
\begin{aligned}
B &= (Q^n - \check{Q}^n)(U_n^j \tilde{\ell}^{(j)}) = (Q^n - \check{Q}^n)\left(\overline{\sum_{\rho \in \mathfrak{I}_j^n}} \tilde{\ell}^{(j)}(S_{\rho(1)}, \ldots, S_{\rho(j)})\right) \\
&= \overline{\sum_{\rho \in \mathfrak{I}_j^n}} (Q^n - \check{Q}^n)\left(\tilde{\ell}^{(j)}(S_{\rho(1)}, \ldots, S_{\rho(j)})\right) = (Q^j - \check{Q}^j)(\tilde{\ell}^{(j)}(S_1, \ldots, S_j)) \quad (\because \text{ symmetry}) \\
&\leq B_\ell \int \left| \prod_{i=1}^j q(s_i) - \prod_{i=1}^j \check{q}(s_i) \right| \mathrm{d}s_1 \cdots \mathrm{d}s_j \\
&= B_\ell \int \left| \sum_{i=1}^j q(s_1) \cdots q(s_{i-1}) \cdot (q(s_i) - \check{q}(s_i)) \cdot \check{q}(s_{i+1}) \cdots \check{q}(s_j) \right| \mathrm{d}s_1 \cdots \mathrm{d}s_j \\
&\leq B_\ell \sum_{i=1}^j \int q(s_1) \cdots q(s_{i-1}) \cdot |q(s_i) - \check{q}(s_i)| \cdot \check{q}(s_{i+1}) \cdots \check{q}(s_j) \mathrm{d}s_1 \cdots \mathrm{d}s_j \\
&= B_\ell \sum_{i=1}^j \int |q(s_i) - \check{q}(s_i)| \mathrm{d}s_i = B_\ell \cdot j \left\| q - \check{q} \right\|_{L^1},
\end{aligned}
$$

which proves the assertion. □

Now the following lemmas bound each approximation terms in terms of the difference between f and $\hat{f}$.

Lemma 8 (Loss difference approximation)**.** *Assume Assumption 5. Then we have for any $g \in \mathcal{G}$,*

$$
\left\| \ell(g, f(\cdot)) - \ell(g, \hat{f}(\cdot)) \right\|_{L^1(q)} \leq B_q L_{\ell_\mathcal{G}} \sum_{j=1}^D \left\| f_j - \hat{f}_j \right\|_{W^{1,1}}
$$

Proof.

$$
\begin{aligned}
\left\| \ell(g, f(\cdot)) - \ell(g, \hat{f}(\cdot)) \right\|_{L^1(q)} &= \int |\ell(g, f(s)) - \ell(g, \hat{f}(s))| q(s) \mathrm{d}s \\
&\leq B_q \int L_{\ell_\mathcal{G}} \left\| f(s) - \hat{f}(s) \right\|_{\ell^2} \mathrm{d}s \\
&\leq B_q L_{\ell_\mathcal{G}} \int \left\| f(s) - \hat{f}(s) \right\|_{\ell^1} \mathrm{d}s \leq B_q L_{\ell_\mathcal{G}} \sum_{j=1}^D \left\| f_j - \hat{f}_j \right\|_{W^{1,1}}.
\end{aligned}
$$

□

Lemma 9 (Density difference approximation)**.** *Assume Assumptions 2, 3, and 4. Then we have*

$$
\left\| q - \check{q} \right\|_{L^1} \leq (L_q L_{f^{-1}} + B_q D C_1') \sum_{j=1}^D \left\| f_j - \hat{f}_j \right\|_{W^{1,1}} + B_q \kappa_2(f - \hat{f})
$$

where C_1' and $\kappa_2(f - \hat{f})$ are defined as in Lemma 11.

Proof. Since $\check{q}(s) = q(f^{-1} \circ \hat{f}(s)) \left|(\mathrm{J}f^{-1} \circ \hat{f})(s)\right|$, we have

$$\begin{aligned}
\|q - \check{q}\|_{L^1} &= \int \left|q(s) - q(f^{-1} \circ \hat{f}(s)) \left|(\mathrm{J}f^{-1} \circ \hat{f})(s)\right|\right| \mathrm{d}s \\
&\leq \int |q(s) - q(f^{-1} \circ \hat{f}(s))| \mathrm{d}s + \int q(f^{-1} \circ \hat{f}(s)) \left|1 - \left|(\mathrm{J}f^{-1} \circ \hat{f})(s)\right|\right| \mathrm{d}s \\
&\leq \underbrace{\int |q(s) - q(f^{-1} \circ \hat{f}(s))| \mathrm{d}s}_{\text{(A)}} + B_q \underbrace{\int \left|1 - (\mathrm{J}f^{-1} \circ \hat{f})(s)\right| \mathrm{d}s}_{\text{(B)}}
\end{aligned}$$

where the last line follows from the triangle inequality. Applying Lemma 10 to (A) and Lemma 11 to (B) yields the assertion. □

Lemma 10. *Assume Assumptions 2 and 3. Then,*

$$\int |q(s) - q(f^{-1} \circ \hat{f}(s))| \mathrm{d}s \leq L_q L_{f^{-1}} \sum_{j=1}^{D} \left\|f_j - \hat{f}_j\right\|_{W^{1,1}}$$

Proof. We have

$$\begin{aligned}
\int |q(s) - q(f^{-1} \circ \hat{f}(s))| \mathrm{d}s &= \int |q(f^{-1} \circ f(s)) - q(f^{-1} \circ \hat{f}(s))| \mathrm{d}s \\
&\leq L_q L_{f^{-1}} \int \left\|f(s) - \hat{f}(s)\right\|_{\ell^2} \mathrm{d}s \leq L_q L_{f^{-1}} \int \left\|f(s) - \hat{f}(s)\right\|_{\ell^1} \mathrm{d}s \\
&\leq L_q L_{f^{-1}} \sum_{j=1}^{D} \left\|f_j - \hat{f}_j\right\|_{W^{1,1}}
\end{aligned}$$

□

Lemma 11 (Jacobian difference approximation)**.** *Assume Assumptions 2 and 4. Then,*

$$\int \left|1 - (\mathrm{J}f^{-1} \circ \hat{f})(s)\right| \mathrm{d}s \leq D C_1' \sum_{j=1}^{D} \left\|f_j - \hat{f}_j\right\|_{W^{1,1}} + \kappa_2(f - \hat{f}),$$

where

$$\begin{aligned}
C_d' &:= (D+1)^{\frac{7}{2}d-2} \left((B_{\partial f}^{\infty})^d \left(\sum_{k=1}^{D} \left\|f_k^{-1}\right\|_{C^{1,1}} \right)^d + (B_{\partial f^{-1}}^{\infty})^d \right), \\
\kappa_2(f - \hat{f}) &:= \sum_{d=2}^{D} \binom{D}{d} C_d' \sum_{j=1}^{D} \left\|f_j - \hat{f}_j\right\|_{W^{1,d}}^d.
\end{aligned}$$

Proof. Applying Lemma 12 with $A := (\mathrm{J}f^{-1} \circ f)(s) = I$, we obtain

$$\begin{aligned}
\int \left|1 - (\mathrm{J}f^{-1} \circ \hat{f})(s)\right| \mathrm{d}s &= \int \left|(\mathrm{J}f^{-1} \circ f)(s) - (\mathrm{J}f^{-1} \circ \hat{f})(s)\right| \mathrm{d}s \\
&\leq \int \sum_{d=1}^{D} \binom{D}{d} \left\| \frac{\mathrm{d}f^{-1} \circ f}{\mathrm{d}s}(s) - \frac{\mathrm{d}f^{-1} \circ \hat{f}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}}^d \mathrm{d}s.
\end{aligned}$$

Now, each term in the integrand can be bounded from above as

$$
\begin{aligned}
&\left\| \frac{\mathrm{d}f^{-1}\circ f}{\mathrm{d}s}(s) - \frac{\mathrm{d}f^{-1}\circ \hat{f}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} \\
&= \left\| \left(\frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(f(s)) \right) \left(\frac{\mathrm{d}f}{\mathrm{d}s}(s) \right) - \left(\frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right) \left(\frac{\mathrm{d}\hat{f}}{\mathrm{d}s}(s) \right) \right\|_{\mathrm{op}} \\
&\leq \left\| \left(\frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(f(s)) - \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right) \left(\frac{\mathrm{d}f}{\mathrm{d}s}(s) \right) \right\|_{\mathrm{op}} + \left\| \left(\frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right) \left(\frac{\mathrm{d}f}{\mathrm{d}s}(s) - \frac{\mathrm{d}\hat{f}}{\mathrm{d}s}(s) \right) \right\|_{\mathrm{op}} \\
&\leq \left\| \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(f(s)) - \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right\|_{\mathrm{op}} \left\| \frac{\mathrm{d}f}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} + \left\| \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right\|_{\mathrm{op}} \left\| \frac{\mathrm{d}f}{\mathrm{d}s}(s) - \frac{\mathrm{d}\hat{f}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} \\
&\qquad (\because \text{ submultiplicativity (Golub \& Van Loan, 2013, Section 2.3.2)}) \\
&\leq \left\| \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(f(s)) - \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right\|_{\mathrm{op}} \left(D \cdot \left\| \frac{\mathrm{d}f}{\mathrm{d}s}(s) \right\|_{\infty} \right) + \left(D \cdot \left\| \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right\|_{\infty} \right) \left\| \frac{\mathrm{d}f}{\mathrm{d}s}(s) - \frac{\mathrm{d}\hat{f}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} \\
&\qquad (\because \|\cdot\|_{\mathrm{op}} \leq D \|\cdot\|_{\infty} \text{ (Golub \& Van Loan, 2013, Section 2.3.2)}) \\
&\leq \left\| \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(f(s)) - \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right\|_{\mathrm{op}} \cdot \left(DB^{\infty}_{\partial f} \right) + \left(DB^{\infty}_{\partial f^{-1}} \right) \cdot \left\| \frac{\mathrm{d}f}{\mathrm{d}s}(s) - \frac{\mathrm{d}\hat{f}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} \\
&\leq DB^{\infty}_{\partial f}\sqrt{D} \left\| \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(f(s)) - \frac{\mathrm{d}f^{-1}}{\mathrm{d}z}(\hat{f}(s)) \right\|_{\mathrm{op}(1)} + DB^{\infty}_{\partial f^{-1}}\sqrt{D} \left\| \frac{\mathrm{d}f}{\mathrm{d}s} - \frac{\mathrm{d}\hat{f}}{\mathrm{d}s} \right\|_{\mathrm{op}(1)} \\
&\qquad (\because \|\cdot\|_{\mathrm{op}} \leq \sqrt{D} \|\cdot\|_{\mathrm{op}(1)} \text{ (Golub \& Van Loan, 2013, Section 2.3.1)}) \\
&= D^{\frac{3}{2}} B^{\infty}_{\partial f} \max_{k\in[D]} \sum_{j=1}^{D} \left| \frac{\partial f_j^{-1}}{\partial z_k}(f(s)) - \frac{\partial f_j^{-1}}{\partial z_k}(\hat{f}(s)) \right| + D^{\frac{3}{2}} B^{\infty}_{\partial f^{-1}} \max_{k\in[D]} \sum_{j=1}^{D} \left| \frac{\partial f_j}{\partial s_k}(s) - \frac{\partial \hat{f}_j}{\partial s_k}(s) \right| \\
&\leq D^{\frac{3}{2}} B^{\infty}_{\partial f} \max_{k\in[D]} \sum_{j=1}^{D} \left\| f_j^{-1} \right\|_{C^{1,1}} \left\| f(s) - \hat{f}(s) \right\|_{\ell^2} + D^{\frac{3}{2}} B^{\infty}_{\partial f^{-1}} \sum_{k=1}^{D}\sum_{j=1}^{D} \left| \frac{\partial f_j}{\partial s_k}(s) - \frac{\partial \hat{f}_j}{\partial s_k}(s) \right| \\
&\leq D^{\frac{3}{2}} B^{\infty}_{\partial f} \left(\sum_{j=1}^{D} \left\| f_j^{-1} \right\|_{C^{1,1}} \right) \left\| f(s) - \hat{f}(s) \right\|_{\ell^1} + D^{\frac{3}{2}} B^{\infty}_{\partial f^{-1}} \sum_{k=1}^{D}\sum_{j=1}^{D} \left| \frac{\partial f_j}{\partial s_k}(s) - \frac{\partial \hat{f}_j}{\partial s_k}(s) \right| \\
&\qquad (\because \|\cdot\|_{\ell^2} \leq \|\cdot\|_{\ell^1} \text{ (Golub \& Van Loan, 2013, Section 2.2.2)}).
\end{aligned}
$$

When powered to d, this yields

$$
\begin{aligned}
&\left\| \frac{\mathrm{d}f^{-1}\circ f}{\mathrm{d}s}(s) - \frac{\mathrm{d}f^{-1}\circ \hat{f}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}}^{d} \\
&\leq (D^2+D)^{d-1} \Bigg[\sum_{j=1}^{D} \left(D^{3/2} B^{\infty}_{\partial f} \left(\sum_{k=1}^{D} \left\| f_k^{-1} \right\|_{C^{1,1}} \right) |f_j(s) - \hat{f}_j(s)| \right)^{d} \\
&\qquad + \sum_{k=1}^{D}\sum_{j=1}^{D} \left(D^{3/2} B^{\infty}_{\partial f^{-1}} \left| \frac{\partial f_j}{\partial s_k}(s) - \frac{\partial \hat{f}_j}{\partial s_k}(s) \right| \right)^{d} \Bigg]
\end{aligned}
$$

where we used $(\sum_{i=1}^{L} a_i)^d \leq L^{d-1}(\sum_{i=1}^{L} a_i^d)$ for $a_i \geq 0$, which follows from Hölder inequality. Hence,

$$
\begin{aligned}
&\int \left\| \frac{\mathrm{d}f^{-1} \circ f}{\mathrm{d}s}(s) - \frac{\mathrm{d}f^{-1} \circ \hat{f}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}}^{d} \mathrm{d}s \\
&\leq D^{\frac{5}{2}d-1}(D+1)^{d-1} \Bigg[\left(B_{\partial f}^{\infty} \sum_{k=1}^{D} \left\| f_k^{-1} \right\|_{C^{1,1}} \right)^{d} \sum_{j=1}^{D} \int |f_j(s) - \hat{f}_j(s)|^d \mathrm{d}s \\
&\qquad\qquad + (B_{\partial f^{-1}}^{\infty})^d \sum_{j=1}^{D} \left(\sum_{k=1}^{D} \int \left| \frac{\partial f_j}{\partial s_k}(s) - \frac{\partial \hat{f}_j}{\partial s_k}(s) \right|^d \mathrm{d}s \right) \Bigg] \\
&\leq (D+1)^{\frac{7}{2}d-2} \left((B_{\partial f}^{\infty})^d \left(\sum_{k=1}^{D} \left\| f_k^{-1} \right\|_{C^{1,1}} \right)^{d} \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,d}}^{d} + (B_{\partial f^{-1}}^{\infty})^d \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,d}}^{d} \right) \\
&\leq C_d' \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,d}}^{d}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\int \left| 1 - (\mathrm{J}f^{-1} \circ \hat{f})(s) \right| \mathrm{d}s \\
&\leq \sum_{d=1}^{D} \binom{D}{d} \int \left\| \frac{\mathrm{d}f^{-1} \circ f}{\mathrm{d}s}(s) - \frac{\mathrm{d}f^{-1} \circ \hat{f}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}}^{d} \mathrm{d}s \\
&\leq D C_1' \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,1}} + \underbrace{\sum_{d=2}^{D} \binom{D}{d} C_d' \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,d}}^{d}}_{\kappa_2(f - \hat{f})}
\end{aligned}
$$

□

Lemma 11 used the following lemma to bound the difference in Jacobian determinants.

Lemma 12 (Determinant perturbation bound (Ipsen & Rehman, 2008, Corollary 2.11))**.** *Let A and E be $D \times D$ complex matrices. Then,*

$$
|\det(A) - \det(A + E)| \leq \sum_{d=1}^{D} \binom{D}{d} \|A\|_{\mathrm{op}}^{D-d} \|E\|_{\mathrm{op}}^{d}.
$$

C.8. Comparison of Rademacher complexities

The following consideration demonstrates how the effective complexity measure $\mathfrak{R}$ in Theorem 3 resulting from the proposed method may enjoy a relaxed dependence on the input dimensionality compared to the ordinary empirical risk minimization. To do so, we first recall the definition of the ordinary Rademacher complexity and a standard performance guarantee derived based on it.

Definition 2 (Ordinary Rademacher complexity)**.** *The ordinary empirical risk minimization finds the candidate hypothesis by*

$$
\hat{g} \in \operatorname*{argmin}_{g \in \mathcal{G}} \hat{R}(g),
$$

where

$$
\hat{R}(g) := \frac{1}{n} \sum_{i=1}^{n} \ell(g, Z_i) = \frac{1}{n} \sum_{i=1}^{n} \ell(g, \hat{f}(S_i^{(1)}, \ldots, S_i^{(D)}))
$$

and the corresponding ordinary Rademacher complexity $\mathfrak{R}_{\mathrm{ord}}(\mathcal{G})$ is

$$\mathfrak{R}_{\mathrm{ord}}(\mathcal{G}) := \frac{1}{n}\mathbb{E}_{\check{\mathcal{D}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\sigma_i \ell(S_i^{(1)},\dots,S_i^{(D)})\right|\right]$$

where $\{\sigma_i\}_{i=1}^n$ are independent uniform sign variables and we denoted $\ell(s^{(1)},\dots,s^{(D)}) = \ell(g,\hat{f}(s^{(1)},\dots,s_i^{(D)}))$ by abuse of notation. This yields the standard Rademacher complexity based bound. Applying Lemma 5 and using the same proof technique, we have that with probability at least $1-\delta$,

$$R(\hat{g}) - R(g^*) \leq 2\sup_{g\in\mathcal{G}}|R(g)-\hat{R}(g)| \leq 4\mathfrak{R}_{\mathrm{ord}}(\mathcal{G}) + 2B_\ell\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Therefore, we the corresponding complexity terms are $\mathfrak{R}_{\mathrm{ord}}(\mathcal{G})$ and $D\mathfrak{R}(\mathcal{G})$. In Remark 3, we make a comparison of these two complexity measures by taking an example. To recall, the effective Rademacher complexity can be written as, in terms of the notation in this section,

$$\begin{aligned}\mathfrak{R}(\mathcal{G}) &= \frac{1}{n}\mathbb{E}_{\check{\mathcal{D}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\sigma_i\mathbb{E}_{S_2',\dots,S_D'}\tilde{\ell}(S_i,S_2',\dots,S_D')\right|\right]\\ &= \frac{1}{n}\mathbb{E}_{\check{\mathcal{D}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\sigma_i\mathbb{E}_{S_1',\dots,S_D'}\frac{1}{D}\left(\ell(S_i^{(1)},S_2'^{(2)},\dots,S_D'^{(D)})+\cdots+\ell(S_1'^{(1)},S_2'^{(2)},\dots,S_i^{(D)})\right)\right|\right]\end{aligned}$$

Remark 3 (Comparison of Radmacher complexities)**.** *As an example, consider $\mathcal{H}$, the set of L-Lipschitz functions (with respect to infinity norm) on the unit cube $[0,1]^d$. It is well-known that there exists a constant $C>0$ such that the following holds (Wainwright, 2019, Example 5.10, p.129) for sufficiently small $t>0$:*

$$\log\mathcal{N}(t,\mathcal{H},\|\cdot\|_\infty) \asymp (C/t)^d. \tag{9}$$

Here, $a(t)\asymp b(t)$ indicates that there exist $k_1,k_2>0$ such that, for sufficiently small t, it holds that $k_1 b(t)\leq a(t)\leq k_2 b(t)$. On the other hand, the well-known discretization argument implies that there exist constants c and B such that for any $t\in(0,B]$, the following relation between the Rademacher complexity and the metric entropy holds:

$$\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}) \leq t + c\sqrt{\frac{\log\mathcal{N}(t,\mathcal{H},\|\cdot\|_\infty)}{n}}. \tag{10}$$

Substituting Eq. (9) into Eq. (10), we can find that, for large enough n, the right hand side is minimized at $t = (c\cdot C^{\frac{d}{2}}\cdot\frac{d}{2})^{\frac{2}{2+d}}\cdot n^{-\frac{1}{2+d}}$. This yields

$$\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}) \leq \tilde{C}\cdot n^{-\frac{1}{2+d}} \tag{11}$$

with a new constant $\tilde{C} = \left(c\cdot C^{\frac{d}{2}}\cdot\frac{d}{2}\right)^{\frac{2}{2+d}} + c\cdot C^{\frac{d}{2}}\left(c\cdot C^{\frac{d}{2}}\cdot\frac{d}{2}\right)^{-\frac{d}{2+d}}$. Therefore, by substituting $d=D$ in Eq. (11), the metric-entropy based bound on the ordinary Rademacher complexity exhibits exponential dependence on the input dimension as

$$\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}) \leq \mathcal{O}\left(n^{-\frac{1}{2+D}}\right),$$

which is a manifestation of the curse of dimensionality. On the other hand, by following a similar calculation, we can see that the effective Rademacher complexity $\mathfrak{R}(\mathcal{H})$ avoids an exponential dependence on the input dimension D. By substituting $d=1$ in Eq. (11), we can see

$$D\mathfrak{R}(\mathcal{H}) \leq \mathfrak{R}_{\mathrm{ord}}(\mathcal{H}_1)+\cdots+\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}_D) \leq \mathcal{O}\left(n^{-\frac{1}{3}}\right),$$

where $\mathcal{H}_j := \{\mathbb{E}_{S'_1,\ldots,S'_D} h(S'_1{}^{(1)}, \ldots, S'_{j-1}{}^{(j-1)}, (\cdot)^{(j)}, S'_{j+1}{}^{(j+1)}, \ldots, S'_D{}^{(D)}) : h \in \mathcal{H}\}$*. This is because the Lipschitz constant of functions in* $\mathcal{H}_j$ *is at most* L *(i.e., the Lipschitz constant does not increase by the marginalization procedure) because for any* $h \in \mathcal{H}_j$*,*

$$
\begin{aligned}
&|h(x) - h(y)| \\
&= |\mathbb{E}_{S'_1,\ldots,S'_D}[h(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, x, S'^{(j+1)}_{j+1}, \ldots, S'^{(D)}_D) - h(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, y, S'^{(j+1)}_{j+1}, \ldots, S'^{(D)}_D)]| \\
&\leq \mathbb{E}_{S'_1,\ldots,S'_D} |h(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, x, S'^{(j+1)}_{j+1}, \ldots, S'^{(D)}_D) - h(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, y, S'^{(j+1)}_{j+1}, \ldots, S'^{(D)}_D)| \\
&\leq \mathbb{E}_{S'_1,\ldots,S'_D} L \cdot \|(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, x, S'^{(j+1)}_{j+1}, \ldots, S'^{(D)}_D) - (S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, y, S'^{(j+1)}_{j+1}, \ldots, S'^{(D)}_D)\| \\
&= \mathbb{E}_{S'_1,\ldots,S'_D} L \cdot \|(0, \ldots, 0, x - y, 0, \ldots, 0)\| \\
&= L \cdot |x - y|.
\end{aligned}
$$

C.9. Remark on higher order Sobolev norms

Here, we comment on how the term $\kappa_2(f - \hat{f})$ is treated as a higher order term of $f - \hat{f}$.

Remark 4 (Higher order Sobolev norms)**.** *Let us assume that* $\operatorname{supp} q \cup \operatorname{supp} \breve{q}$ *is contained in a compact set* $\tilde{\mathcal{S}}$ *for all* $\hat{f}$ *considered. Note that for* $d \in [D]$*,*

$$
\int_{\tilde{\mathcal{S}}} |h(s)|^d \mathrm{d}s \leq (V_{\tilde{\mathcal{S}}})^{\frac{d}{d-D}} \left(\int_{\tilde{\mathcal{S}}} |h(s)|^D \mathrm{d}s \right)^{d/D}
$$

by Hölder 's inequality, where we defined $V_{\tilde{\mathcal{S}}} := \int_{\tilde{\mathcal{S}}} 1 \mathrm{d}s$*, hence we have* $\|\cdot\|_{L^d(\tilde{\mathcal{S}})} \leq (V_{\tilde{\mathcal{S}}})^{\frac{1}{d-D}} \|\cdot\|_{L^D(\tilde{\mathcal{S}})}$*. By applying the relation to each term in the definition of* $\|\cdot\|_{W^{1,d}}$*, we obtain*

$$
\|f\|^d_{W^{1,d}} \leq (V_{\tilde{\mathcal{S}}})^{\frac{d}{d-D}} \|f\|^d_{W^{1,D}}
$$

Thus we obtain

$$
\begin{aligned}
\kappa_2(f - \hat{f}) &= \sum_{d=2}^{D} \binom{D}{d} C'_d \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|^d_{W^{1,d}} \\
&\leq \sum_{d=2}^{D} \binom{D}{d} (V_{\tilde{\mathcal{S}}})^{\frac{d}{d-D}} C'_d \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|^d_{W^{1,D}} \\
&\leq \mathcal{O}\left(\sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|^2_{W^{1,D}} \right) \quad (\hat{f} \to f).
\end{aligned}
$$

By also replacing $\sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,1}}$ *with* $\sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,D}}$ *in Theorem 3, we can see more clearly that* $\kappa_2(f - \hat{f})$ *is a higher order term of* $\sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,D}}$*.*

D. Details and Proofs of Theorem 1

Here, we provide the proof of Theorem 1. We reuse the notation and terminology from Section C of this Supplementary Material. We prove the uniformly minimum variance property of the proposed risk estimator under the ideal situation of $\hat{f} = f$.

Theorem 4 (Known causal mechanism case)**.** *Assume* $\hat{f} = f$*. Then, for all* $g \in \mathcal{G}$*, we have that* $\breve{R}(g)$ *is the uniformly minimum variance unbiased estimator of* $R(g)$*. As a special case, it has a smaller variance than the ordinary empirical risk estimator:* $\forall q \in \mathcal{Q}, \forall g \in \mathcal{G}, \mathrm{Var}(\breve{R}(g)) \leq \mathrm{Var}(\hat{R}(g))$*.*

Proof. The proof is a result of the following two facts. When $\breve{q} \in \mathcal{Q}$, the estimator $\breve{R}(g)$ becomes the generalized U-statistic of the statistical functional Eq. (6). Furthermore, when $\hat{f} = f$, Eq. (6) coincides with $R(g)$ because the approximation error

is zero. Since we assume $\hat{f} = f$ we have $\check{q} = q \in \mathcal{Q}$ and hence both of the statements above hold. Therefore, by Lemma 13, the first assertion of the theorem follows. The last assertion of the theorem follows as a special case as $\hat{R}(g)$ is an unbiased estimator of $R(g)$ for $q \in \mathcal{Q}$.

From here, we confirm the above statements by calculation. We first show that $\check{R}(g)$ is the generalized U-statistic. To see this, observe that the statistical functional Eq. (6) allows the following expression given $\check{q} \in \mathcal{Q}$:

$$
\begin{aligned}
&\int \tilde{\ell}(s_1, \ldots, s_D)\check{q}(s_1)\cdots\check{q}(s_D)\mathrm{d}s_1\cdots\mathrm{d}s_D \\
&= \int \overline{\sum_{\pi\in\mathfrak{S}_D}} \ell(g, \hat{f}(s^{(1)}_{\pi(1)}, \ldots, s^{(D)}_{\pi(D)}))\check{q}(s_1)\cdots\check{q}(s_D)\mathrm{d}s_1\cdots\mathrm{d}s_D \\
&= \int \overline{\sum_{\pi\in\mathfrak{S}_D}} \ell(g, \hat{f}(s^{(1)}_{\pi(1)}, \ldots, s^{(D)}_{\pi(D)}))\prod_d \check{q}^{(d)}(s^{(d)}_1)\cdots\prod_d \check{q}^{(d)}(s^{(d)}_D)\mathrm{d}s_1\cdots\mathrm{d}s_D \\
&= \int \overline{\sum_{\pi\in\mathfrak{S}_D}} \ell(g, \hat{f}(s^{(1)}_1, \ldots, s^{(D)}_1))\prod_d \check{q}^{(d)}(s_1)\mathrm{d}s_1 \\
&= \int \ell(g, \hat{f}(s^{(1)}, \ldots, s^{(D)}))\check{q}_1(s^{(1)})\cdots\check{q}_D(s^{(D)})\mathrm{d}s^{(1)}\cdots\mathrm{d}s^{(D)}.
\end{aligned}
$$

This is a regular statistical functional of degrees $(1, \ldots, 1)$ with the kernel $\ell(g, \hat{f}(\cdot, \ldots, \cdot))$. On the other hand, we have

$$
\check{R}(g) = \frac{1}{n^D}\sum_{(i_1,\ldots,i_D)\in[n]^D} \tilde{\ell}(S_{i_1}, \ldots, S_{i_D}) = \frac{1}{n^D}\sum_{(i_1,\ldots,i_D)\in[n]^D} \ell(g, \hat{f}(S^{(1)}_{i_1}, \ldots, S^{(D)}_{i_D}))
$$

because the summations run through all combinations with replacement. This combined with the fact that $\{S^{(d)}_i\}_{i,d}$ are jointly independent when $\check{q} \in \mathcal{Q}$ yields that $\check{R}(g)$ is the generalized U-statistic for Eq. (6).

Now we show that Eq. (6) coincides $R(g)$. Given $\hat{f} = f$, we have

$$
\begin{aligned}
R(g) &= \int q(s)\ell(g, f(s))\mathrm{d}s \\
&= \int q(s)\ell(g, \hat{f}(s))\mathrm{d}s \quad (\text{By } f = \hat{f}.) \\
&= \int q_1(s^{(1)})\cdots q_D(s^{(D)})\ell(g, \hat{f}(s^{(1)}, \ldots, s^{(D)}))\mathrm{d}s^{(1)}\cdots\mathrm{d}s^{(D)} \quad (\text{by } q \in \mathcal{Q}) \\
&= \int \check{q}_1(s^{(1)})\cdots\check{q}_D(s^{(D)})\ell(g, \hat{f}(s^{(1)}, \ldots, s^{(D)}))\mathrm{d}s^{(1)}\cdots\mathrm{d}s^{(D)} \quad (\text{by } q = \check{q}) \\
&= \int \tilde{\ell}(s_1, \ldots, s_D)\check{q}(s_1)\cdots\check{q}(s_D)\mathrm{d}s_1\cdots\mathrm{d}s_D. \quad (\because \text{ symmetry})
\end{aligned}
$$

□

The following well-known lemma states that a generalized U-statistic is a uniformly minimum variance unbiased estimator.

Lemma 13 (Uniformly minimum variance property of a generalized U-statistic)**.** *Let $\theta : \mathcal{Q} \to \mathbb{R}$ be a regular statistical functional with kernel $\psi : \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_L} \to \mathbb{R}$ (Clémençon et al., 2016), i.e.,*

$$
\theta(q) = \int \psi((x^{(1)}_1, \ldots, x^{(1)}_{k_1}), \ldots, (x^{(L)}_1, \ldots, x^{(L)}_{k_L}))\prod_{j=1}^{k_1} q_1(x^{(1)}_j)\mathrm{d}x^{(1)}_j \cdots \prod_{j=1}^{k_L} q_L(x^{(L)}_j)\mathrm{d}x^{(L)}_j.
$$

Given samples $\{x^{(l)}_i\}_{i=1}^{n_l} \overset{i.i.d.}{\sim} q_l (n_l \geq k_l \text{ and } l = 1, \ldots, L)$, let $\mathrm{GU}^{(k_1,\ldots,k_L)}_{(n_1,\ldots,n_L)}\psi$ be the corresponding generalized U-statistic

$$
\mathrm{GU}^{(k_1,\ldots,k_L)}_{(n_1,\ldots,n_L)}\psi := \frac{1}{\prod_l \binom{n_l}{k_l}}\sum \psi\left(\left(x^{(1)}_{i^{(1)}_1}, \ldots, x^{(1)}_{i^{(1)}_{k_1}}\right), \ldots, \left(x^{(L)}_{i^{(L)}_1}, \ldots, x^{(L)}_{i^{(L)}_{k_L}}\right)\right).
$$

where $\sum$ denotes that the indices run through all possible combinations (without replacement) of the indices. Then, $\mathrm{GU}^{(k_1,\ldots,k_L)}_{(n_1,\ldots,n_L)}\psi$ is the uniformly minimum variance unbiased estimator of θ on $\mathcal{Q}$.

Proof. The assertion can be proved in a parallel manner as the proof of (Lee, 1990, Section 1.1, Lemma B) □

Remark 5 (Relation to the UMVUE property of $\hat{R}(g)$)**.** *The result in Theorem 4 is not contradictory to the fact that the sample average $\hat{R}(g)$ is a U-statistic of degree-1 and hence the minimum variance among all unbiased estimator of $R(g)$ on $\mathcal{P}$, where $\mathcal{P}$ is a set of distributions containing all absolutely continuous distributions (Lee, 1990). Specifically, $\check{R}(g)$ is not generally an unbiased estimator of $R(g)$ on $\mathcal{P} \setminus \mathcal{Q}$, even if $\hat{f} = f$. While $\check{R}(g)$ satisfies the D-sample symmetry condition, the same does not hold for $\hat{R}(g)$. By restricting the attention to $\mathcal{Q}$, the estimator $\check{R}(g)$ achieves a smaller variance than $\hat{R}(g)$.*

E. Further Comparison with Related Work

Here, we provide an additional detailed comparison with the related work to complement Section 5 of the main text.

E.1. Comparison with Magliacane et al. (2018)

Magliacane et al. (2018) considered domain adaptation among different interventional states by using SCMs. Their problem setting and ours do not strictly include each other (the two settings are somewhat complementary), and their assumption may be more suitable for application fields with interventional experiments such as genomics, while ours may be more suited for fields with observational data such as health record analysis or economics. At the methodological level, Magliacane et al. (2018) takes a variable selection approach to find a subset so that the conditional distribution is invariant, whereas our paper takes a data augmentation approach via the estimation of the SEMs (in the reduced form).

The essential assumptions of Magliacane et al. (2018) are the existence of a separating set (with small "incomplete information bias") and the identifiability of such a set (yielded from Proposition 1, Assumption 1, and Assumption 2 (iii) in Magliacane et al. (2018)). A particularly plausible application conforming to the assumptions is, for example (but not limited to), genomics experiments. Part of the reason is that Assumption 2 (ii) and (iii) are likely to hold for well-targeted experiments (Magliacane et al., 2018). The following is a detailed comparison.

(1) Modeling assumption and problem setup. The two problem settings do not strictly include one another, and they are of complementing relations where ours corresponds to the intervention-free case and Magliacane et al. (2018) corresponds to the intervention case. If we try to express the problem setting of Magliacane et al. (2018) within our formulation, we would be expressing the interventions as alterations to the SEMs. We assume that such alterations do not occur in our setting since our focus is on observational data; therefore, the problem formulation of Magliacane et al. (2018) is not a subset of ours. On the other hand, if we try to express our problem setting within the formulation of Magliacane et al. (2018), our problem setup would only have C_1 as the context variable, and C_1 would be a parent of all observed variables, e.g., C_1 switches the distribution of S by switching different quantile functions to perform inverse transform. This potentially allows the existence of the effect $C_1 \to Y$ and diverges from Assumption 2 (iii) in Magliacane et al. (2018). Also, even if such an edge does not exist, it is acceptable that there are no separating sets (in the extreme case) if Y is a parent of all X_i's. In this case, conditioning on any of the X_i's would result in making C_1 and Y dependent. From this consideration, our problem setting is not a subset of that of Magliacane et al. (2018), either.

(2) Plausible applications. The problem setup of Magliacane et al. (2018) is suitable especially for applications in which various experiments are conducted such as genomics (Magliacane et al., 2018), whereas our problem setting may be more suitable for some fields with observational data such as health record analysis or economics.

(3) Methodology. Our proposed method actually estimates the SEMs (though in the reduced-form) and exploits the estimated SEMs in the domain adaptation algorithm. In fact, directly using the estimated SEMs as a tool to realize domain adaptation can be seen as the first attempt to fully leverage the structural causal models in the DA algorithm. On the other hand, Magliacane et al. (2018) approaches the problem of domain adaptation via variable selection to find a subset so that the conditional distribution is invariant.

E.2. Comparison with Gong et al. (2018)

In the present paper, we assumed an invariance of structural equations between domains. Here, we clarify the difference from a related but different assumption considered by Causal Generative Domain Adaptation Network (CG-DAN; Gong

et al., 2018).

(1) Problem setup. Gong et al. (2018) presumes the *anticausal* scenario (i.e., Y is the cause of X) and that X given Y follows a structural equation model, whereas our paper considers more general SEMs of X and Y.

(2) Theoretical justification. The approach of Gong et al. (2018) does not have a theoretical guarantee in terms of the identifiability of f, i.e., there has been no known theoretical condition under which the learned generator is applicable across different domains. On the other hand, our method enjoys a strong theoretical justification of nonlinear ICA including the identifiability of f under known theoretical conditions.

(3) Methodology. The method of Gong et al. (2018) estimates the GCM of X given Y using source domain data and uses it to design a generator neural network. On the other hand, we more directly exploit the estimated reduced-form SEM in the method.

E.3. Comparison with Arjovsky et al. (2020)

Arjovsky et al. (2020) proposed *invariant risk minimization* (IRM) for the *out-of-distribution (OOD) generalization* problem. The IRM approach tries to learn a feature extractor that makes the optimal predictor invariant across domains, and its theoretical validity is argued based on SCMs. Here, we compare it with the present work in terms of the problem setup, theoretical justification, and the methodology.

(1) Basic assumption and problem setup. The OOD generalization problem tackled in Arjovsky et al. (2020) assumes no access to the target domain data. In this respect, the problem is different and intrinsically more difficult than the one considered in this paper, where a small labeled sample from the target domain is assumed to be available. In order to solve the OOD generalization problem, in a nutshell, Arjovsky et al. (2020) essentially assumes the existence of a feature extractor that *elicits an invariant predictor*, i.e., one that makes the optimal predictors of the different domains to be identical after the feature transformation. This can be seen as a variant of the representation learning approach for domain adaptation where we assume there exists $\mathcal{T}$ such that $p(Y|\mathcal{T}(X))$ is invariant across domains. Indeed, for example, when the loss function is the cross-entropy, the condition corresponds to the invariance of $P(Y|\mathcal{T}(X))$ across domains (Arjovsky et al., 2020). More technically, in addition, (Arjovsky et al., 2020, Definition 7(ii)) requires the condition $\mathbb{E}_1[Y|\mathrm{Pa}(Y)] = \mathbb{E}_2[Y|\mathrm{Pa}(Y)]$, which can be violated when the latent factors corresponding to Y have different distributions across domains. On the other hand, our assumption can be seen as the existence of a feature extractor that can simultaneously estimate the independent components in all domains, which does not necessarily imply the existence of a common feature transformer that induces a unique optimal predictor.

(2) Theoretical justification. Arjovsky et al. (2020) formulated a condition under which the IRM principle leads to an appropriate predictor for OOD generalization, but only under a certain linearity assumption which is essentially a relaxation of linear SEMs. Furthermore, in the theoretical guarantee, the feature extractor is restricted to be linear. In addition, Arjovsky et al. (2020) only provides the population-level analysis that the solution of the IRM objective formulated using the underlying distributions enjoys OOD generalization, and it does not discuss the condition under which the ideal feature extractor can be properly estimated by the empirical IRM. The requirement for the strong assumption of linearity likely stems from the intrinsic difficulty of the OOD problem in Arjovsky et al. (2020), namely, its formulation does not assume specific types of interventions. On the other hand, our method enjoys a stronger theoretical guarantee of an excess risk bound without such parametric assumptions on the models or the data generating process, by focusing on the case that the causal mechanisms are indifferent across the domains.

(3) Methodology. The methodology of IRM estimates a single predictor that generalizes well to all domains by finding a feature extractor that makes the predictor optimal in all domains. The approach shares the same spirit as the representation learning approaches to domain adaptation, which try to find a feature extractor that induces invariant conditional distributions, such as transfer component analysis (Pan et al., 2011). On the other hand, our method estimates the SEMs (in the reduced-form) and exploits it to make the training on the few target domain data more efficient through data augmentation.