Minimax Analysis for Inverse Risk in Nonparametric Invertible Regression

(joint work with M. Imaizumi, arXiv:2112.00213)

Akifumi Okuno

Institute of Statistical Mathematics and RIKEN AIP

Sep. 2022

Summary of This Talk

This study focuses on **invertibility** of the function.
We estimate *invertible* regression function $\hat{\mathbf{f}}_n : [-1,1]^d \to [-1,1]^d$ and evaluate *invertible risk*

$$\mathrm{R}_{\mathrm{INV}}(\hat{\mathbf{f}}_n, \mathbf{f}_*) := \|\hat{\mathbf{f}}_n - \mathbf{f}_*\|^2_{L^2(P_X)} + \psi(\|\hat{\mathbf{f}}_n^{\dagger} - \mathbf{f}_*^{-1}\|_{L^2(P_X)}).$$

Our contribution ($d = 2$; planer invertible regression; OI2021)

With $\psi(z) = z^4$,

$$\inf_{\bar{\mathbf{f}}_n} \sup_{\mathbf{f}_* \in \mathcal{F}^{\mathrm{Lip}}_{\mathrm{Inv}}} \mathrm{R}_{\mathrm{INV}}(\hat{\mathbf{f}}_n, \mathbf{f}_*) \asymp n^{-2/(2+d)}$$

up to logarithmic factors, **same as the (standard) Lipschitz function estimation!**

- We can employ this minimax rate as **a baseline of efficiency!**
- Generalized to $d \in \mathbb{N}, \psi(z) = z^2$ by assuming C^2 in OI (in prep.)

Background

Invertible Functions

Let $I = [-1, 1]$. A function $\mathbf{f} : I^d \to I^d$ is *invertible* iff

$$\mathbf{f}^{-1}(\mathbf{y}) := \{\mathbf{x} \in I^d \mid \mathbf{f}(\mathbf{x}) = \mathbf{y}\}$$

is a unique point, for any $\mathbf{y} \in I^d$. We consider Lipschitz invertible functions $\mathbf{f} \in \mathcal{F}_{\mathrm{Inv}}^{\mathrm{Lip}}$.

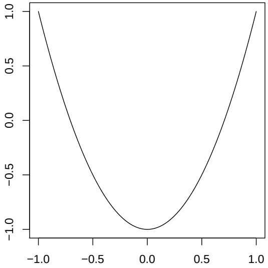


Figure: Non-Invertible $f(x) = 2x^2 - 1$

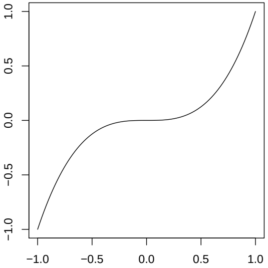
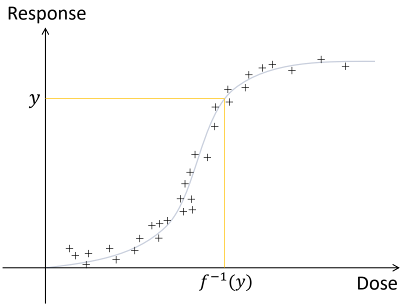


Figure: *Invertible* $f(x) = x^3$

Invertible Function Estimation ($d = 1$)

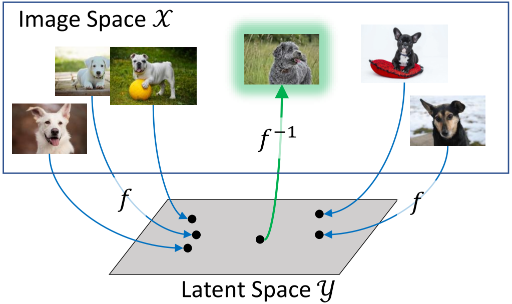
Invertiblility $=$ (Strict) **Monotonicity**



- Many papers on application/theory of monotone func. estimation in econ/stats.
  - Nonparametric statistical calibration (e.g., Tang et al., 2011, 2015)
  - Nonparametric instrumental variable regression (e.g., Krief, 2017)

Invertible Function Estimation ($d \in \mathbb{N}$)

Invertiblility = **One-to-one correspondence**



Usually, it is quite difficult to define *invertibile* and *expressive* estimator for $d \geq 2$.
Recent way: Invertible Neural Network = *Normalizing Flow* (Dinh et al., 2014).

Types of Normalizing Flows and Universality

There are various types of normalizing flows (NF), where they are basically in the form of

$$\mathbf{f}(\mathbf{x}) = (\boldsymbol{\phi}_1 \circ \boldsymbol{\psi}_1 \circ \boldsymbol{\phi}_2 \circ \boldsymbol{\psi}_2 \cdots \circ \boldsymbol{\psi}_{L-1} \circ \boldsymbol{\phi}_L)(\mathbf{x})$$

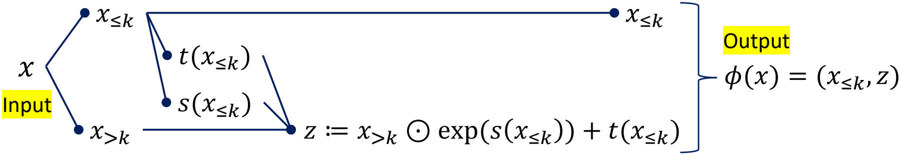
with invertible $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \cdots, \boldsymbol{\phi}_L : \mathbb{R}^d \to \mathbb{R}^d$ and Affine mappings $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \cdots, \boldsymbol{\psi}_{L-1}$.

(i) **Simple ones:** Non-universal
- Planar flow $\boldsymbol{\phi}_j(\mathbf{x}) = \mathbf{x} + \mathbf{a}_j \mathbf{h}(\mathbf{B}_j^\top \mathbf{x} + \mathbf{c}_j)$,
- Householder flow $\boldsymbol{\phi}_j(\mathbf{x}) = \mathbf{x} - 2\mathbf{v}_j \mathbf{v}_j^\top \mathbf{x}$, etc...

(ii) **Triangular map-based:** Universal (in the sense of distribution matching)
- Sum-of-Squares (SoS; Huang et al., 2018),
- Neural Autoregressive (NAF; Huang et al., 2018), etc...

(iii) **Real NVP:** Universal (in the usual sense)
- Affine-coupling flow (ACF; Dinh et al., 2014) $\boldsymbol{\phi}_j(\mathbf{x}) = (\mathbf{x}_{\le k}, \mathbf{x}_{>k} \odot \exp(\mathbf{s}_j(\mathbf{x}_{\le k})) + \mathbf{t}_j(\mathbf{x}_{\le k}))$ equipped with NNs $\mathbf{s}_j, \mathbf{t}_j : \mathbb{R}^k \to \mathbb{R}^{d-k}$ and $k \in [d]$.

Affine-Coupling Flow (ACF)



- **ACF is invertible**: $\mathbf{f}^{-1}(\mathbf{y}) = (\boldsymbol{\phi}_L^{-1} \circ \boldsymbol{\psi}_{L-1}^{-1} \cdots \circ \boldsymbol{\psi}_2^{-1} \circ \boldsymbol{\phi}_2^{-1} \circ \boldsymbol{\psi}_1^{-1} \circ \boldsymbol{\phi}_1^{-1})(\mathbf{y})$ with

$$\boldsymbol{\phi}_j^{-1}(\mathbf{y}) = \left(\mathbf{y}_{\le k}, \frac{\mathbf{y}_{>k} - \mathbf{t}_j(\mathbf{y}_{\le k})}{\exp(\mathbf{s}_j(\mathbf{y}_{\le k}))}\right).$$

- With increasing number of layers $L \to \infty$,
  ACF universally approximates C^2 invertible functions (Teshima et al., 2020).

Still difficult to evaluate the *efficiency, for* $d \geq 2$.

- Teshima et al. (2020) assumes $L \to \infty$.
- Even the (simple) minimax optimal convergence rate is not obtained.

- $d = 1$ is OK: monotonicity is easy enough to handle. $\exists$Many studies.
- $d \geq 2$ is very difficult: monotonicity is no longer available.
  Even the characterization of the invertible function is not known:
  nonparametric estimator (for theory) is not known.

There is a *HUGE* gap from $d = 1$ to $d \geq 2$:
we evaluate the efficiency for $d = 2$.

Conventional Theory and Our Problem Setup: Inverse Risk

Regression Problem

$$\mathcal{F}_{\mathrm{Inv}} := \{\mathbf{f} : I^2 \to I^2 \mid \forall \mathbf{y} \in I^2,\ !\exists \mathbf{x} \in I^2 \text{ such that } \mathbf{f}(\mathbf{x}) = \mathbf{y}\} \quad (I := [-1, 1]),$$
$$\mathcal{F}_{\mathrm{Inv}}^{\mathrm{Lip}} := \{\mathbf{f} \in \mathcal{F}_{\mathrm{Inv}} \mid \mathbf{f}, \mathbf{f}^{-1} \text{ are Lipschitz}\}.$$

Assume we have observations $\mathfrak{D}_n := \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n \subset I^2 \times \mathbb{R}^2$ that independently follow

$$\mathbf{Y}_i = \mathbf{f}_*(\mathbf{X}_i) + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} N_2(\mathbf{0}, \sigma^2 \mathbf{I}_2), \quad i = 1, 2, \ldots, n,$$

for a true function $\mathbf{f}_* \in \mathcal{F}_{\mathrm{Inv}}^{\mathrm{Lip}}$ and $\sigma^2 > 0$.

- $\hat{\mathbf{f}}_n$ estimates $\mathbf{f}_*$, using the observations $\mathfrak{D}_n$.
- Note: $d = 2$ is assumed throughout this talk.

Consistency

Definition (Risk)

For any estimator $\bar{\mathbf{f}}_n$, we define a L^2-risk:

$$\mathrm{R}(\bar{\mathbf{f}}_n, \mathbf{f}_*) := |||\bar{\mathbf{f}}_n - \mathbf{f}_*|||_{L^2(P_X)}^2,$$

where $|||\mathbf{f}|||_{L^2(P_{\mathbf{X}})} := (\sum_{j=1}^2 \int |f_j|^2 \mathrm{d}P_{\mathbf{X}})^{1/2}$ is an L^2-norm.

Definition (Consistency)

A estimator $\bar{\mathbf{f}}_n$ is *consistent* if

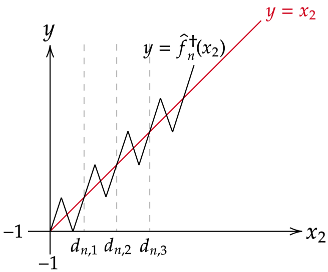
$$\mathbb{P}(\mathrm{R}(\bar{\mathbf{f}}_n, \mathbf{f}_*) \leq C r_n) \geq 1 - \delta_n$$

holds for some $C \in (0, \infty)$ and decreasing sequences $r_n, \delta_n \searrow 0$. r_n is also called *convergence rate*.

Kernel smoother is consistent with $r_n = n^{-2/(2+d)}$, for Lipschitz $\mathbf{f}_*$.

Consistency $\neq$ Invertibility: An Example

$$\mathbf{f}_*(\mathbf{x}) = \mathbf{x}, \quad \hat{\mathbf{f}}_n(\mathbf{x}) = (x_1, \hat{f}_n^\dagger(x_2)),$$



With $d_{n,j} = -1 + j\gamma_n$, $\hat{\mathbf{f}}_n^\dagger$ is consistent with the (arbitrarily fast) rate γ_n, whereas it is *NOT* invertible over entire $I^2 = [-1, 1]^2$.

Inverse Risk Measures both Consistency and Invertibility

Definition (Empirical inverse function)

Let $\mathbf{c} \in \mathbb{R}^2 \setminus I^2$ be a constant vector. An inverse function for the estimator $\bar{\mathbf{f}}_n : I^2 \to I^2$ is:

$$\bar{\mathbf{f}}_n^{\dagger}(\mathbf{y}) := \begin{cases} \mathbf{x} & (\text{if } !\exists \mathbf{x} \in I^2 \text{ such that } \bar{\mathbf{f}}_n(\mathbf{x}) = \mathbf{y}) \\ \mathbf{c} & (\text{otherwise}) \end{cases}, \quad \forall \mathbf{y} \in I^2.$$

Definition (Inverse risk)

$$\mathrm{R}_{\mathrm{INV}}(\bar{\mathbf{f}}_n, \mathbf{f}_*) := \mathrm{R}(\bar{\mathbf{f}}_n, \mathbf{f}_*) + \psi(\mathrm{R}(\bar{\mathbf{f}}_n^{\dagger}, \mathbf{f}_*^{-1})), \quad \text{for } \bar{\mathbf{f}}_n : I^2 \to I^2.$$

- Inverse risk measures both invertibility (a.e.) and consistency (for both $\bar{\mathbf{f}}_n, \bar{\mathbf{f}}_n^{\dagger}$).
- The previous approximation example: $\mathrm{R}(\bar{\mathbf{f}}_n, \mathbf{f}_*) \to^p 0$, $\mathrm{R}_{\mathrm{INV}}(\bar{\mathbf{f}}_n, \mathbf{f}_*) > \exists c > 0$.

Level-Set Representation

Level-Set Representation

Definition (Level-Set Representation)

For $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x})) \in \mathcal{F}_{\mathrm{Inv}}$, we define a level-set $L_{f_j}(y_j) := \{\mathbf{x} \in I^2 \mid f_j(\mathbf{x}) = y_j\}$ and the level-set representation

$$\mathbf{f}^{-1}(\mathbf{y}) = \{\mathbf{x} \in I^2 \mid \mathbf{f}(\mathbf{x}) = \mathbf{y}\} = L_{f_1}(y_1) \cap L_{f_2}(y_2), \quad \forall \mathbf{y} = (y_1, y_2) \in I^2.$$

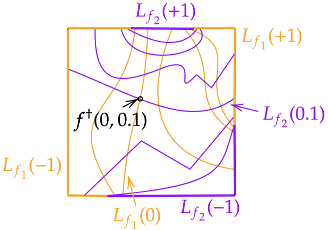


Figure: $\mathbf{f}^{-1}(0, 0.1) = L_{f_1}(0) \cap L_{f_2}(0.1)$

- Example: for $\mathbf{f}(\mathbf{x}) = \mathbf{x}$, $L_{f_1}(y_1) = (y_1, I)$, $L_{f_2}(y_2) = (I, y_2)$.

An Real Example

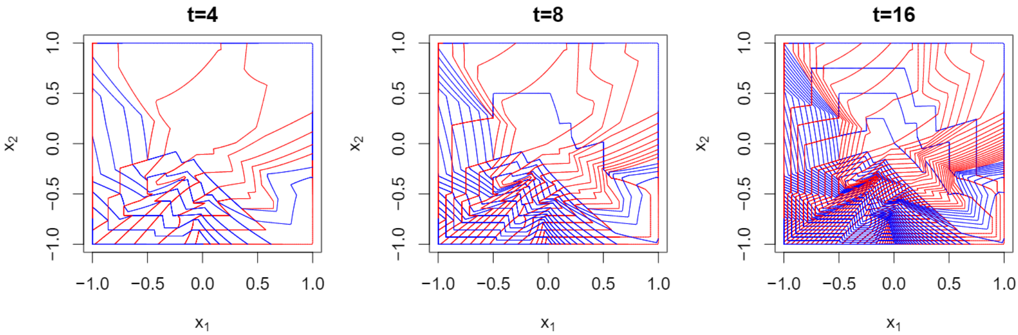


Figure: $\{L_{f_j}(\pm k/t)\}_{k=0,1,2,\ldots,t}$ (red for $j=1$, blue for $j=2$).

Level-Set Properties (in Theory)

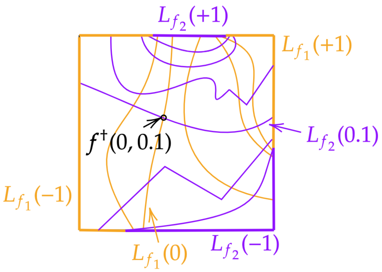


Figure: $\mathbf{f}^{-1}(0, 0.1) = L_{f_1}(0) \cap L_{f_2}(0.1)$

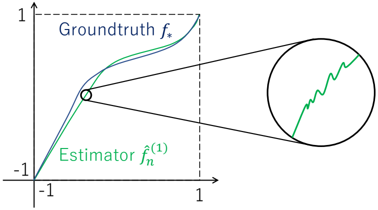
For any $\mathbf{f} = (f_1, f_2) \in \mathcal{F}^{\mathrm{Lip}}_{\mathrm{Inv}}$,

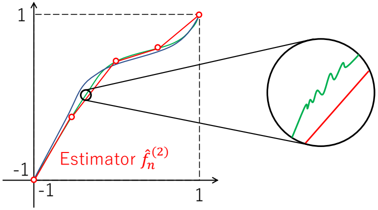
- $L_{f_1}(y_1) = \cup_{\alpha \in I} \mathbf{f}^{-1}(y_1, \alpha)$ and $L_{f_1}(y_2) = \cup_{\alpha \in I} \mathbf{f}^{-1}(\alpha, y_2)$.
- $d_{\mathrm{Hausdorff}}(L_{f_j}(y), L_{f_j}(y')) \le \exists C|y - y'|, \forall y, y' \in I, j = 1, 2.$
- $L_{f_j}(\pm 1) \subset \partial I^2, j = 1, 2.$ (more specifically, $\mathbf{f}(\partial I^2) = \partial I^2 = \mathbf{f}^{-1}(\partial I^2)$)

Proposed (Asymptotically A.E.) Invertible Estimator

Basic Idea: Two-Step Estimation

Example: in the case $d = 1$.

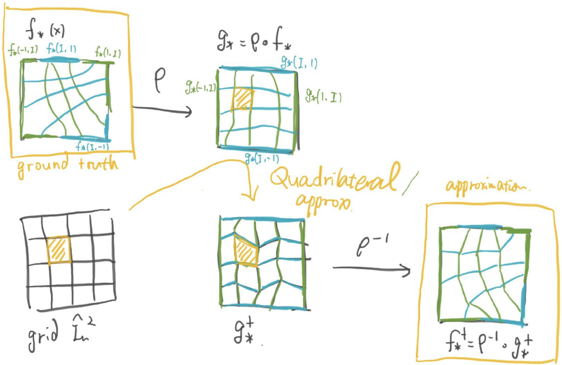




1. Compute $\hat{\mathbf{f}}_n^{(1)}$ over the grid
2. Interpolate them using the *line* (as the second-step estimator $\hat{\mathbf{f}}_n^{(2)}$).

Planer Invertible Regression ($d = 2$)

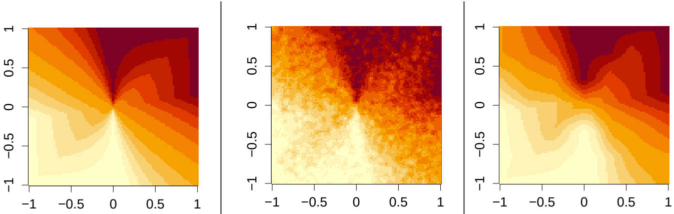
Level set representation of $\mathbf{f}_*^{-1}$ yields $\mathbf{f}_*(\mathbf{x}) = (\mathbf{f}_*^{-1})^{-1}(\mathbf{x}) = \mathbf{f}_*(x_1, I) \cap \mathbf{f}_*(I, x_2)$.



1. Compute $\hat{\mathbf{f}}_n^{(1)}$ over the grid
2. Interpolate them using the *quadrilateral* (as the second-step estimator $\hat{\mathbf{f}}_n^{(2)}$).

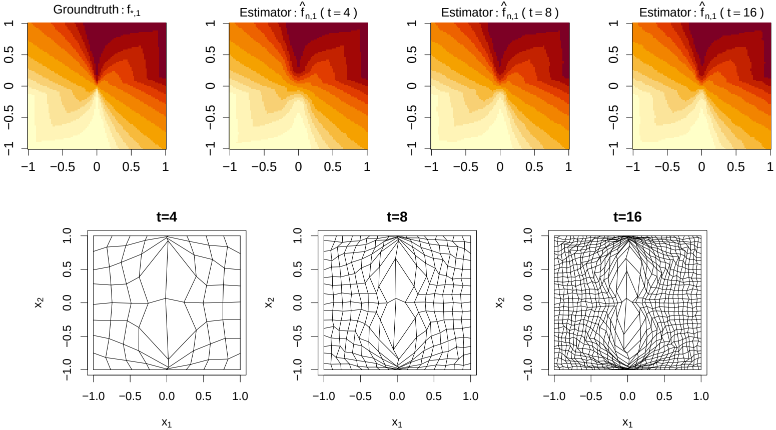
Numerical Experiments: Approximation

- $n = 10^4, \sigma^2 = 10^{-1}$(larger noise), $\mathbf{X}_i \sim U(I^2)$.
- $t = 3$.

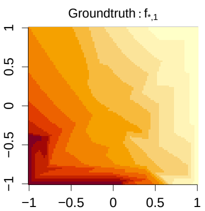


(a) $f_{*,1}$ (b) $\hat{f}_{n,1}^{(1)}$ (c) $\hat{f}_{n,1}^{(2)}$

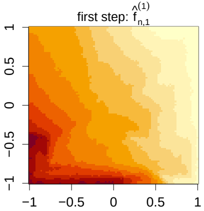
Numerical Experiments: Number of Splits $t = t_n$



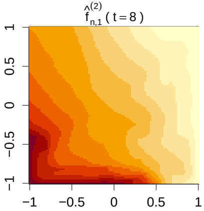
Numerical Experiments: Other Functions



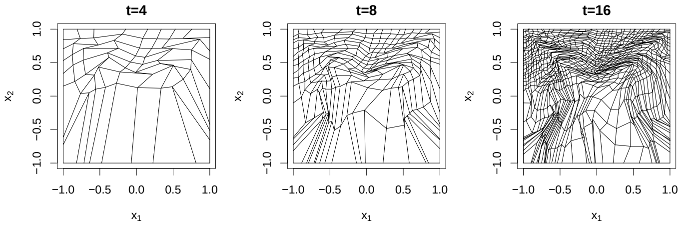
(a) $f_{*,1}$



(b) $\hat{f}^{(1)}_{n,1}$



(c) $\hat{f}^{(2)}_{n,1}$



Lower/Upper Bound Analysis

Lower Bound Analysis

Let $d = 2, \psi(z) = z^4$.

Theorem (Lower Bound)

$$C_* n^{-2/(2+d)} \le \inf_{\hat{\mathbf{f}}_n} \sup_{\mathbf{f}_* \in \mathcal{F}_{Inv}^{Lip}} \mathrm{R}_{INV}(\hat{\mathbf{f}}_n, \mathbf{f}_*)$$

with probability larger than $1/2$.

Theorem (Upper Bound)

$$\inf_{\hat{\mathbf{f}}_n} \sup_{\mathbf{f}_* \in \mathcal{F}_{Inv}^{Lip}} \mathrm{R}_{INV}(\hat{\mathbf{f}}_n, \mathbf{f}_*) \le \bar{C} n^{-2/(2+d)} (\log n)^{2\alpha'}$$

w.p. $1 - \delta_n$ ($\nearrow 1$), *for any* $\alpha' > 0$.

See OI (2021) for details.

Ongoing Work and Conclusion

Ongoing Work

- Generalization to $d \in \mathbb{N}$ (OI, in prep.) by assuming C^q-smoothness ($q \geq 2$).

Theorem

Let $d \in \mathbb{N}$. There exists $\bar{C} \in (0, \infty)$ such that,

$$\inf_{\hat{\mathbf{f}}_n} \sup_{\mathbf{f}_* \in \mathcal{F}^q_{Inv}} \tilde{R}_{INV}(\hat{\mathbf{f}}_n, \mathbf{f}_*) \leq \bar{C} n^{-2q/(2q+d)} (\log n)^{2\alpha'} \quad w.p.a.l.\ 1 - \delta_n$$

Table: Studies on minimax optimality of the estimation of invertible functions $\mathbf{f} \in C^q([-1, 1]^d)$.

	$d = 1$	$d = 2$	$d = 3, 4, 5, 6, \ldots$
$q < 1$		$\times$	$\times$
Lipschitz (nearly $q = 1$)	Existing	OI (2021)	$\times$
$1 < q < 2$		$\times$	$\times$
$2 \leq q$		**OI (in prep.)**	

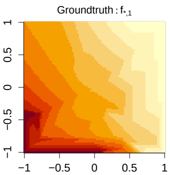
Conclusion

- We proved for $d = 2$ that

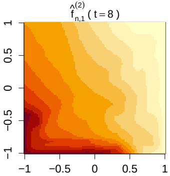
$$\inf_{\hat{\mathbf{f}}_n} \sup_{\mathbf{f}_* \in \mathcal{F}_{\mathrm{Inv}}^{\mathrm{Lip}}} \mathrm{R}_{\mathrm{INV}}(\hat{\mathbf{f}}_n, \mathbf{f}_*) \asymp n^{-2/(2+d)}$$

in probability, up to logarithmic factors.

- We proposed a minimax optimal (whereby asymptotically a.e. invertible) estimator $\hat{\mathbf{f}}_n$.



(a) Groundtruth



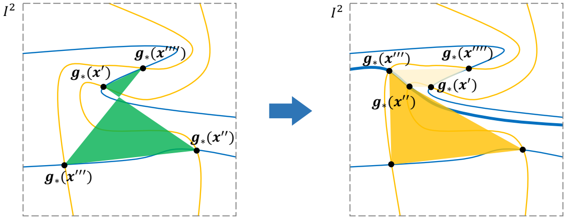
(b) Estimator

https://arxiv.org/abs/2112.00213

Some Remarks

Problem: Quadrilateral Approximation and Twists

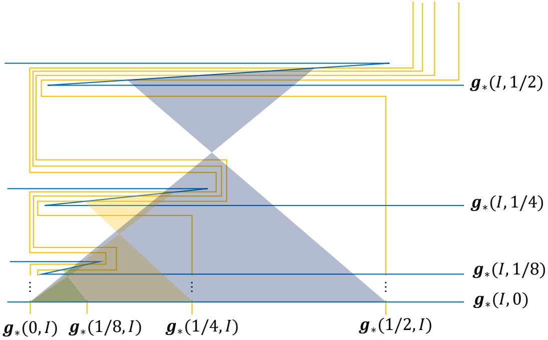
- If $L_{\mathbf{g}_*} \in [1, 2^{1/4})$, no twist appears when approximating quadrilaterals.
- Otherwise, there can be twists.



Each twist vanishes by increasing the number of division (for most suitable cases).
Daneri and Pratelli (2014) Proposition 4.1 proves

$$\mathcal{L}(\text{twisted region}) \to^p 0.$$

Pathological Example



Whereas each twist is decomposed into smaller quadrilaterals (by increasing $t = t_n$), twists can appear indefinitely in some pathological examples. (They are ignored in the sense of Lebesgue measure, in our theory)

Which is better to assume: Lipschitz or C^2?

- **Nonparametric statistics** usually assumes that $\mathbf{f}_*$ is Lipschitz:
  - 👍 Less restrictive
  - 👎 Includes **pathological** examples

  This study assumes Lipschitz (with $d = 2$): as we are researchers of statistics...
  Almost impossible to extend to general $d \geq 3$.

- **Geometry** usually assumes that $\mathbf{f}_*$ is C^2:
  - 👍 Theoretically tractable (tangent space can be defined)
  - 👎 More restrictive

  Our ongoing work assumes C^2 (and generalize to $d \in \mathbb{N}$).

References I

Daneri, S. and Pratelli, A. (2014). Smooth approximation of bi-lipschitz orientation-preserving homeomorphisms. *Annales de l'IHP Analyse non linéaire*, 31(3):567–589.

Dinh, L., Krueger, D., and Bengio, Y. (2014). NICE: Non-Linear Independent Components Estimation. *arXiv preprint arXiv:1410.8516*.

Donaldson, S. K. and Sullivan, D. P. (1989). Quasiconformal 4-manifolds. *Acta Mathematica*, 163(none):181 – 252.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR.

Kobyzev, I., Prince, S., and Brubaker, M. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979.

References II

Kong, Z. and Chaudhuri, K. (2020). The expressive power of a class of normalizing flow models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3599–3609. PMLR.

Krief, J. M. (2017). Direct instrumental nonparametric estimation of inverse regression functions. *Journal of Econometrics*, 201(1):95–107.

Tang, R., Banerjee, M., and Michailidis, G. (2011). A two-stage hybrid procedure for estimating an inverse regression function. *The Annals of Statistics*, 39(2):956–989.

Tang, R., Banerjee, M., Michailidis, G., and Mankad, S. (2015). Two-stage plans for estimating the inverse of a monotone function. *Technometrics*, 57(3):395–407.

Teshima, T., Ishikawa, I., Tojo, K., Oono, K., Ikeda, M., and Sugiyama, M. (2020). Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators. In *Advances in Neural Information Processing Systems*, volume 33, pages 3362–3373. Curran Associates, Inc.

References III

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.