Some Applications of Mixed Volumes in Data Science

Eliza O'Reilly

Caltech

Application #1: Prediction with Random Tessellation Forests

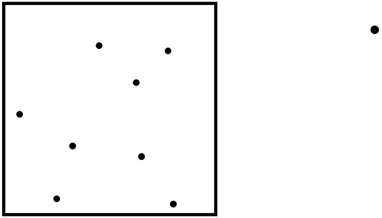
Joint work with Ngoc Mai Tran (UT Austin)

Goal: Given example input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, obtain estimator $\hat{f}_n$ to **predict** output from new input x: $\hat{y} = \hat{f}_n(x)$

Goal: Given example input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, obtain estimator $\hat{f}_n$ to **predict** output from new input x: $\hat{y} = \hat{f}_n(x)$

Randomized Decision Trees:

- Recursively split data along random feature of input
- Induce a hierarchical axis-aligned partition of input space



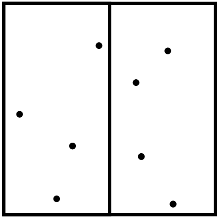
- **Random tree (regression) estimator**: average over cell/leaf

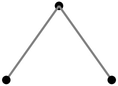
$$\hat{f}_n(x) = \frac{\sum_{i=1}^n y_i 1_{\{x_i \text{ in same cell as } x\}}}{\sum_{i=1}^n 1_{\{x_i \text{ in same cell as } x\}}}$$

Goal: Given example input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, obtain estimator $\hat{f}_n$ to **predict** output from new input x: $\hat{y} = \hat{f}_n(x)$

Randomized Decision Trees:

- Recursively split data along random feature of input
- Induce a hierarchical axis-aligned partition of input space





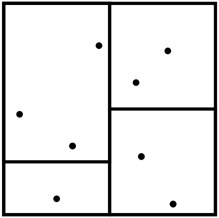
- **Random tree (regression) estimator**: average over cell/leaf

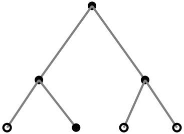
$$\hat{f}_n(x) = \frac{\sum_{i=1}^n y_i \mathbb{1}_{\{x_i \text{ in same cell as } x\}}}{\sum_{i=1}^n \mathbb{1}_{\{x_i \text{ in same cell as } x\}}}$$

Goal: Given example input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, obtain estimator $\hat{f}_n$ to **predict** output from new input x: $\hat{y} = \hat{f}_n(x)$

Randomized Decision Trees:

- Recursively split data along random feature of input
- Induce a hierarchical axis-aligned partition of input space





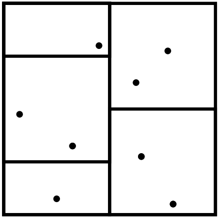
- **Random tree (regression) estimator**: average over cell/leaf

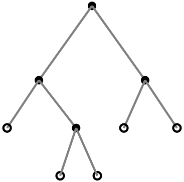
$$\hat{f}_n(x) = \frac{\sum_{i=1}^n y_i 1_{\{x_i \text{ in same cell as } x\}}}{\sum_{i=1}^n 1_{\{x_i \text{ in same cell as } x\}}}$$

Goal: Given example input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, obtain estimator $\hat{f}_n$ to **predict** output from new input x: $\hat{y} = \hat{f}_n(x)$

Randomized Decision Trees:

- Recursively split data along random feature of input
- Induce a hierarchical axis-aligned partition of input space





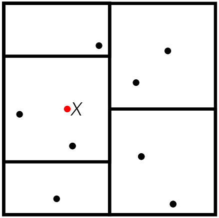
- **Random tree (regression) estimator**: average over cell/leaf

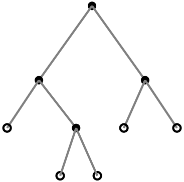
$$\hat{f}_n(x) = \frac{\sum_{i=1}^n y_i 1_{\{x_i \text{ in same cell as } x\}}}{\sum_{i=1}^n 1_{\{x_i \text{ in same cell as } x\}}}$$

Goal: Given example input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, obtain estimator $\hat{f}_n$ to **predict** output from new input x: $\hat{y} = \hat{f}_n(x)$

Randomized Decision Trees:

- Recursively split data along random feature of input
- Induce a hierarchical axis-aligned partition of input space





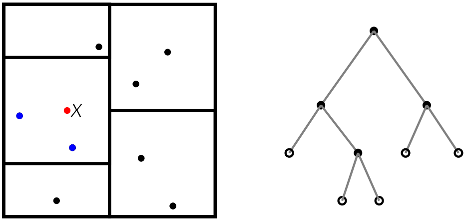
- **Random tree (regression) estimator**: average over cell/leaf

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n y_i 1_{\{x_i \text{ in same cell as } x\}}}{\sum_{i=1}^n 1_{\{x_i \text{ in same cell as } x\}}}$$

Goal: Given example input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, obtain estimator $\hat{f}_n$ to **predict** output from new input x: $\hat{y} = \hat{f}_n(x)$

Randomized Decision Trees:

- Recursively split data along random feature of input
- Induce a hierarchical axis-aligned partition of input space

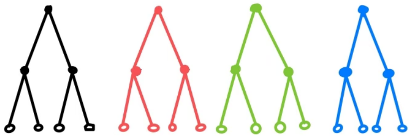


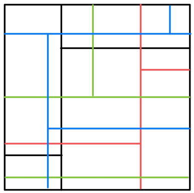
- **Random tree (regression) estimator**: average over cell/leaf

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n y_i 1_{\{x_i \text{ in same cell as } x\}}}{\sum_{i=1}^n 1_{\{x_i \text{ in same cell as } x\}}}$$

Random Forest (RF) Estimator

- Average of M i.i.d. tree estimators

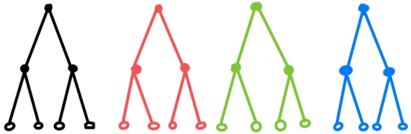


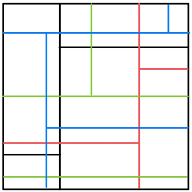


[1][Ho, 1995; 1998; Amit and Geman, 1997; Breiman, 2001]
[2][Caruana and Niculescu-Mizil, 2006; Fernandez-Delgado et al., 2014]
[3][Scornet et al., 2015; Wager and Athey, 2018; Chi et al., 2020; Klusowski and Tian, 2022]
[4][Breiman, 2004; Genuer, 2012]

Random Forest (RF) Estimator

- Average of M i.i.d. tree estimators



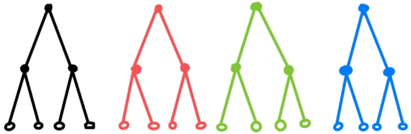


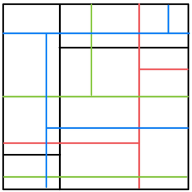
- Original RF algorithm[1]: splits *dependent* on data
- State-of-the-art empirical performance for many tasks[2]

[1] [Ho, 1995; 1998; Amit and Geman, 1997; Breiman, 2001]
[2] [Caruana and Niculescu-Mizil, 2006; Fernandez-Delgado et al., 2014]
[3] [Scornet et al., 2015; Wager and Athey, 2018; Chi et al., 2020; Klusowski and Tian, 2022]
[4] [Breiman, 2004; Genuer, 2012]

Random Forest (RF) Estimator

- Average of M i.i.d. tree estimators





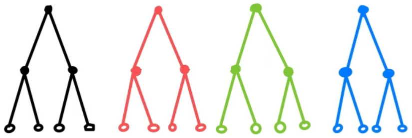
- Original RF algorithm[1]: splits *dependent* on data
- State-of-the-art empirical performance for many tasks[2]
- Difficult to analyze; still a lack of theoretical understanding[3]

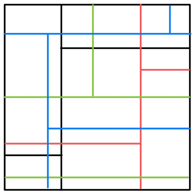
---

[1] [Ho, 1995; 1998; Amit and Geman, 1997; Breiman, 2001]
[2] [Caruana and Niculescu-Mizil, 2006; Fernandez-Delgado et al., 2014]
[3] [Scornet et al., 2015; Wager and Athey, 2018; Chi et al., 2020; Klusowski and Tian, 2022]
[4] [Breiman, 2004; Genuer, 2012]

Random Forest (RF) Estimator

- Average of M i.i.d. tree estimators



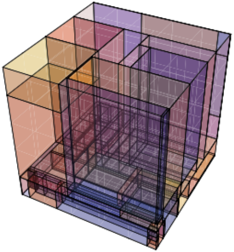


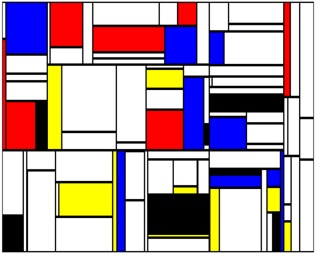
  - Original RF algorithm[1]: splits *dependent* on data
  - State-of-the-art empirical performance for many tasks[2]
  - Difficult to analyze; still a lack of theoretical understanding[3]
  - Purely RF variants[4]: splits *independent* of data

[1] [Ho, 1995; 1998; Amit and Geman, 1997; Breiman, 2001]
[2] [Caruana and Niculescu-Mizil, 2006; Fernandez-Delgado et al., 2014]
[3] [Scornet et al., 2015; Wager and Athey, 2018; Chi et al., 2020; Klusowski and Tian, 2022]
[4] [Breiman, 2004; Genuer, 2012]

Mondrian process

- Introduced by Roy and Teh in 2008
- Stochastic process that recursively builds an axis-aligned hierarchical partition in $\mathbb{R}^d$





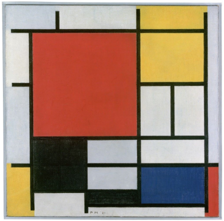


Figure: Piet Mondrian (1921).

Mondrian process construction in $\mathbb{R}^d$

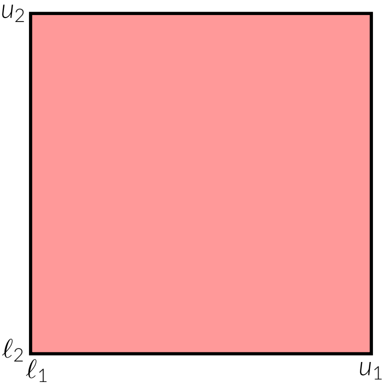
1. Fix **lifetime parameter** $\lambda > 0$

2. Draw

$$\Delta \sim Exp\left[\sum_{i=1}^{d}(u_i - \ell_i)\right]$$

3. IF $\Delta > \lambda$ stop,
   ELSE sample a split:
   - *Dimension*: j with probability proportional to $u_j - \ell_j$
   - *Location*: uniform on $[\ell_j, u_j]$

4. Recurse independently on each subrectangle with lifetime $\lambda - \triangle$

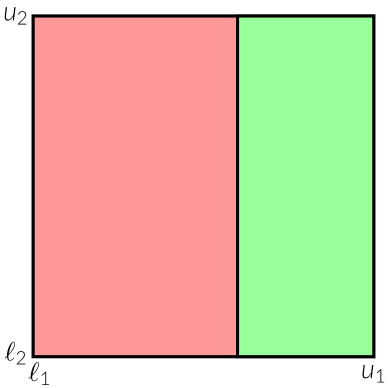


Mondrian process construction in $\mathbb{R}^d$

1. Fix **lifetime parameter** $\lambda > 0$
2. Draw

$$\Delta \sim Exp\left[\sum_{i=1}^{d}(u_i - \ell_i)\right]$$

3. IF $\Delta > \lambda$ stop,
   ELSE sample a split:
   - *Dimension*: j with probability proportional to $u_j - \ell_j$
   - *Location*: uniform on $[\ell_j, u_j]$
4. Recurse independently on each subrectangle with lifetime $\lambda - \triangle$

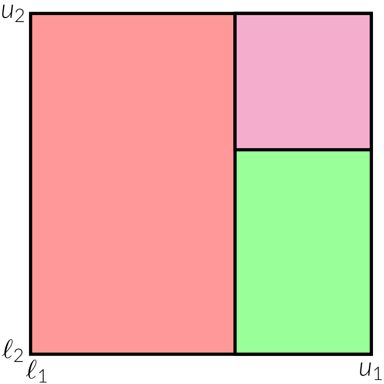


Mondrian process construction in $\mathbb{R}^d$

1. Fix **lifetime parameter** $\lambda > 0$
2. Draw

$$\Delta \sim Exp\left[\sum_{i=1}^{d}(u_i - \ell_i)\right]$$

3. IF $\Delta > \lambda$ stop,
   ELSE sample a split:
   - *Dimension*: j with probability proportional to $u_j - \ell_j$
   - *Location*: uniform on $[\ell_j, u_j]$
4. Recurse independently on each subrectangle with lifetime $\lambda - \triangle$



Mondrian process construction in $\mathbb{R}^d$

1. Fix **lifetime parameter** $\lambda > 0$

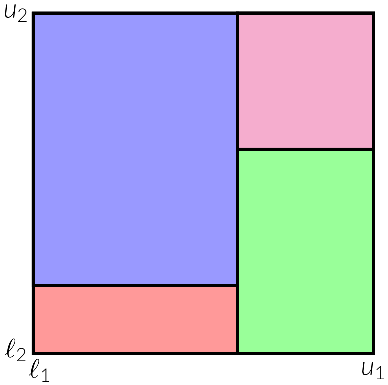
2. Draw

$$\Delta \sim Exp\left[\sum_{i=1}^{d}(u_i - \ell_i)\right]$$

3. IF $\Delta > \lambda$ stop,
   ELSE sample a split:

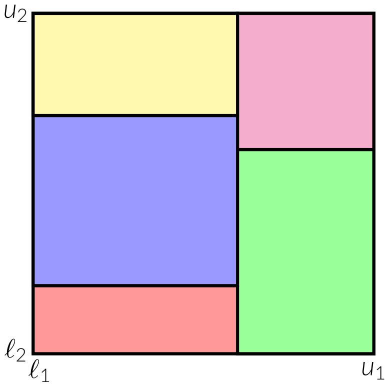
   - *Dimension*: j with probability proportional to $u_j - \ell_j$
   - *Location*: uniform on $[\ell_j, u_j]$

4. Recurse independently on each subrectangle with lifetime $\lambda - \triangle$



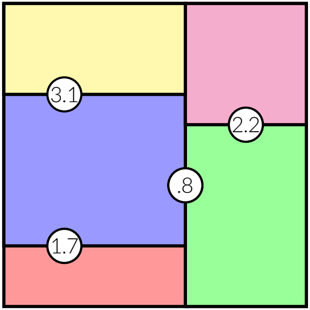
Mondrian process construction in $\mathbb{R}^d$

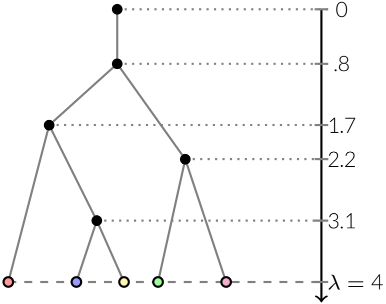
1. Fix **lifetime parameter** $\lambda > 0$
2. Draw
$$\Delta \sim Exp\left[\sum_{i=1}^{d}(u_i - \ell_i)\right]$$
3. IF $\Delta > \lambda$ stop,
   ELSE sample a split:
   - *Dimension*: j with probability proportional to $u_j - \ell_j$
   - *Location*: uniform on $[\ell_j, u_j]$
4. Recurse independently on each subrectangle with lifetime $\lambda - \Delta$



Mondrian Random Forests

- Comparable empirical performance to RF for some tasks[5]
- Minimax rates under nonparametric assumptions in arbitrary dimension[6]

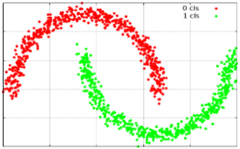


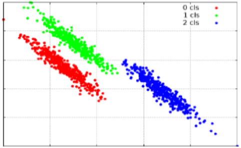


[5] [Lakshminarayanan, Roy, and Teh, 2014]
[6] [Mourtada, Gaïffas, Scornet, 2020]

Beyond axis-aligned partitions

- Non-axis-aligned splits can capture dependencies between features



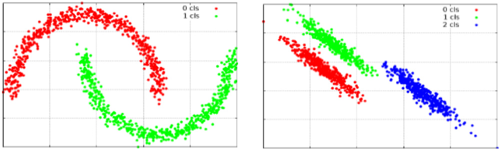


- Non-axis-aligned RF variants show **improved empirical performance**[7]
- Lack of theoretical analysis, computational efficiency

[7] [Breiman, 2001; Fan, Li, and Sisson, 2019; Tomita et al., 2020]

Beyond axis-aligned partitions

- Non-axis-aligned splits can capture dependencies between features



- Non-axis-aligned RF variants show **improved empirical performance**[7]

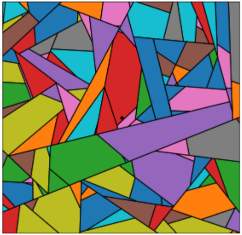
- Lack of theoretical analysis, computational efficiency

Question: Is there a generalization of the Mondrian process with non-axis-aligned cuts?

[7] [Breiman, 2001; Fan, Li, and Sisson, 2019; Tomita et al., 2020]

Stable Under Iteration Processes

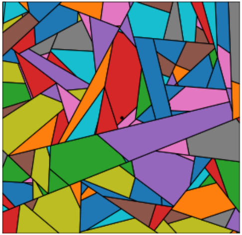
- **Yes!** Mondrian is special case of the **STIT process** in stochastic geometry
- Introduced by Nagel and Weiss in 2003
- Indexed by a **directional distribution** ϕ on $\mathbb{S}^{d-1}$



[8] [Ge, Wang, Teh, Wang, and Elliott, 2019]

Stable Under Iteration Processes

- **Yes!** Mondrian is special case of the **STIT process** in stochastic geometry
- Introduced by Nagel and Weiss in 2003
- Indexed by a **directional distribution** ϕ on $\mathbb{S}^{d-1}$



- Improved empirical performance with uniform STIT over Mondrian[8]
- General cell shapes introduce computational and **theoretical** challenges

[8] [Ge, Wang, Teh, Wang, and Elliott, 2019]

Theoretical Challenge: Minimax Rates

Theoretical Challenge: Minimax Rates

- **Assumption**: $\{(x_i, y_i)\}_{i=1}^n$ i.i.d. samples of $(X, Y) \in W \times \mathbb{R}$ such that

$$Y = f(X) + \varepsilon,$$

  where $W \subset \mathbb{R}^d$ is a compact and convex window

- $\hat{f}_{\boldsymbol{\lambda},n,M}$: STIT forest estimator of size M; lifetime parameter λ

Theoretical Challenge: Minimax Rates

- **Assumption**: $\{(x_i, y_i)\}_{i=1}^n$ i.i.d. samples of $(X, Y) \in W \times \mathbb{R}$ such that

$$Y = f(X) + \varepsilon,$$

  where $W \subset \mathbb{R}^d$ is a compact and convex window

- $\hat{f}_{\boldsymbol{\lambda},n,M}$: STIT forest estimator of size M; lifetime parameter λ

- The quality of the estimator $\hat{f}_{\boldsymbol{\lambda},n,M}$ is measured by the **quadratic risk**

$$\mathbb{E}[(\hat{f}_{\boldsymbol{\lambda},n,M}(X) - f(X))^2]$$

Theoretical Challenge: Minimax Rates

- **Assumption**: $\{(x_i, y_i)\}_{i=1}^n$ i.i.d. samples of $(X, Y) \in W \times \mathbb{R}$ such that

$$Y = f(X) + \varepsilon,$$

  where $W \subset \mathbb{R}^d$ is a compact and convex window

- $\hat{f}_{\boldsymbol{\lambda},n,M}$: STIT forest estimator of size M; lifetime parameter λ

- The quality of the estimator $\hat{f}_{\boldsymbol{\lambda},n,M}$ is measured by the **quadratic risk**

$$\mathbb{E}[(\hat{f}_{\boldsymbol{\lambda},n,M}(X) - f(X))^2]$$

- The **minimax risk** for a function class $\mathcal{F}$ is

$$\min_{\hat{f}_n} \max_{f \in \mathcal{F}} \mathbb{E}[(\hat{f}_n(X) - f(X))^2]$$

Theoretical Challenge: Minimax Rates

Theorem (Tran and O.)

(i) If f is Lipschitz, letting $\lambda_n \asymp n^{1/(d+2)}$ gives

$$\mathbb{E}[(\hat{f}_{\lambda_n,n,M}(X) - f(X))^2] \leq O\left(n^{-2/(d+2)}\right)$$

Theoretical Challenge: Minimax Rates

Theorem (Tran and O.)

(i) If f is Lipschitz, letting $\lambda_n \asymp n^{1/(d+2)}$ gives

$$\mathbb{E}[(\hat{f}_{\lambda_n,n,M}(X) - f(X))^2] \le O\left(n^{-2/(d+2)}\right)$$

(ii) If f is $\mathcal{C}^2$ and X has positive and Lipschitz density, letting $\lambda_n \asymp n^{1/(d+4)}$ and $M_n \gtrsim n^{2/(d+4)}$ gives

$$\mathbb{E}[(\hat{f}_{\lambda_n,n,M_n}(X) - f(X))^2] \le O\left(n^{-4/(d+4)}\right)$$

Theoretical Challenge: Minimax Rates

Theorem (Tran and O.)

(i) If f is Lipschitz, letting $\lambda_n \asymp n^{1/(d+2)}$ gives

$$\mathbb{E}[(\hat{f}_{\lambda_n,n,M}(X) - f(X))^2] \leq O\left(n^{-2/(d+2)}\right)$$

(ii) If f is $\mathcal{C}^2$ and X has positive and Lipschitz density, letting $\lambda_n \asymp n^{1/(d+4)}$ and $M_n \gtrsim n^{2/(d+4)}$ gives

$$\mathbb{E}[(\hat{f}_{\lambda_n,n,M_n}(X) - f(X))^2] \leq O\left(n^{-4/(d+4)}\right)$$

- STIT random forests are **minimax optimal** for Lipschitz and $\mathcal{C}^2$ functions

Rates for multi-index models[9]

- Suppose for $g : \mathbb{R}^s \to \mathbb{R}$ and $a_i \in \mathbb{R}^d$, $i = 1, \ldots, s$,

$$f(x) = g(\langle a_1, x\rangle, \ldots, \langle a_s, x\rangle), \quad x \in B_d(0, R).$$

[9] [Li,1991; Fukumizu et al., 2004; Dalalyan et al., 2008]

Rates for multi-index models[9]

- Suppose for $g : \mathbb{R}^s \to \mathbb{R}$ and $a_i \in \mathbb{R}^d, i = 1, \ldots, s$,

$$f(x) = g(\langle a_1, x\rangle, \ldots, \langle a_s, x\rangle), \quad x \in B_d(0, R).$$

- $S := \mathrm{span}(a_1, \ldots, a_s) \subseteq \mathbb{R}^d$ is the s-dimensional *relevant feature subspace*

[9] [Li,1991; Fukumizu et al., 2004; Dalalyan et al., 2008]

Rates for multi-index models[9]

- Suppose for $g : \mathbb{R}^s \to \mathbb{R}$ and $a_i \in \mathbb{R}^d, i = 1, \ldots, s$,

$$f(x) = g(\langle a_1, x\rangle, \ldots, \langle a_s, x\rangle), \quad x \in B_d(0, R).$$

- $S := \operatorname{span}(a_1, \ldots, a_s) \subseteq \mathbb{R}^d$ is the s-dimensional *relevant feature subspace*
- Let $\hat{f}_{\boldsymbol{\lambda},n,M}$ be a STIT forest estimator with directional distribution

$$\phi_n = (1 - \varepsilon_n)\phi_S + \varepsilon_n \phi_{S^\perp},$$

for $\varepsilon_n \in (0, 1)$ where $\operatorname{supp}(\phi_S) = S \cap \mathbb{S}^{d-1}$, $\operatorname{supp}(\phi_S) = S^\perp \cap \mathbb{S}^{d-1}$

[9] [Li,1991; Fukumizu et al., 2004; Dalalyan et al., 2008]

Rates for multi-index models[9]

- Suppose for $g : \mathbb{R}^s \to \mathbb{R}$ and $a_i \in \mathbb{R}^d, i = 1, \ldots, s$,

$$f(x) = g(\langle a_1, x\rangle, \ldots, \langle a_s, x\rangle), \quad x \in B_d(0, R).$$

- $S := \text{span}(a_1, \ldots, a_s) \subseteq \mathbb{R}^d$ is the s-dimensional *relevant feature subspace*
- Let $\hat{f}_{\lambda,n,M}$ be a STIT forest estimator with directional distribution

$$\phi_n = (1 - \varepsilon_n)\phi_S + \varepsilon_n \phi_{S^\perp},$$

for $\varepsilon_n \in (0, 1)$ where $\text{supp}(\phi_S) = S \cap \mathbb{S}^{d-1}$, $\text{supp}(\phi_S) = S^\perp \cap \mathbb{S}^{d-1}$

Theorem

(i) If g is Lipschitz, letting $\lambda_n \asymp n^{1/(s+2)}$ and $\varepsilon_n \asymp n^{-1/(s+2)}$ gives

$$\mathbb{E}[(\hat{f}_{\lambda_n,n,M}(X) - f(X))^2] \leq O\left(n^{-2/(s+2)}\right)$$

(ii) A similar extension holds if g is $\mathcal{C}^2$.

[9] [Li,1991; Fukumizu et al., 2004; Dalalyan et al., 2008]

Proof Idea: Bias-Variance Decomposition of the Risk

- We have the following **bias-variance decomposition**:

$$\mathbb{E}[(\hat{f}_{\lambda,n,1}(X) - f(X))^2] = \mathbb{E}[(f(X) - \bar{f}_\lambda(X))^2] + \mathbb{E}[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2],$$

where

$$\bar{f}_\lambda(x) = \mathbb{E}_X[f(X) | X \in Z_x], \quad x \in W,$$

is the orthogonal projection of $f \in L^2(W, \mu)$ onto the subspace of functions that are constant within the cell of the STIT tessellation

Proof Idea: Bias-Variance Decomposition of the Risk

- We have the following **bias-variance decomposition**:

$$\mathbb{E}[(\hat{f}_{\lambda,n,1}(X) - f(X))^2] = \mathbb{E}[(f(X) - \bar{f}_\lambda(X))^2] + \mathbb{E}[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2],$$

  where

$$\bar{f}_\lambda(x) = \mathbb{E}_X[f(X) | X \in Z_x], \quad x \in W,$$

  is the orthogonal projection of $f \in L^2(W, \mu)$ onto the subspace of functions that are constant within the cell of the STIT tessellation

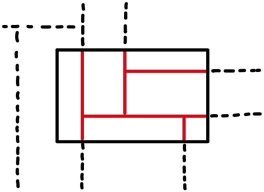
- **Bias** is controlled by the *diameter* of the cell containing X

- **Variance** is controlled by the *expected number of cells* in W

Stationary STIT Tessellation on $\mathbb{R}^d$

- $\mathcal{Y}(\lambda, W)$: STIT in compact and convex $W \subset \mathbb{R}^d$ with lifetime λ
- **Consistency**: For $W_1 \subset W_2$,

$$\mathcal{Y}(\lambda, W_1) \stackrel{d}{=} \mathcal{Y}(\lambda, W_2) \cap W_1$$

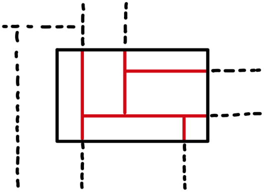


[10] [Nagel and Weiss, 2005]

Stationary STIT Tessellation on $\mathbb{R}^d$

- $\mathcal{Y}(\lambda, W)$: STIT in compact and convex $W \subset \mathbb{R}^d$ with lifetime λ
- **Consistency**: For $W_1 \subset W_2$,

$$\mathcal{Y}(\lambda, W_1) \stackrel{d}{=} \mathcal{Y}(\lambda, W_2) \cap W_1$$



- There exists a **stationary** STIT tessellation $\mathcal{Y}(\lambda)$ on $\mathbb{R}^d$ such that[10]
  - $\mathcal{Y}(\lambda) \cap W \stackrel{d}{=} \mathcal{Y}(\lambda, W)$ for all compact W

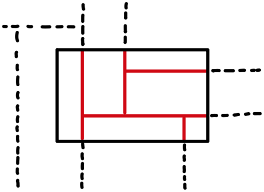
[10] [Nagel and Weiss, 2005]

Stationary STIT Tessellation on $\mathbb{R}^d$

- $\mathcal{Y}(\lambda, W)$: STIT in compact and convex $W \subset \mathbb{R}^d$ with lifetime λ

- **Consistency**: For $W_1 \subset W_2$,

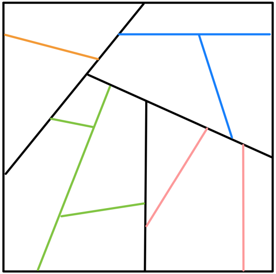
$$\mathcal{Y}(\lambda, W_1) \stackrel{d}{=} \mathcal{Y}(\lambda, W_2) \cap W_1$$



- There exists a **stationary** STIT tessellation $\mathcal{Y}(\lambda)$ on $\mathbb{R}^d$ such that[10]
  - $\mathcal{Y}(\lambda) \cap W \stackrel{d}{=} \mathcal{Y}(\lambda, W)$ for all compact W
  - **Stable Under Iteration**: for all $\lambda > 0$,

$$\mathcal{Y}(\lambda) \stackrel{d}{=} n(\mathcal{Y}(\lambda) \boxplus \cdots \boxplus \mathcal{Y}(\lambda)),$$

    s.t. $\mathcal{Y} \boxplus \mathcal{Y} := \mathcal{Y} \cup \bigcup_{c \in \text{cells}(\mathcal{Y})} (\mathcal{Y}(c) \cap c)$
    where $\{\mathcal{Y}(c) : c \in \mathcal{Y}\}$ are i.i.d. copies of $\mathcal{Y}$
  - **Scaling property**: $\lambda \mathcal{Y}(\lambda) \stackrel{d}{=} \mathcal{Y}(1)$



[10] [Nagel and Weiss, 2005]

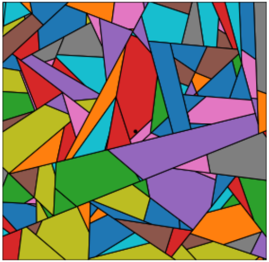
Cells of stationary random tessellations

- Cells of $\mathcal{Y}(\lambda)$ form a stationary point process on space $\mathcal{K}$ of compact convex polytopes

Cells of stationary random tessellations

- Cells of $\mathcal{Y}(\lambda)$ form a stationary point process on space $\mathcal{K}$ of compact convex polytopes

- **Typical cell** is a *centered* random polytope Z such that for all $A \in \mathcal{B}(\mathcal{K})$,



$$\mathbb{E}\left[\sum_{C \in \mathrm{cells}(\mathcal{Y})} \mathbf{1}\{C \in A\}\right] = \frac{1}{\mathbb{E}[\mathrm{vol}_d(Z)]}\mathbb{E}\left[\int_{\mathbb{R}^d} \mathbf{1}\{Z + y \in A\}\mathrm{d}y\right]$$

Proof Idea: Variance Bound

- **Recall**: Variance is controlled by the expected number of cells/leaves
- Let $W \subset \mathbb{R}^d$ be a compact and convex set
- Let $\mathcal{Y}(\lambda)$ be a STIT tessellation in $\mathbb{R}^d$ with lifetime λ

Lemma

Let Z be the typical cell of $\mathcal{Y}(1)$. Then,

$$\mathbb{E}\left[\sum_{C \in \mathrm{cells}(\mathcal{Y}(\lambda))} \mathbf{1}\{C \cap W \neq \emptyset\}\right] = \sum_{k=0}^{d} \binom{d}{k} \lambda^k \frac{\mathbb{E}[V(W[k], Z[d-k])]}{\mathbb{E}[\mathrm{vol}_d(Z)]},$$

where $\mathbb{E}[V(W[k], Z[d-k])] := \mathbb{E}[V(\underbrace{W, \dots, W}_{k}, \underbrace{Z, \dots, Z}_{d-k})]$.

Proof Idea: Variance Bound

- **Recall**: Variance is controlled by the expected number of cells/leaves
- Let $W \subset \mathbb{R}^d$ be a compact and convex set
- Let $\mathcal{Y}(\lambda)$ be a STIT tessellation in $\mathbb{R}^d$ with lifetime λ

Lemma

Let Z be the typical cell of $\mathcal{Y}(1)$. Then,

$$\mathbb{E}\left[\sum_{C \in \mathrm{cells}(\mathcal{Y}(\lambda))} \mathbf{1}\{C \cap W \neq \emptyset\}\right] = \sum_{k=0}^{d} \binom{d}{k} \lambda^k \frac{\mathbb{E}[V(W[k], Z[d-k])]}{\mathbb{E}[\mathrm{vol}_d(Z)]},$$

where $\mathbb{E}[V(W[k], Z[d-k])] := \mathbb{E}[V(\underbrace{W, \ldots, W}_{k}, \underbrace{Z, \ldots, Z}_{d-k})]$.

- Proof: apply Steiner's formula to $\mathbb{E}\left[\sum_{C \in \mathcal{P}(\lambda)} 1_{\{C \cap W \neq \emptyset\}}\right] = \frac{\mathbb{E}[\mathrm{vol}_d(W - Z_\lambda)]}{\mathbb{E}[\mathrm{vol}_d(Z_\lambda)]}$

Proof Idea: Risk Bound

- Let $\hat{f}_{\boldsymbol{\lambda},n,M}$ be the forest estimator of f corresponding to the STIT $\mathcal{Y}(\lambda)$
- Let Z_0 the zero cell (cell containing the origin), and Z the typical cell of $\mathcal{Y}(1)$

Assume $f : W \to \mathbb{R}$ is L-Lipschitz. Then,

$$\begin{aligned}
&\mathbb{E}[(\hat{f}_{\boldsymbol{\lambda},n,M}(X) - f(X))^2] \\
&\leq \frac{L\mathbb{E}[\mathrm{diam}(Z_0)^2]}{\lambda^2} + \frac{(5\|f\|_\infty^2 + 2\sigma^2)}{n} \sum_{k=0}^{d} \binom{d}{k} \lambda^k \frac{\mathbb{E}[V(W[k], Z[d-k])]}{\mathbb{E}[\mathrm{vol}_d(Z)]}
\end{aligned}$$

- Letting $\lambda = \lambda_n \asymp n^{1/(d+2)}$ gives the minimax rate for Lipschitz functions

Proof Idea: Risk Bound for Multi-Index Model

- Suppose for $g : \mathbb{R}^s \to \mathbb{R}$ and $a_i \in \mathbb{R}^d, i = 1, \dots, s$

$$f(x) := g(\langle a_1, x \rangle, \dots, \langle a_s, x \rangle)$$

- $S := \mathrm{span}(a_1, \dots, a_s) \subseteq \mathbb{R}^d$ is the s-dimensional *relevant feature subspace*
- Let $\hat{f}_{\boldsymbol{\lambda},n}$ be a STIT forest estimator with directional distribution

$$\phi_n = (1 - \varepsilon_n)\phi_S + \varepsilon_n \phi_{S^\perp}$$

Proof Idea: Risk Bound for Multi-Index Model

- Suppose for $g : \mathbb{R}^s \to \mathbb{R}$ and $a_i \in \mathbb{R}^d, i = 1, \dots, s$

$$f(x) := g(\langle a_1, x \rangle, \dots, \langle a_s, x \rangle)$$

- $S := \mathrm{span}(a_1, \dots, a_s) \subseteq \mathbb{R}^d$ is the s-dimensional *relevant feature subspace*
- Let $\hat{f}_{\lambda,n}$ be a STIT forest estimator with directional distribution

$$\phi_n = (1 - \varepsilon_n)\phi_S + \varepsilon_n \phi_{S^\perp}$$

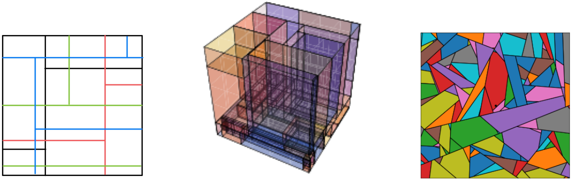
Assume $g : \mathbb{R}^s \to \mathbb{R}$ is L-Lipschitz. Then,

$$\mathbb{E}[(\hat{f}_{\lambda,n}(X) - f(X))^2] \leq \frac{L^2 \|A\|_{op}}{\lambda^2 (1 - \varepsilon_n)^2} \mathbb{E}[\mathrm{diam}(Z_0 \cap S)^2]$$
$$+ \frac{(5\|f\|_\infty^2 + 2\sigma^2)}{n} \left(\lambda^d \varepsilon_n^{d-s} \binom{d}{s} R^d \kappa_d V(\Pi_S[s], \Pi_{S^\perp}[d-s]) + o(\lambda^d \varepsilon_n^{d-s}) \right).$$

- $\Pi_S, \Pi_{S^\perp}$ are convex bodies (zonoids) defined by ϕ_S and $\phi_{S^\perp}$

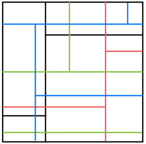
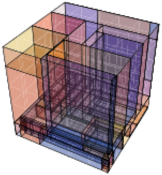
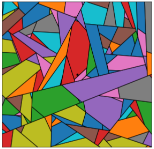
Summary and future work

- We have proved **minimax optimal** rates for a large class of random forest/partition estimators with general split directions
- **Theory of stationary random tessellations** is a powerful and flexible framework for understanding and developing random partition methods
- **Performance** depends on **geometry** of the cells (e.g. **mixed volumes** of typical cell)



Summary and future work

- We have proved **minimax optimal** rates for a large class of random forest/partition estimators with general split directions
- **Theory of stationary random tessellations** is a powerful and flexible framework for understanding and developing random partition methods
- **Performance** depends on **geometry** of the cells (e.g. **mixed volumes** of typical cell)

- How to learn directional distribution from data?
- Other applications: clustering, random feature models

Application #2:
Optimal regularizers for a data source

Joint work with Oscar Leong (Caltech), Yong Sheng Soh (National University of Singapore), and Venkat Chandrasekaran (Caltech)

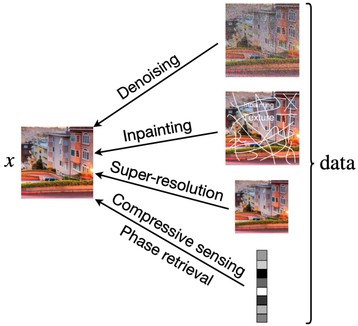
Inverse Problems and regularization

- Goal is to recover signal x from:

$$y = \mathcal{A}(x) + \varepsilon$$

  where $\mathcal{A}$ is a known forward map and ε is observation noise

- Problem may be *ill-posed*



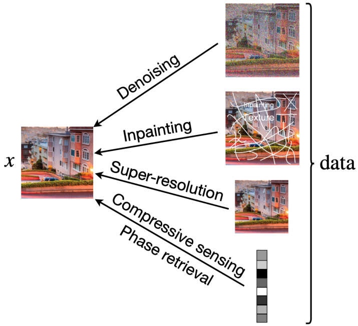
Inverse Problems and regularization

- Goal is to recover signal x from:

$$y = \mathcal{A}(x) + \varepsilon$$

  where $\mathcal{A}$ is a known forward map and ε is observation noise

- Problem may be *ill-posed*



Functional Analytic Regularization:

- Recover x with the following optimization problem:

$$\operatorname*{argmin}_{x} \mathrm{loss}(\mathcal{A}(x), y) + \textcolor{blue}{\mathrm{regularizer}(x)}$$

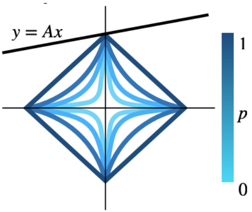
- Regularizer function promotes structure in the solution

Variety of regularizers

Hand-crafted:

Sparsity is promoted by ℓ_1 norm (convex) and by ℓ_p norm for $p \in [0, 1)$ (non-convex)

$$\underset{x}{\operatorname{argmin}} \operatorname{loss}(\mathcal{A}(x), y) + \|x\|_p$$

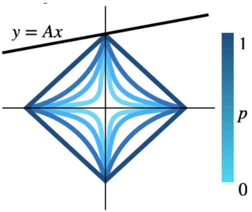


Variety of regularizers

Hand-crafted:

Sparsity is promoted by ℓ_1 norm (convex) and by ℓ_p norm for $p \in [0, 1)$ (non-convex)

$$\operatorname*{argmin}_x \operatorname{loss}(\mathcal{A}(x), y) + \|x\|_p$$

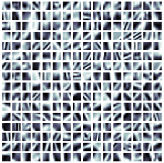


Data-driven:

Dictionary Learning (convex)

- Learn $A \in \mathbb{R}^{d \times p}$ such that $x \approx Az$, where z is sparse

$$\|A^T x\|_1 \iff \|x\|_{A(B_{\ell_1})}$$

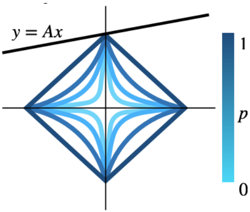


Variety of regularizers

Hand-crafted:

Sparsity is promoted by ℓ_1 norm (convex) and by ℓ_p norm for $p \in [0, 1)$ (non-convex)

$$\underset{x}{\operatorname{argmin}}\ \operatorname{loss}(\mathcal{A}(x), y) + \|x\|_p$$

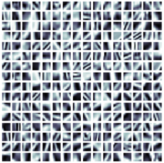


Data-driven:

Dictionary Learning (convex)

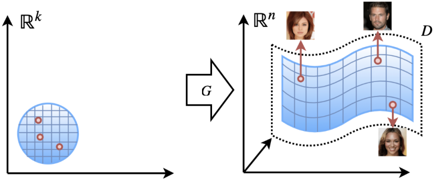
- Learn $A \in \mathbb{R}^{d \times p}$ such that $x \approx Az$, where z is sparse

$$\|A^T x\|_1 \iff \|x\|_{A(B_{\ell_1})}$$



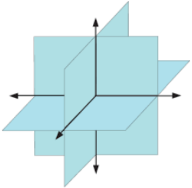
Generative models (non-convex)

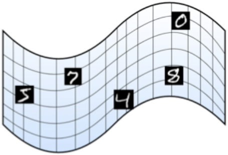
- Neural network based regularization



Main question considered in this work

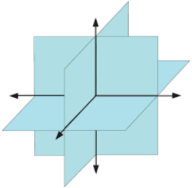
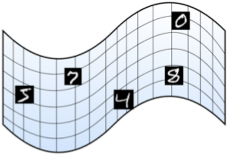
Which regularizer should one choose?





Main question considered in this work

Which regularizer should one choose?


- What is the optimal regularizer to impose for a given data source?
- Convex versus nonconvex?

Set-up and assumptions

- Let P be a probability distribution on $\mathbb{R}^d$ modeling a data source
- Define optimal regularizer f from a family $\mathcal{F}$ as a solution to:

$$\operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_P[f(x)]$$

- Conditions on $f \in \mathcal{F}$:
  - Positively homogenous: $f(\alpha x) = \alpha f(x), \alpha \geq 0$
  - $f \geq 0$ and continuous

Set-up and assumptions

- Let P be a probability distribution on $\mathbb{R}^d$ modeling a data source
- Define optimal regularizer f from a family $\mathcal{F}$ as a solution to:

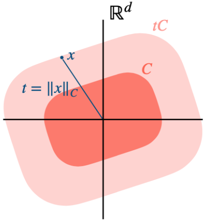
$$\operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_P[f(x)]$$

- Conditions on $f \in \mathcal{F}$:
  - Positively homogenous: $f(\alpha x) = \alpha f(x), \alpha \geq 0$
  - $f \geq 0$ and continuous

$f \in \mathcal{F} \iff f = \|\cdot\|_K$ is the Minkowski functional of a **star body** K

Minkowski functional of a compact set $C \subset \mathbb{R}^d$:

$$\|x\|_C := \inf\{t > 0 : x \in tC\}$$



Star bodies and radial functions

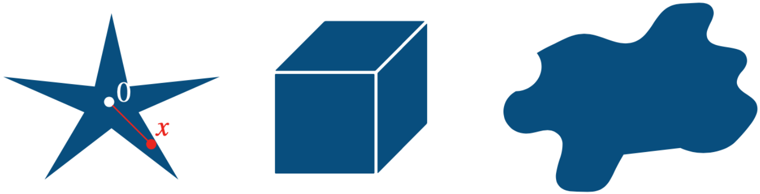
- The **radial function** of a compact set $K \subset \mathbb{R}^d$ is defined by

$$\rho_K(x) := \sup\{t > 0 : tx \in K\} = \|x\|_K^{-1}$$

- A compact set $K \subset \mathbb{R}^d$ is a **star body** if ρ_K is continuous and it is *starshaped* (with respect to the origin)

$$x \in K \Rightarrow [0, x] \subseteq K$$

- Star bodies are uniquely determined by their radial functions



Unique optimal regularizer

Theorem (Leong, Soh, Chandrasekaran, O., 2022+)

Let P be a distribution on $\mathbb{R}^d$ with density p and assume $\mathbb{E}_P[\|x\|_{\ell_2}] < \infty$.

Unique optimal regularizer

Theorem (Leong, Soh, Chandrasekaran, O., 2022+)

Let P be a distribution on $\mathbb{R}^d$ with density p and assume $\mathbb{E}_P[\|x\|_{\ell_2}] < \infty$. Suppose the following function is continuous:

$$\rho_P(u) := \left(\int_0^\infty r^d p(ru)\mathrm{d}r\right)^{1/(d+1)}, \qquad u \in \mathbb{S}^{d-1}. \tag{1}$$

Unique optimal regularizer

Theorem (Leong, Soh, Chandrasekaran, O., 2022+)

Let P be a distribution on $\mathbb{R}^d$ with density p and assume $\mathbb{E}_P[\|x\|_{\ell_2}] < \infty$. Suppose the following function is continuous:

$$\rho_P(u) := \left(\int_0^\infty r^d p(ru) \mathrm{d}r\right)^{1/(d+1)}, \qquad u \in \mathbb{S}^{d-1}. \tag{1}$$

Then $\exists$ a unique star body L_P with radial function ρ_P, and

$$K_* := \mathrm{vol}_d(L_P)^{-1/d} L_P$$

is the unique solution to

$$\underset{K \in \mathcal{S}^d : \mathrm{vol}_d(K) = 1}{\operatorname{argmin}} \quad \mathbb{E}_P[\|x\|_K]$$

Unique optimal regularizer

Theorem (Leong, Soh, Chandrasekaran, O., 2022+)

Let P be a distribution on $\mathbb{R}^d$ with density p and assume $\mathbb{E}_P[\|x\|_{\ell_2}] < \infty$. Suppose the following function is continuous:

$$\rho_P(u) := \left(\int_0^\infty r^d p(ru) \mathrm{d}r \right)^{1/(d+1)}, \qquad u \in \mathbb{S}^{d-1}. \tag{1}$$

Then $\exists$ a unique star body L_P with radial function ρ_P, and

$$K_* := \mathrm{vol}_d(L_P)^{-1/d} L_P$$

is the unique solution to

$$\underset{K \in \mathcal{S}^d : \mathrm{vol}_d(K)=1}{\operatorname{argmin}} \quad \mathbb{E}_P[\|x\|_K]$$

- If L_P is convex, then the optimal regularizer is convex!

Examples

(i) Densities induced by star bodies:

$$p(x) = \psi(\|x\|_L) \Rightarrow L_P = c_\psi L$$

Examples

(i) Densities induced by star bodies:

$$p(x) = \psi(\|x\|_L) \Rightarrow L_P = c_\psi L$$

(ii) Gaussian Mixtures:

$$P = \frac{1}{2}\mathcal{N}(0, \mathbf{\Sigma}_1) + \frac{1}{2}\mathcal{N}(0, \mathbf{\Sigma}_2)$$

where $\mathbf{\Sigma}_1 := [1, 0; 0, \varepsilon] \in \mathbb{R}^{2\times 2}$ and $\mathbf{\Sigma}_2 := [\varepsilon, 0; 0, 1] \in \mathbb{R}^{2\times 2}$

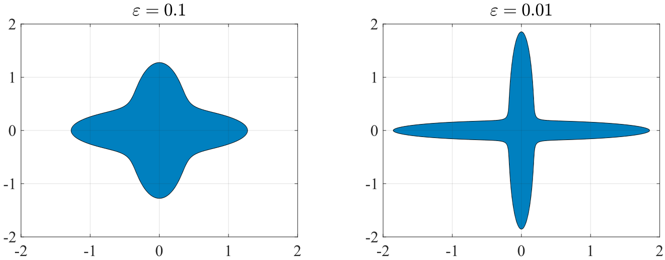


Figure: Plots of L_P for $\varepsilon = 0.1$ (left) and $\varepsilon = 0.01$ (right).

Proof

Goal: Characterize unique solution to

$$\operatorname*{argmin}_{K \in \mathcal{S}^d : \mathrm{vol}_d(K) = 1} \mathbb{E}_P[\|x\|_K]$$

By change to polar coordinates,

$$\begin{aligned} \mathbb{E}_P[\|x\|_K] &= \int_{\mathbb{R}^d} \|x\|_K p(x) \mathrm{d}x = \int_{\mathbb{S}^{d-1}} \|u\|_K \int_0^\infty r^d p(ru) \mathrm{d}r \mathrm{d}u \\ &= \int_{\mathbb{S}^{d-1}} \rho_K(u)^{-1} \rho_P(u)^{d+1} \mathrm{d}u := d\tilde{V}_{-1}(K, L_P) \end{aligned}$$

Proof

Goal: Characterize unique solution to

$$\underset{K \in \mathcal{S}^d : \mathrm{vol}_d(K)=1}{\mathrm{argmin}} \quad \mathbb{E}_P[\|x\|_K]$$

By change to polar coordinates,

$$\begin{aligned} \mathbb{E}_P[\|x\|_K] &= \int_{\mathbb{R}^d} \|x\|_K p(x) \mathrm{d}x = \int_{\mathbb{S}^{d-1}} \|u\|_K \int_0^\infty r^d p(ru) \mathrm{d}r \mathrm{d}u \\ &= \int_{\mathbb{S}^{d-1}} \rho_K(u)^{-1} \rho_P(u)^{d+1} \mathrm{d}u := d\tilde{V}_{-1}(K, L_P) \end{aligned}$$

Theorem (Dual Mixed Volume Inequality (Lutwak, 1975))

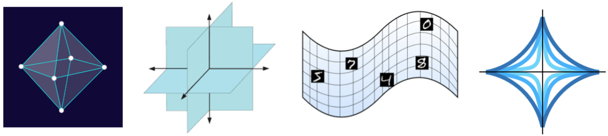
For star bodies K and L,

$$\tilde{V}_{-1}(K, L)^d \geq \mathrm{vol}_d(K)^{-1} \mathrm{vol}_d(L)^{d+1},$$

and equality hold if and only if L and K are dilates, i.e. $L = \lambda K$ *for some* $\lambda > 0$

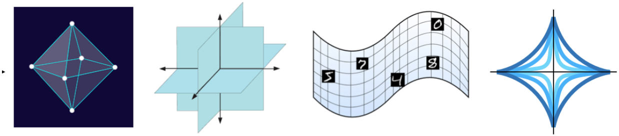
Summary and future work

- Dual Brunn-Minkowski theory provides tools for characterizing optimal functional for imposing structure on a dataset for inverse problems
- Other results: convergence of empirical minimizers and generalization error bounds



Summary and future work

- Dual Brunn-Minkowski theory provides tools for characterizing optimal functional for imposing structure on a dataset for inverse problems
- Other results: convergence of empirical minimizers and generalization error bounds



- How do optimal regularizers perform in downstream tasks?
- How to efficiently compute the optimal regularizer?

Papers

"Minimax Rates for High-Dimensional Random Tessellation Forests"
Joint with Ngoc Mai Tran. https://arxiv.org/abs/2109.10541

"Optimal Convex and Nonconvex Regularizers for a Data Source"
Joint with Oscar Leong, Yong Sheng Soh, and Venkat Chandrasekaran. In preparation.

Thank you!