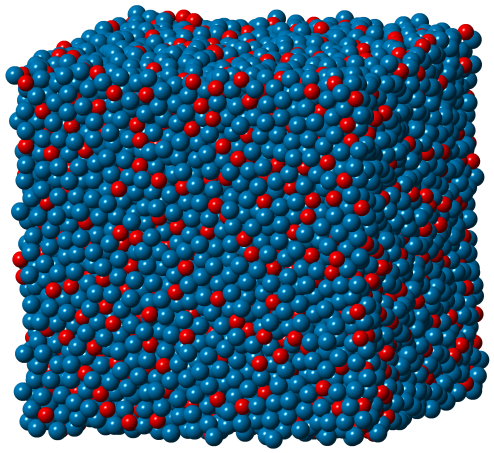
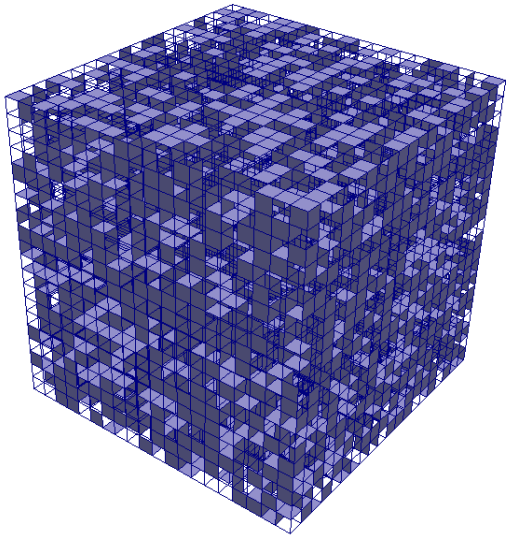
Materials TDA and random-statistical topology



Yasu Hiraoka

WPI-AIMR, Tohoku University



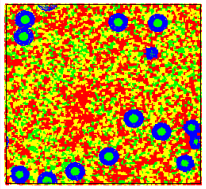
Supported by

JST CREST
SIP Structural Materials for Innovation
JST Innovation Hub MI^2I, NIMS
NEDO

Materials TDA

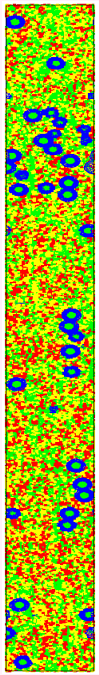
Supported by AIMR, CREST, SIP, MI^2I, NEDO

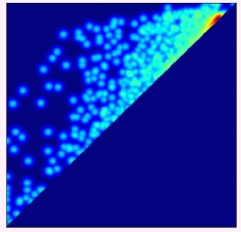
Polymer

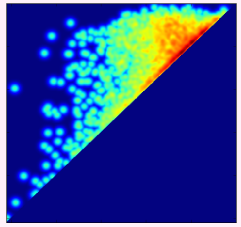


Atomic Force Microscopy image (by Nakajima)

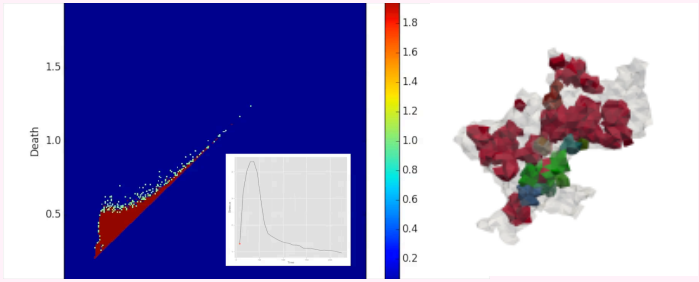
expansion







craze formation



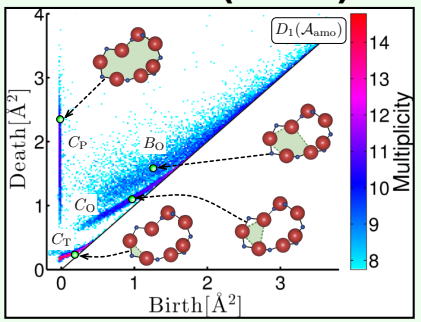
PRE (2017)

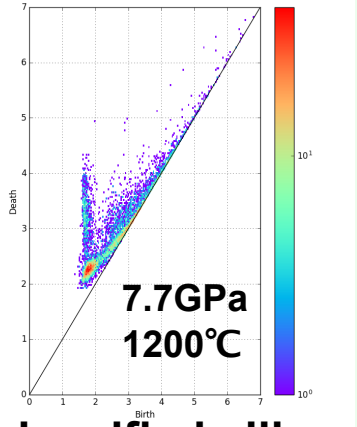
Glass

Nanotechnology (2015)

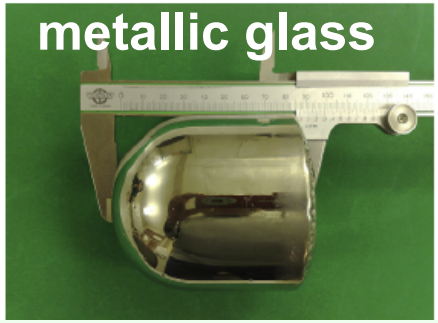
PNAS (2016)

silica

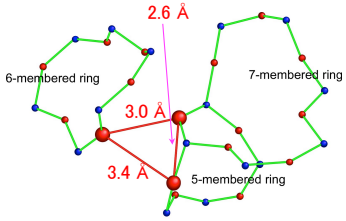




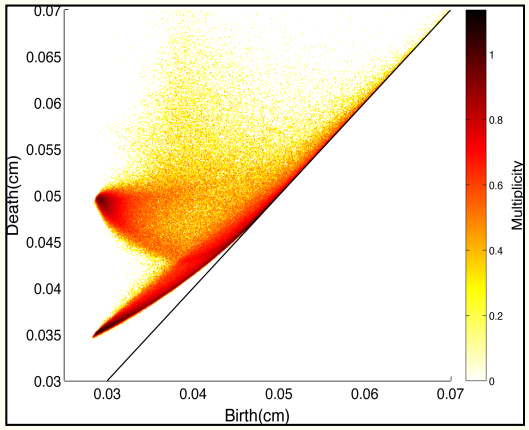
metallic glass



densified silica

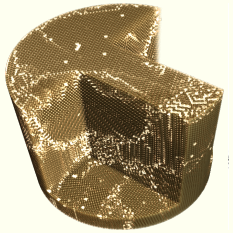


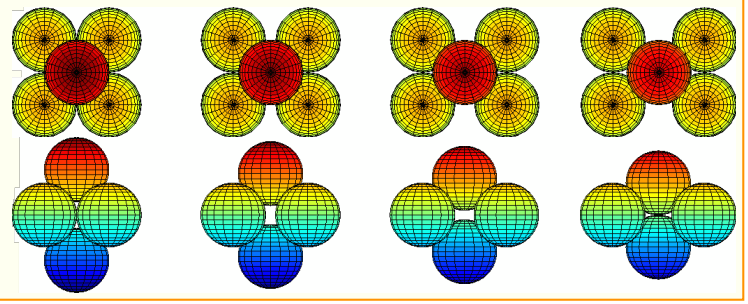
Grain



Nature Communications (2017)

deformation of octa.

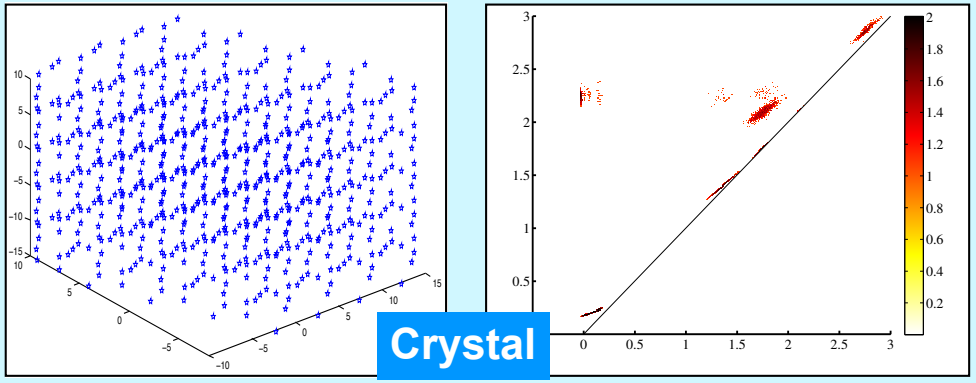




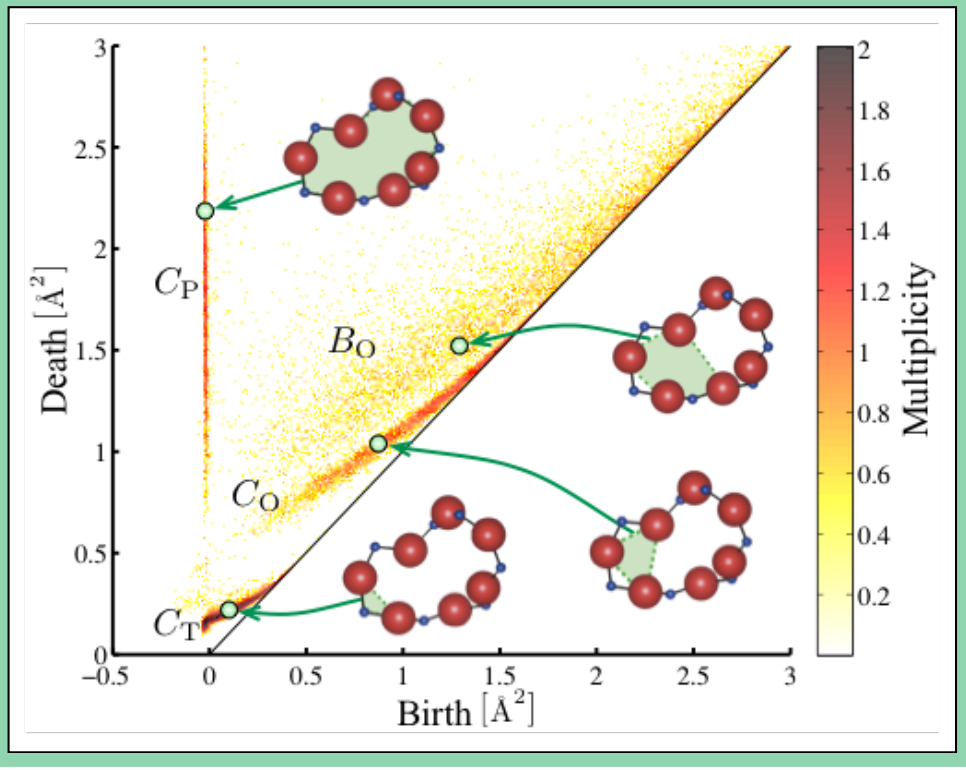
Hierarchical Structural Analysis of Silica Glass

with Nakamura, Hirata, Escolar, Matsue, Nishiura PNAS (2016) CREST TDA, SIP

MD and PD_1



Inverse Analysis of glass PD



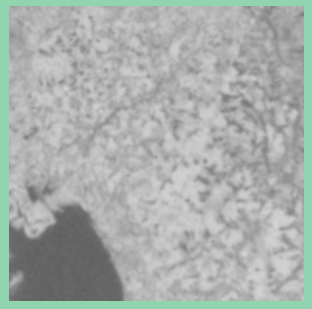
- Glass contains curves in PD
- Curves express geometric constraints (orders) of atomic configurations
- Inverse analysis reveals hierarchical ring structures
- PD multi-scale analysis characterizes inter-tetrahedral O-O orders (curve Co)
- Useful tool for structural analysis

Materials Informatics: Machine Learning on PDs

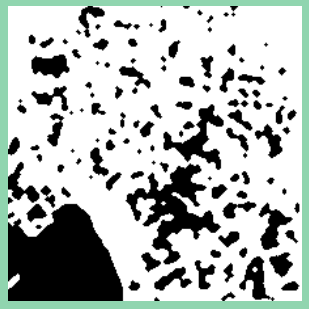
with Kimura (KEK), Obayashi (AIMR)

SIP, CREST TDA

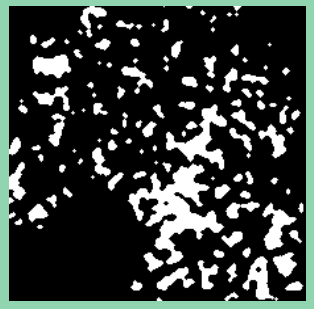
X-CT of iron-ore sinters



original



iron oxide



calcium ferrite (CF)

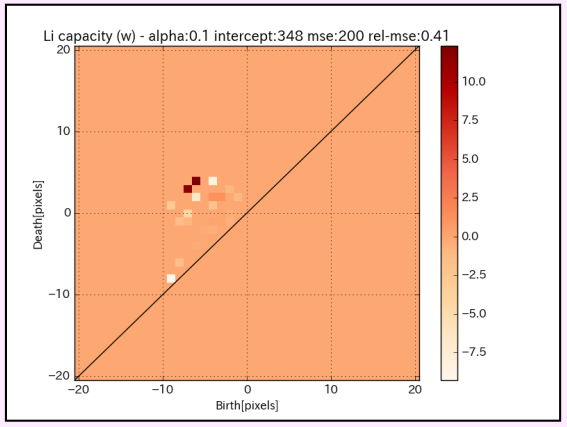
Trigger site of micro cracks are supposed to be related to hetero-structure of iron oxide and CF. No descriptors have been developed so far.

background

- large amount of experimental images are available
- want to find a compact descriptor to connect images to materials properties (conductivity, cracks, elasticity, etc)

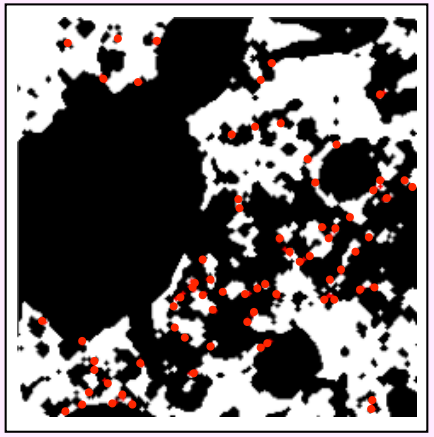
our approach

- PD for compact descriptor of images
- ML for combining with big data



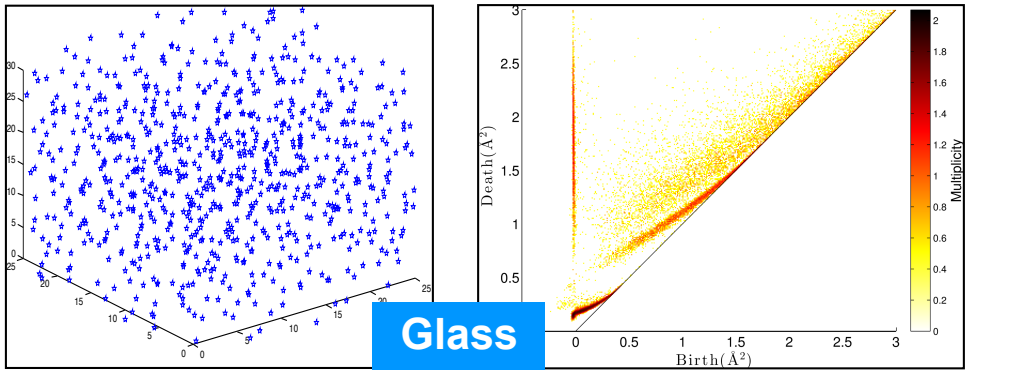
LASSO (Sparse PD)

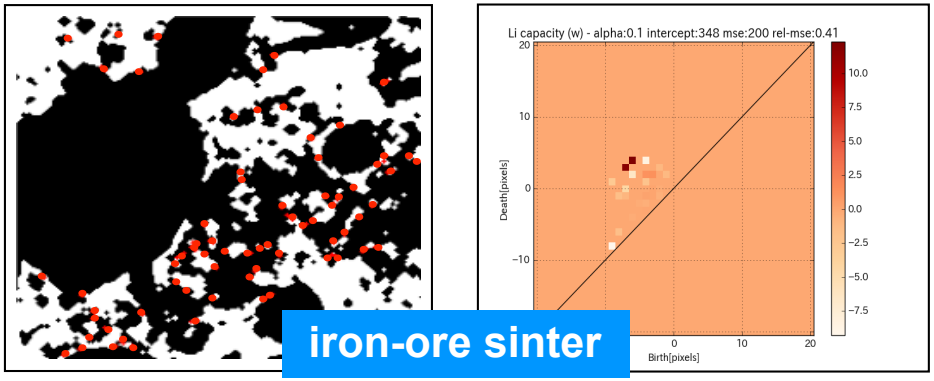
inverse



detected trigger site of cracks

Mathematical Motivation





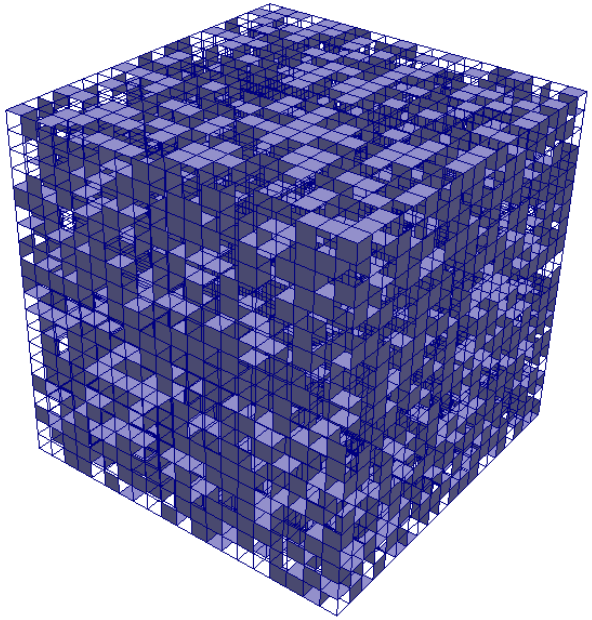
- **PDs from simulation/experiment depend on the system size L**
- **Those system sizes are usually very small scale ($\leq \mu\mathrm{m}$) comparing to the real materials ($\approx \mathrm{m}$)**
- **We need to consider a scaling limit to study universality**
- **Does there exists a limiting PD as the system size $L \to \infty$?**

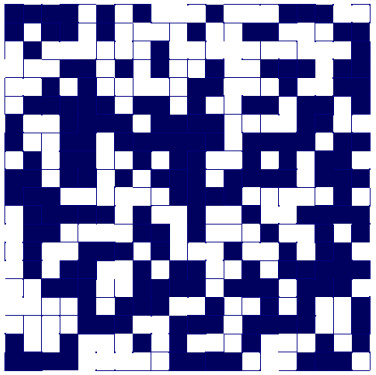
Content of today's talk

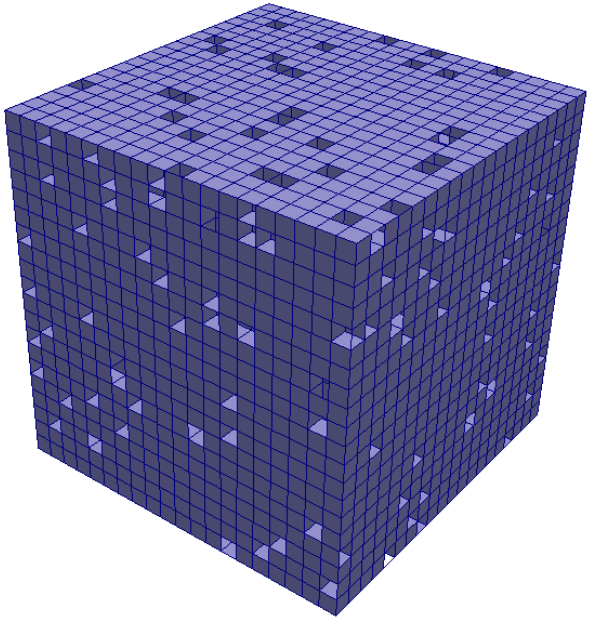
1. **Study limiting behaviors of Betti numbers and persistence diagrams on random point processes and cubical sets in R^d**
2. **Machine learnings on PDs: sparse PDs as dual vectors**

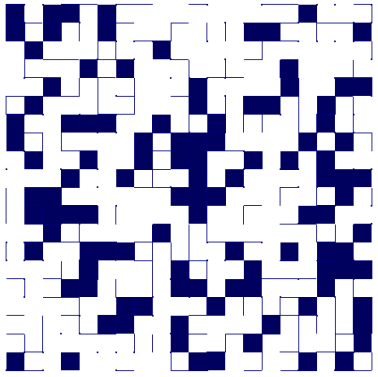
Limit theorems for random cubical homology

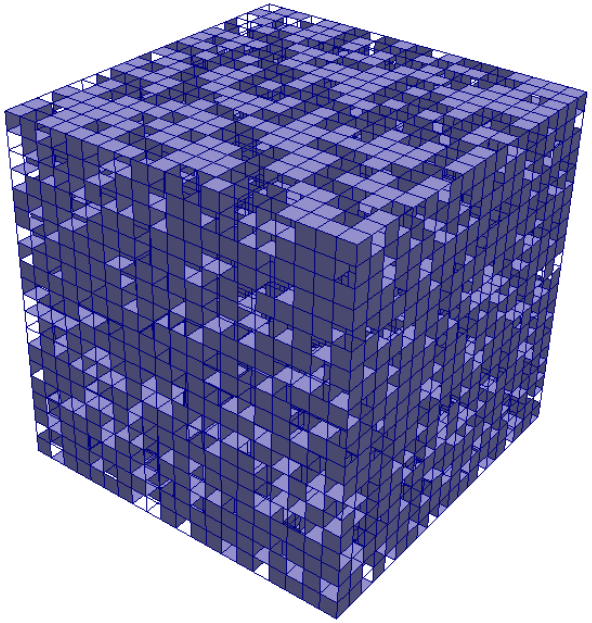
Y. H. and K. Tsunoda, arXiv:1612.08485

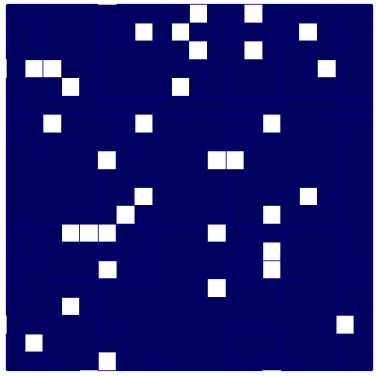






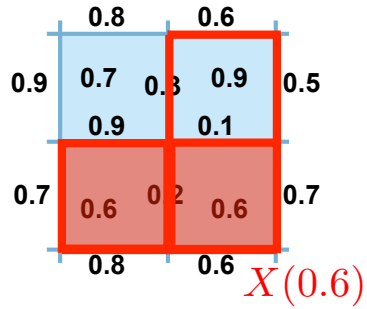




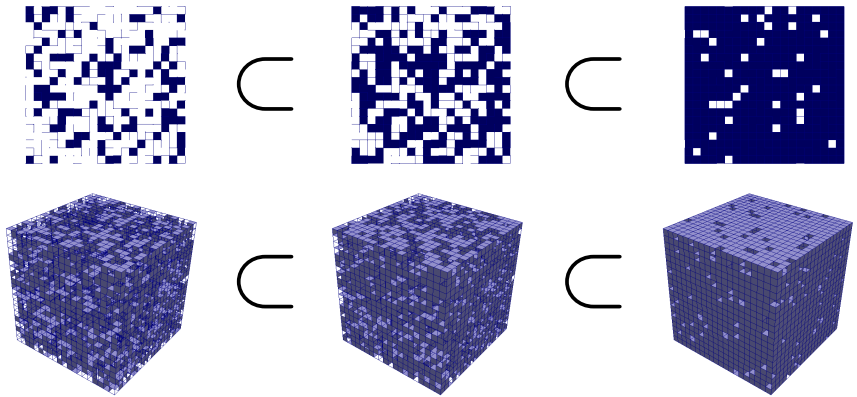


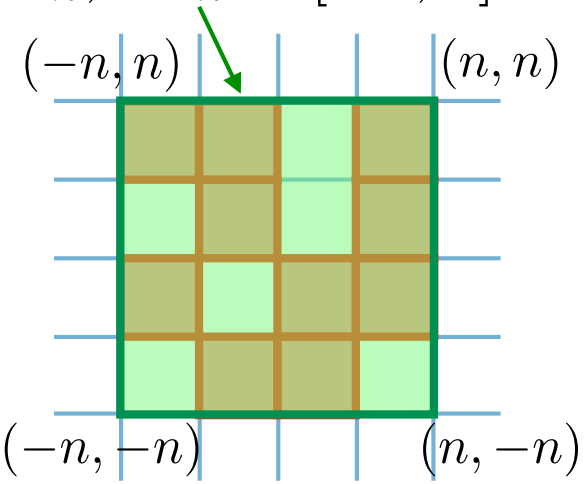
Random cubical set in $\mathbf{R}^d$

- **an elementary cube:** $Q = I_1 \times \cdots \times I_d,\ I_k = [l_k, l_k+1]$ or $I_k = [l_k, l_k]$ $l_k \in \mathbf{Z}$
- $\mathcal{K}^d$ **: the set of all elementary cubes in** $\mathbf{R}^d$
- $\Omega = [0,1]^{\mathcal{K}^d}$**: configuration space,** $\Omega \ni \omega = \{\omega_Q\}_{Q \in \mathcal{K}^d}$**: a configuration**
- P**: Probability measure on** $(\Omega, \mathcal{F})$ **satisfying**
  - **stationarity:** $P(\tau_x^{-1}A) = P(A),\ \forall x \in \mathbf{Z}^d, A \in \mathcal{F}$
  - **ergodicity:** $\tau_x^{-1}A = A, \forall x \in \mathbf{Z}^d \Longrightarrow P(A) = 0 \text{ or } 1$



- **ex: product measure, Bernoulli (Linial-Meshulam), Costa-Farber, etc**
- **random cubical set of** ω **at** t $(0 \le t \le 1)$**:** $X(t) = \bigcup \{Q \in \mathcal{K}^d : \omega_Q \le t\}$
  $X^n(t) = X(t) \cap \Lambda_n, \quad \Lambda_n = [-n, n]^d$
- **random filtration:** $\mathbb{X} = \{X(t)\}_{0 \le t \le 1}$





Limit theorems for random cubical homology

Y. H. and K. Tsunoda, arXiv:1612.08485

Theorem (LLN): Let $\beta_q^n(t) = \beta_q(X^n(t))$**. For each** $0 \leq q < d$ **and** $t \in [0,1]$ **there exists a non-random constant** $\hat{\beta}_q(t)$ **s.t.**

$$\frac{\beta_q^n(t)}{|\Lambda_n|} \longrightarrow \hat{\beta}_q(t) \quad \text{as} \quad n \longrightarrow \infty \text{ almost surely.}$$

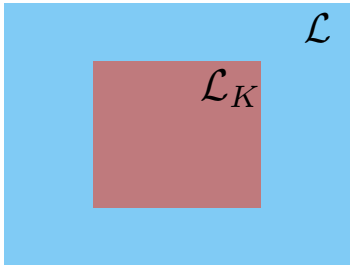
Let $\mathcal{L}_K = \{Q \subset \Lambda_K\}$**. For** $\mathcal{L} \supset \mathcal{L}_K$**, define random cubical sets**

$$X_{\mathcal{L}}(t) := \bigcup \{Q \in \mathcal{L} \colon \omega_Q \leq t\} \qquad X_{\mathcal{L}}^K(t) := \bigcup \{Q \in \mathcal{L} \setminus \mathcal{L}_K \colon \omega_Q \leq t\}$$

Let $\Omega_q(K,t) \subset \Omega$ **be the set of all configurations satisfying**

$$\beta_q(X_{\mathcal{L}}(t)) \geq 1 + \beta_q(X_{\mathcal{L}}^K(t))$$

for any finite subset $\mathcal{L} \subset \mathcal{K}^d$ **with** $\mathcal{L}_K \subset \mathcal{L}$**.**



Proposition (positivity): If there exists $K > 0$ **with** $P(\Omega_q(K,t)) > 0$**, then** $\hat{\beta}_q(t) > 0$**.**

Limit theorems for random cubical homology

Y. H. and K. Tsunoda, arXiv:1612.08485

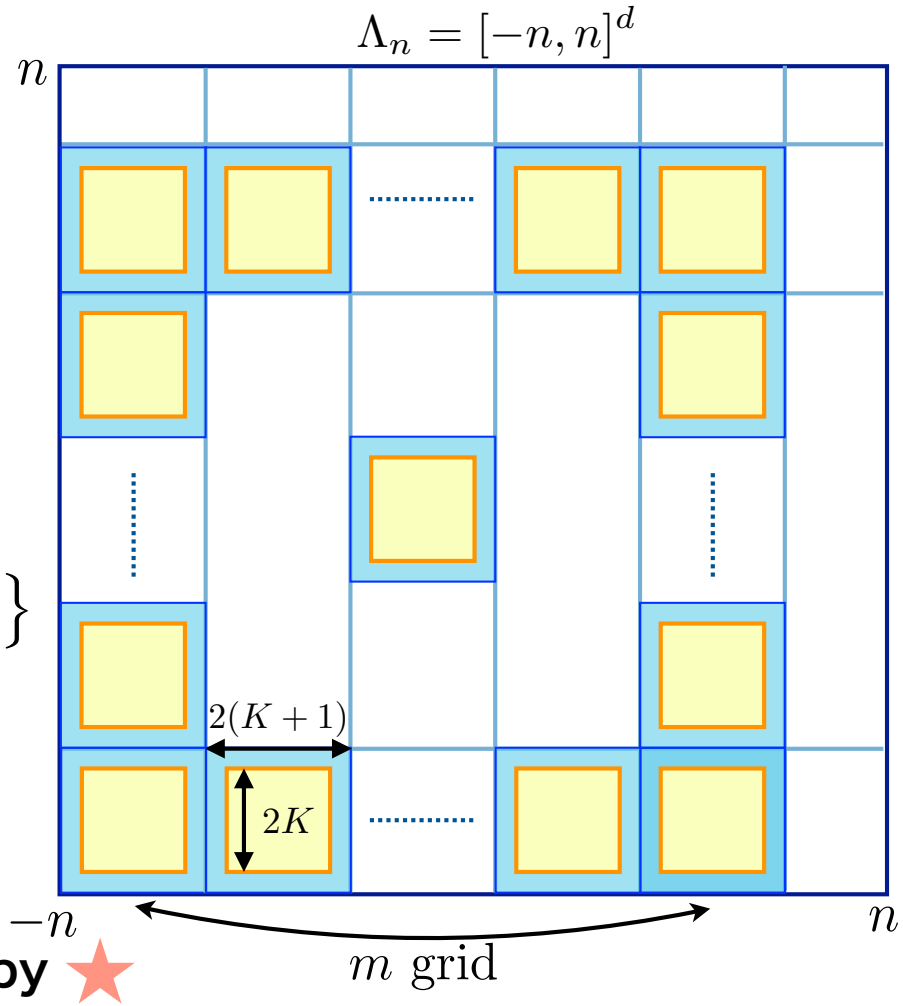
Proof of LLN $\beta_q^n(t)/|\Lambda_n| \to \hat{\beta}_q(t)$**:**

- **Set** $\hat{\beta}_q(t) := \overline{\lim} \dfrac{1}{|\Lambda_n|} E[\beta_q^n(t)]$
- **Fix** $K \in \mathbb{N}$ **and take** $m \in \mathbb{N}$ **s.t.**

$$(K+1)m \le n < (K+1)(m+1)$$ ★

- **Set** $X^{m,K}(t) := \{Q \in \mathcal{K}^d : \omega_Q \le t, Q \subset$ ▢$\}$

$$\beta_q^{m,K}(t) := \beta_q(X^{m,K}(t))$$



Then

$$\left|\frac{\beta_q^n(t)}{|\Lambda_n|} - \hat{\beta}_q(t)\right| \le \left|\frac{\beta_q^n(t)}{|\Lambda_n|} - \frac{\beta_q^n(t)}{|\Lambda_K| m^d}\right| \to 0 \text{ by } ★$$

$$+\left|\frac{\beta_q^n(t)}{|\Lambda_K| m^d} - \frac{\beta_q^{m,K}(t)}{|\Lambda_K| m^d}\right| \to 0 \text{ by } |\beta_q(Y) - \beta_q(X)| \le \#Y - \#X \quad (X \subset Y)$$

$$\le 3^d |Z| \quad (Y \setminus X \subset Z)$$

$$+\left|\frac{\beta_q^{m,K}(t)}{|\Lambda_K| m^d} - \hat{\beta}_q(t)\right| \to 0 \text{ by multivariate ergodic theorem}$$

Limit theorems for random cubical homology

Y. H. and K. Tsunoda, arXiv:1612.08485

Lifetime sum on $\Lambda_n = [-n, n]^d$**:** $L_q^n = \int_0^1 \beta_q^n(t)dt$ $\qquad \beta_q^n(t) = \beta_q(X^n(t))$

- **= lifetime sum of the q-th persistent homology on** $\mathbb{X}^n = \{X^n(t)\}_{0 \le t \le 1}$
- **plays an important role for higher dim generalization of Frieze's** $\zeta(3)$**-theorem (H. and Shirai. Rand. Str. Alg. 2017)**
- **If we ignore the exceptional sets of a.s. convergence,**

$$\lim_n \frac{L_q^n}{|\Lambda_n|} = \lim_n \int_0^1 \frac{\beta_q^n(t)}{|\Lambda_n|} dt = \int_0^1 \lim_n \frac{\beta_q^n(t)}{|\Lambda_n|} dt = \int_0^1 \hat{\beta}_q(t)dt$$

LLN for Betti

dominated convergence

but not a.s. convergence in general ($\because$ **continuous parameter** t **)**

- **For LLN, we need a uniform convergence.**

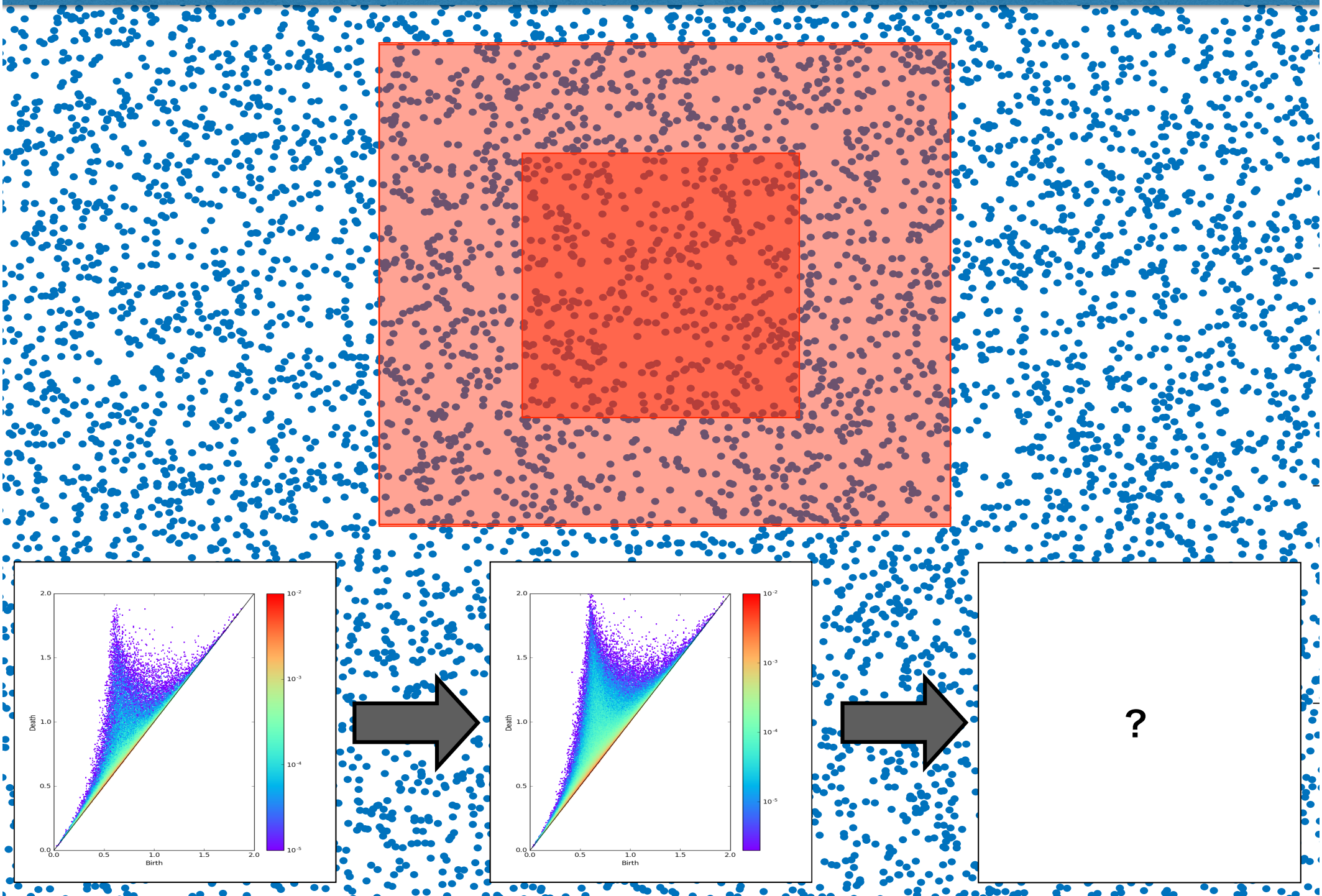
Theorem: Assume the marginal distribution function $P(\omega_Q \le t)$ **is continuous in** t**. Then,**

$$\lim_{n \to \infty} \sup_{t \in [0,1]} \left| \frac{1}{|\Lambda_n|} \beta_q^n(t) - \hat{\beta}_q(t) \right| = 0 \quad \textbf{almost surely.}$$

Corollary (LLN): $\lim_{n \to \infty} \frac{L_q^n}{|\Lambda_n|} = \int_0^1 \hat{\beta}_q(t)dt$ **almost surely.**

Limit theorem for persistence diagrams

T. K. Duy, T. Shirai, and Y. H., arXiv:1612.08371



Limit theorem for Betti number

- $\Lambda_L = [-L/2, L/2)^d$: **window in** $\mathbf{R}^d$
- Φ : **point process on** $\mathbf{R}^d$ **(locally finite random counting measure)**
- Φ_A **: restriction of** Φ **on** $A \subset \mathbf{R}^d$
- $\mathbb{C}(\Phi_A, r)$ **: Čech complex built on** Φ_A **with radius** r

Theorem (Yogeshwaran, Subag, Adler. 2015):

Assume that Φ **is a stationary point process having all finite moments. Then, there exists a constant** $\hat{\beta}_q^r$ **such that**

$$\frac{\mathbb{E}[\beta_q(\mathbb{C}(\Phi_{\Lambda_L}, r))]}{L^d} \longrightarrow \hat{\beta}_q^r \qquad \text{as } L \to \infty.$$

In addition, if Φ **is ergodic, then**

$$\frac{\beta_q(\mathbb{C}(\Phi_{\Lambda_L}, r))}{L^d} \longrightarrow \hat{\beta}_q^r \qquad \text{as } L \to \infty$$

holds almost surely.

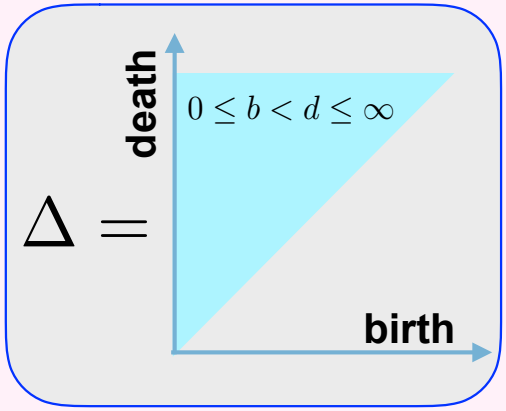
Limit theorem for persistence diagrams

T. K. Duy, T. Shirai, and Y. H., arXiv:1612.08371

Theorem (LLN):

Assume that Φ is a stationary point process having all finite moments. Let $\xi_{q,L}$ be the point process on Δ corresponding to the q-th PD for $\mathbb{K}(\Phi_{\Lambda_L})$. Then, there exists a unique Radon measure ν_q on Δ s.t.

$$\frac{1}{L^d}\mathbb{E}[\xi_{q,L}] \xrightarrow{v} \nu_q \quad \text{as } L \to \infty.$$



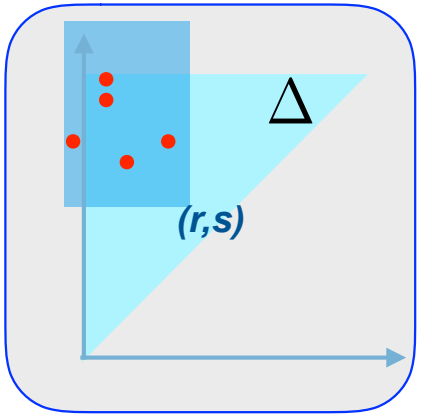
In addition, if Φ is ergodic, then

$$\frac{1}{L^d}\xi_{q,L} \xrightarrow{v} \nu_q \quad \text{as } L \to \infty$$

holds almost surely.

Sketch of proof

1. **Show a LLN for persistence Betti numbers $\beta_q^{r,s}$**
2. **Apply random measure theory**



Geometric model

- $\mathscr{F}(\mathbf{R}^d)$ **: the collection of all finite subsets in** $\mathbf{R}^d$
- **Let** $\kappa : \mathscr{F}(\mathbf{R}^d) \to [0, \infty)$ **be a function satisfying**

  1. $\kappa(\sigma) \leq \kappa(\tau)$ **for** $\sigma \subset \tau$

  2. translation invariant: $\kappa(\sigma + x) = \kappa(\sigma)$ **for** $\forall x \in \mathbf{R}^d$

  3. there is an increasing function $\rho : [0, \infty) \to [0, \infty)$ **s.t.**

$$\|x - y\| \leq \rho(\kappa\{x, y\})$$

- **Given** κ**, define a filtration** $\mathbb{K}(\Phi) = \{K(\Phi, t) \colon 0 \leq t < \infty\}$ **of simplicial complexes on a point process** Φ **by**

$$K(\Phi, t) = \{\sigma \subset \Phi \colon \kappa(\sigma) \leq t\} \qquad \textbf{(called } \kappa\textbf{-filtration)}$$

- **Čech filtration:** $\kappa(\{x_0, \ldots, x_k\}) = \inf_{w \in \mathbf{R}^d} \max_{0 \leq i \leq k} \|x_i - w\|$
- **Rips filtration:** $\kappa(\{x_0, \ldots, x_k\}) = \max_{0 \leq i < j \leq k} \frac{\|x_i - x_j\|}{2}$

Stability for κ-filtration and support of limiting measure

Theorem (Stability):

Suppose that κ is Lipschitz w.r.t d_H, i.e., there exists $\gamma > 0$ s.t.

$$|\kappa(\sigma) - \kappa(\sigma')| \leq \gamma d_H(\sigma, \sigma') \quad \textbf{for } \forall \sigma, \sigma' \in \mathscr{F}(\mathbf{R}^d).$$

Then, $d_B(D_q(\kappa, \Phi), D_q(\kappa, \Phi')) \leq \gamma d_H(\Phi, \Phi')$ **for** $\Phi, \Phi' \in \mathscr{F}(\mathbf{R}^d)$.

- $(b, d) \in \Delta$ **is realizable** $\Leftrightarrow$ $\exists \Phi \in \mathscr{F}(\mathbf{R}^d)$ **s.t.** $(b, d) \in D_q(\kappa, \Phi)$
- $R_q = R_q(\kappa)$ **: the set of all realizable points**
- ν_q **: the limiting persistence diagram (unique Radon measure in Theorem (LLN))**

Theorem (Support and realizability):

Let κ be Lipschitz, Φ be a stationary point process, and ν_q be its limiting PD. If Φ satisfies conditions about absolute continuity w.r.t Poisson p.p., then $\operatorname{supp} \nu_q = \overline{R_q(\kappa)}$.

Corollary (Support of Čech limiting PD)

For Čech PD generated by Poisson/Ginibre in $\mathbf{R}^d$,

$$\operatorname{supp} \nu_q = \Delta, \quad q = 1, \ldots, d-1$$

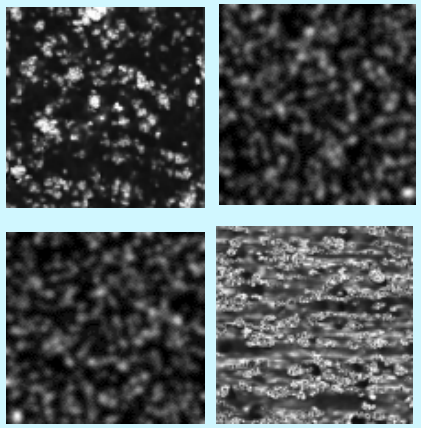
Statistical inverse analysis on persistence diagram

with Obayashi (AIMR) arXiv:1706.10082 CREST TDA, SIP, NEDO, MI^2

Background

- PDs are good descriptors in materials science
- Want to extract statistical features in the dataset of PDs
- Vectorization of PDs are necessary for applying machine learnings (persistence landscape, persistence image, PSSK, PWGK, etc)
- Want to study the original data space (inverse problems)

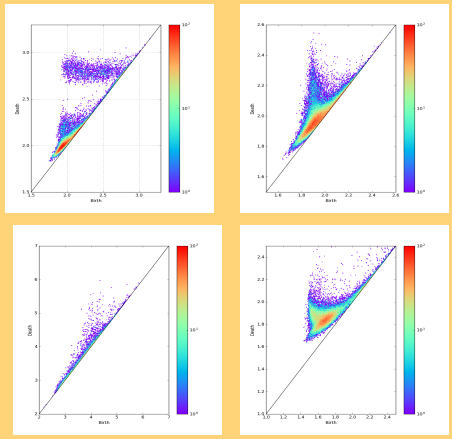
Dataset of inputs



direct

inverse

Dataset of PDs



direct

inverse

Machine learning

- SVM
- PCA
- Regression
- Anomaly detection
- Time series
- etc

Study machine learning models based on persistence diagrams

Vectorization: persistence image
ML: Logistic regression, Linear regression (LASSO/RIDGE)

Logistic regression of persistent homology

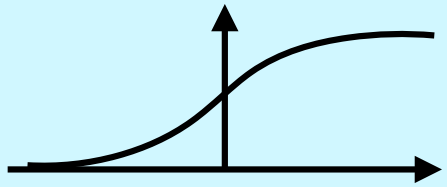
Logistic regression:

Given a training set $\{(x_i, y_i) : x_i \in \mathbf{R}^n, y_i \in \{0,1\}\}_{i=1}^M$ **,**

find optimal $w \in \mathbf{R}^n$ **and** $b \in \mathbf{R}$ **for the model**

$$P(y=1 \mid w,b) = g(w \cdot x + b),$$
$$P(y=0 \mid w,b) = 1 - P(y=1 \mid w,b) = g(-w \cdot x - b),$$

$$g(z) = 1/(e^{-z}+1)$$



⟺ **find the minimizer**

$$L(w,b) = -\frac{1}{M}\sum_{i=1}^{M}\{y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\} + \lambda R(w)$$
$$\hat{y}_i = g(w \cdot x_i + b)$$

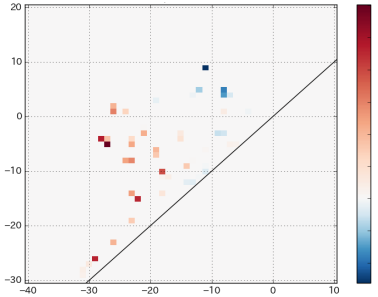
- **explanatory variable** $x \in \mathbf{R}^n$ **: (vectorized) persistence diagram**
- **response variable** $y \in \{0,1\}$ **: (binary) classification**
- **learned vector** w **can be expressed by PD (called learned PD)**

➡ **generators in the learned PD identify the relevant geometric features for classification**

- **regularization**

LASSO: $R(w) = ||w||_1$

(sparse PD analysis)

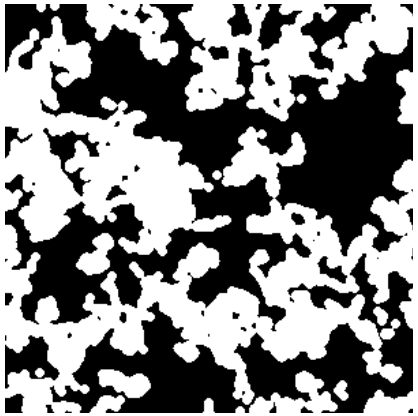


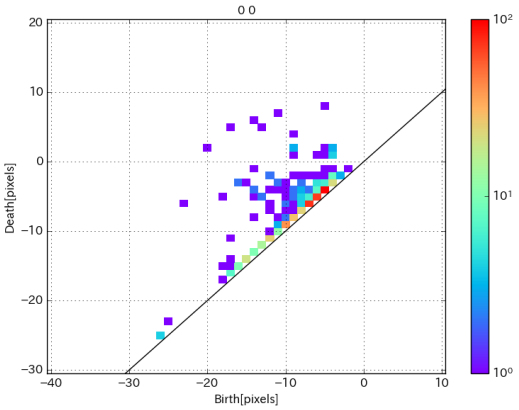
RIDGE: $R(w) = \frac{1}{2}||w||_2^2$

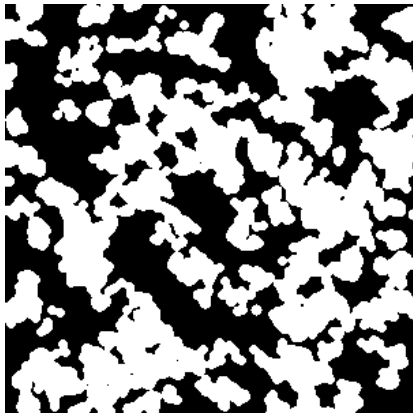
(nice math property)

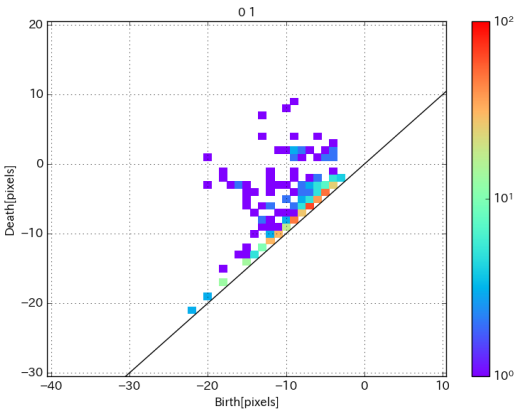
Performance of RIDGE logistic regressions: Easy example

Model A (200 trainings, 100 tests)



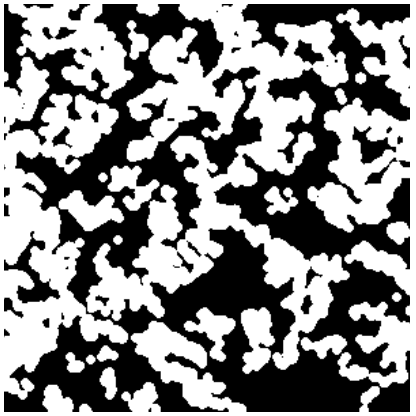


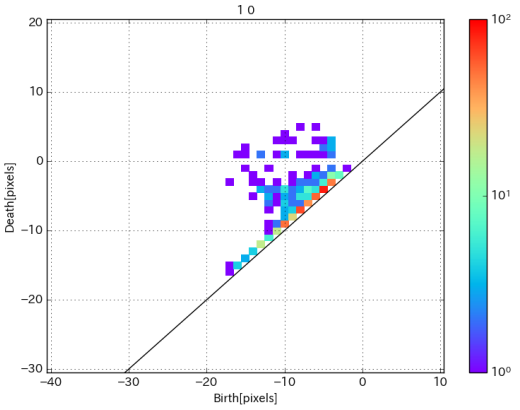


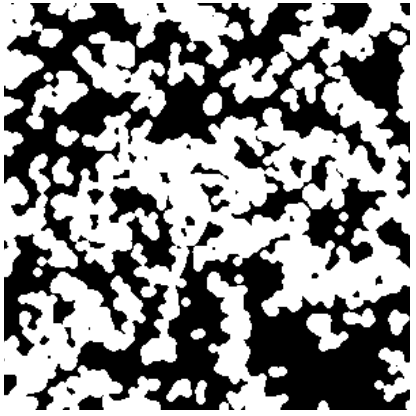


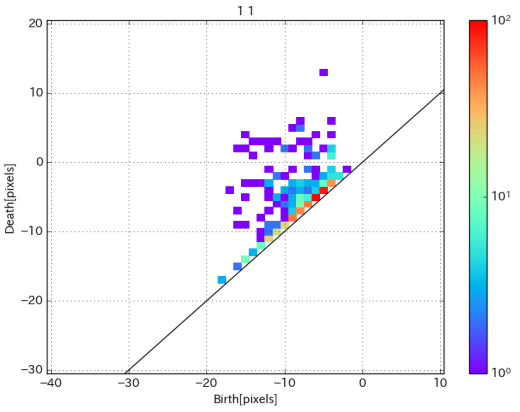
$y = 0$

Model B (200 trainings, 100 tests)







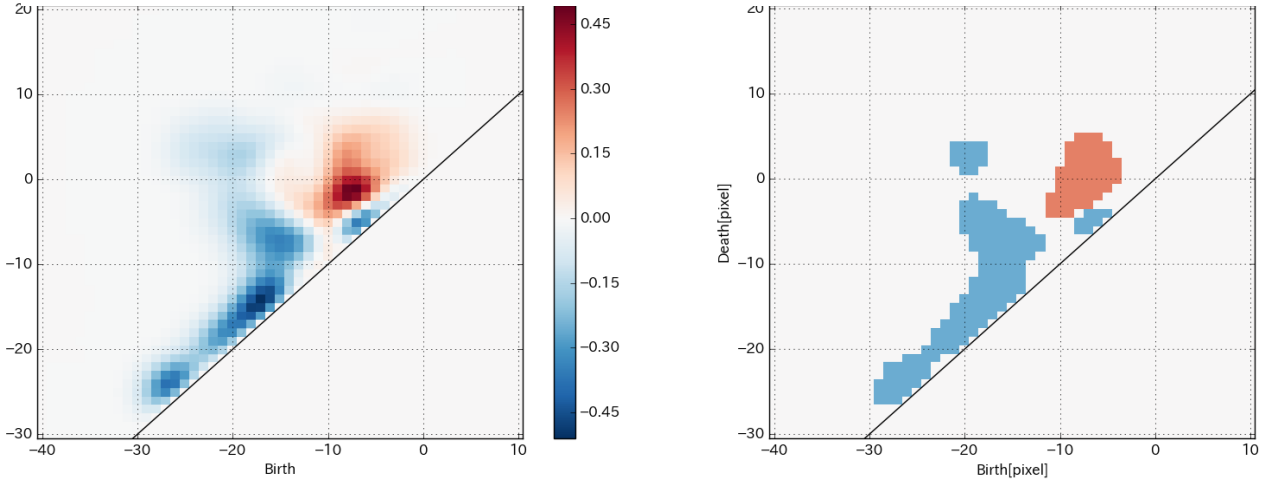


$y = 1$

Classification result (mean accuracy) = 100%

Performance of RIDGE logistic regressions: Easy example

Learned persistence diagram and its thresholding (with RIDGE)

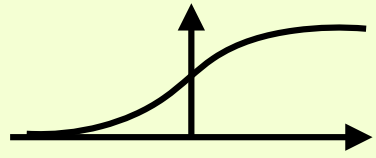


Red (resp. blue) generators contribute to 1 (resp. 0) for classification

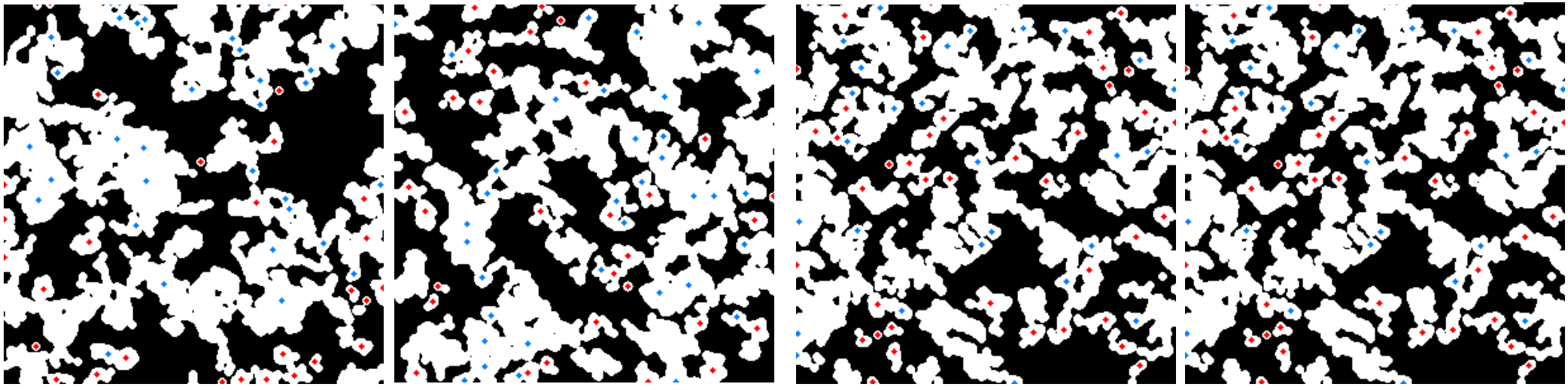
Logistic regression model:

$$P(y = 1 \mid w, b) = g(w \cdot x + b),$$
$$P(y = 0 \mid w, b) = 1 - P(y = 1 \mid w, b) = g(-w \cdot x - b),$$

$$g(z) = 1/(e^{-z} + 1)$$



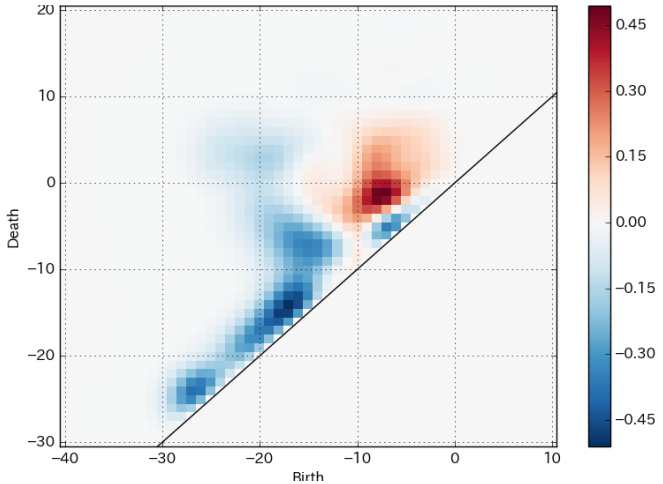
Geometric features contributing for classification (via inverse prob.)

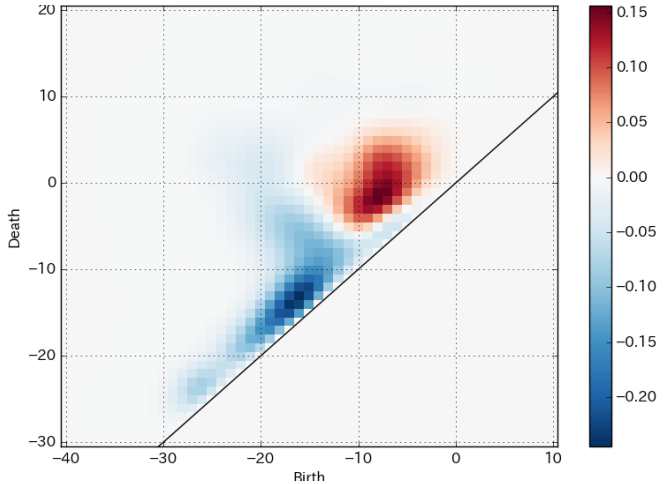


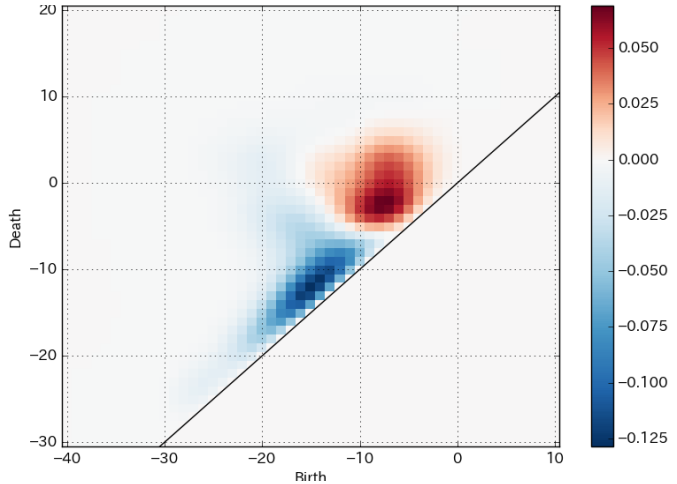
Performance of LASSO/RIDGE logistic regressions: Easy example

RIDGE/LASSO learned PDs and overfitting parameters

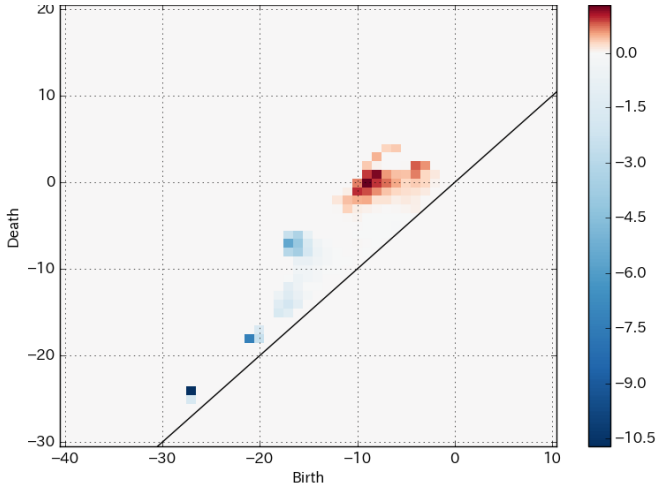
<RIDGE>

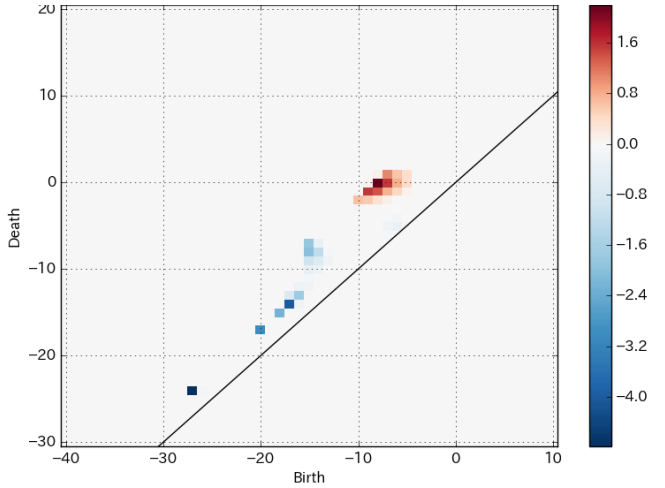


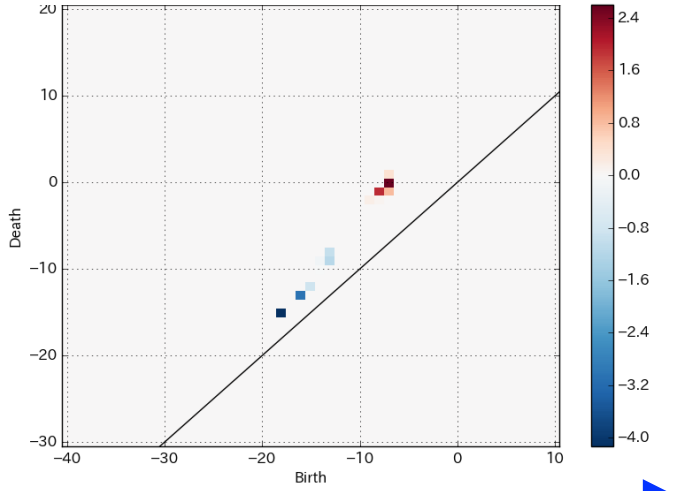




<LASSO>



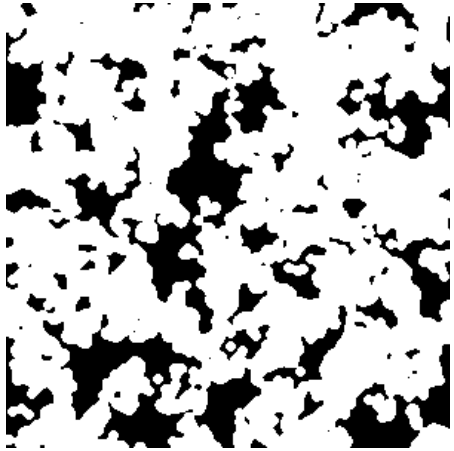


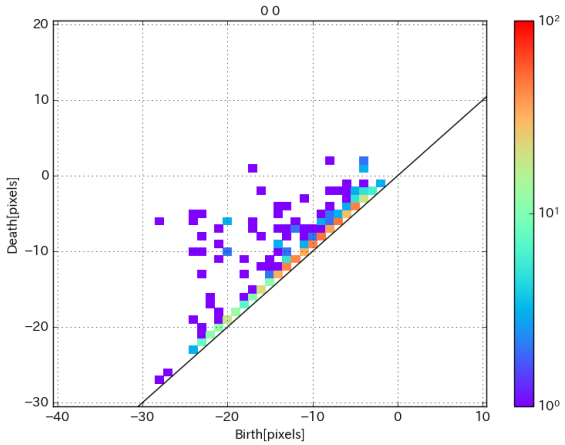


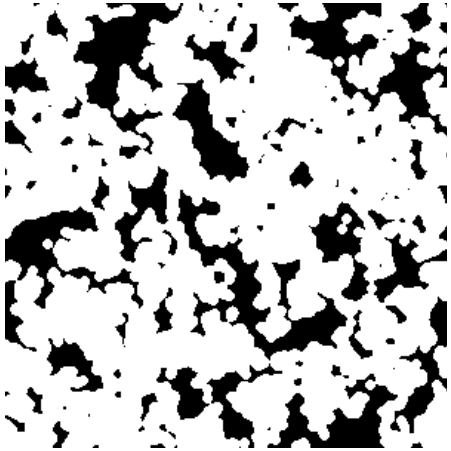
(complex) **(simple)** λ

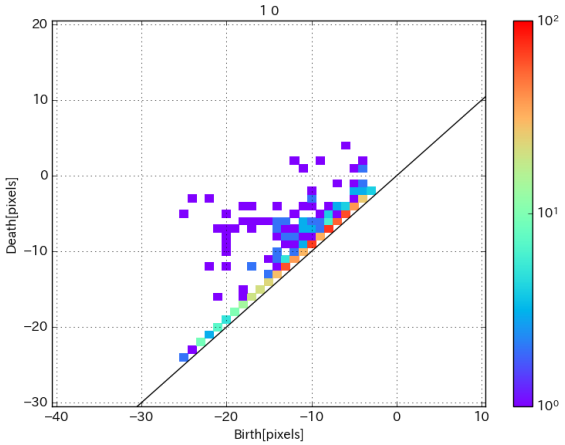
sparse persistence diagram shows most effective generators for learning

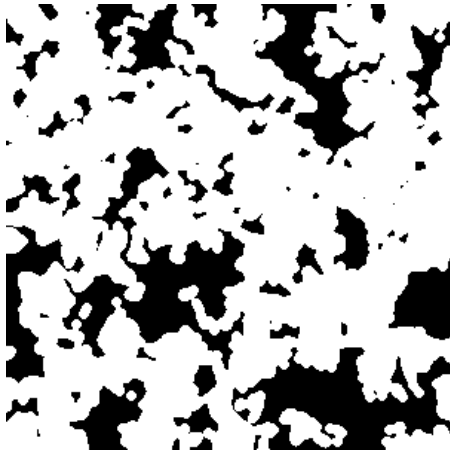
Performance of logistic regressions: Hard example

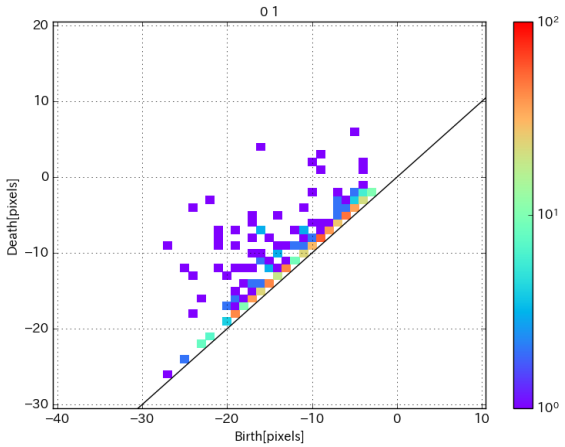


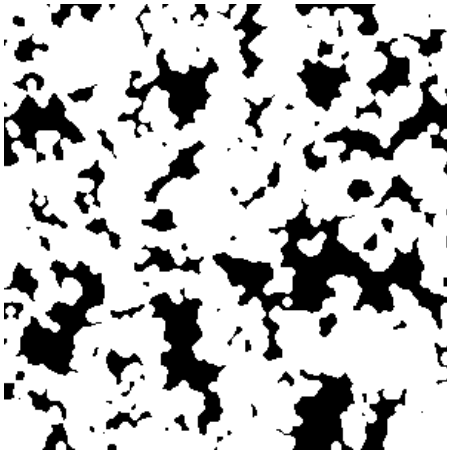


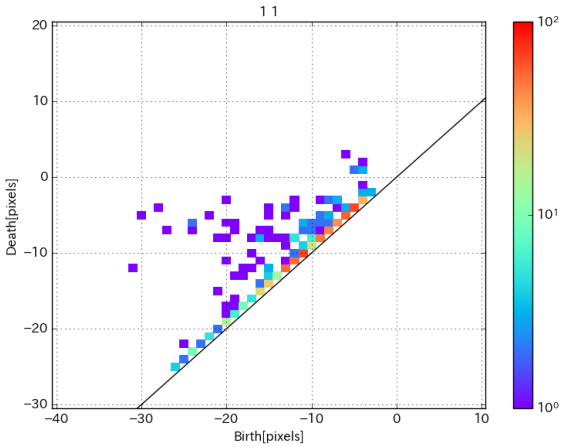










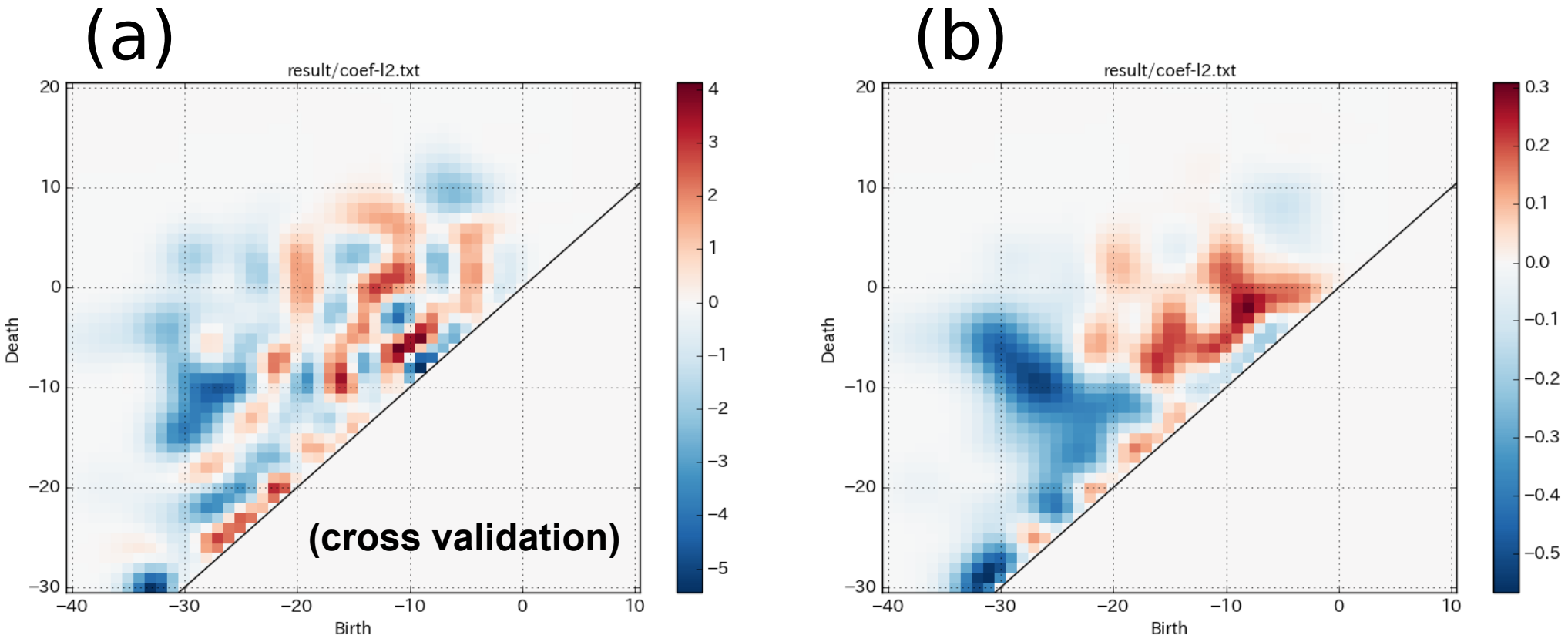


Classification result (mean accuracy) = 92%

Performance of RIDGE logistic regressions: Hard example

RIDGE learned PDs and overfitting parameters

<RIDGE>



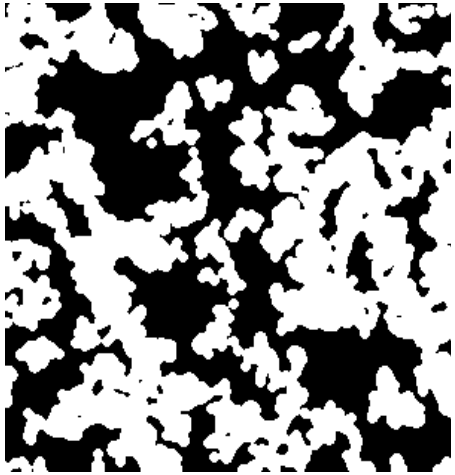
(complex) **(simple)** λ

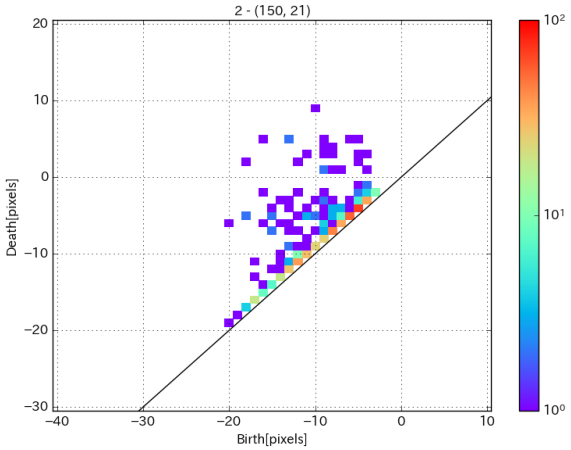
Performance comparison

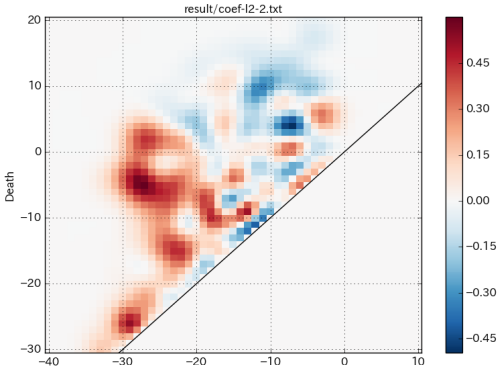
Method	Mean accuracy
PI, logistic regression, ℓ^2-penalty	0.92
PI, SVM classifier with RBF kernel	0.935
Bag of keypoints using sift with grid sampling, SVM classifier with χ^2 kernel	0.85
# of connected components of black pixels	0.73
# of connected components of white pixels	0.50
# of white pixels	0.50

Performance of linear regressions

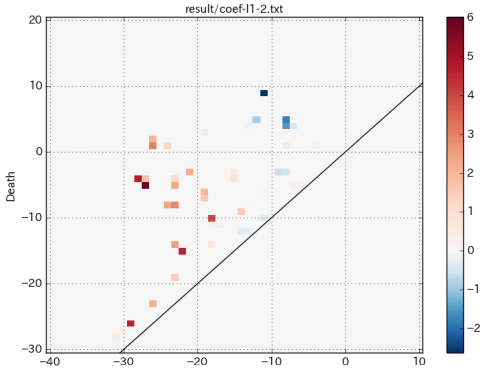
- **random images with parameters** $S = 0, \ldots, 9$
- **predict** S **from the learned PD**



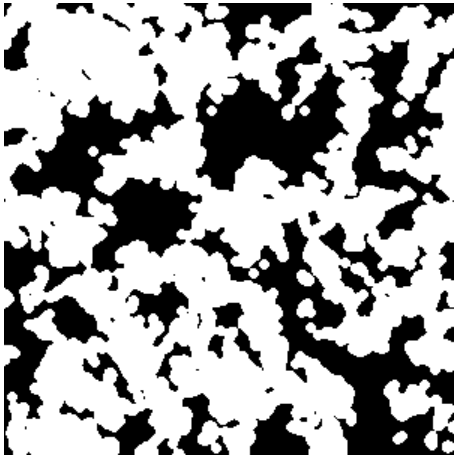


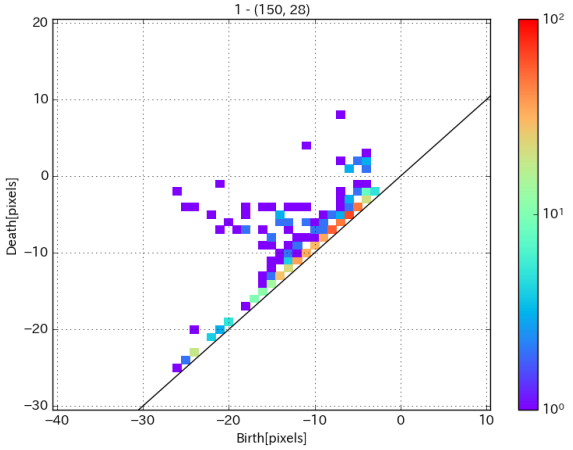


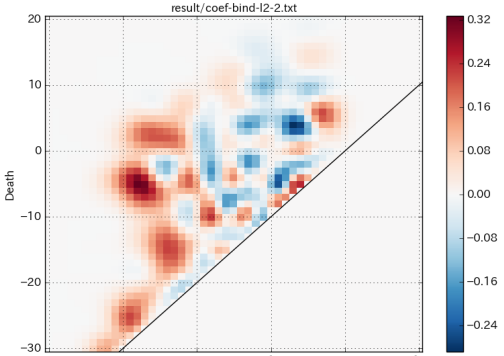
PI-RIDGE
(score=0.86)



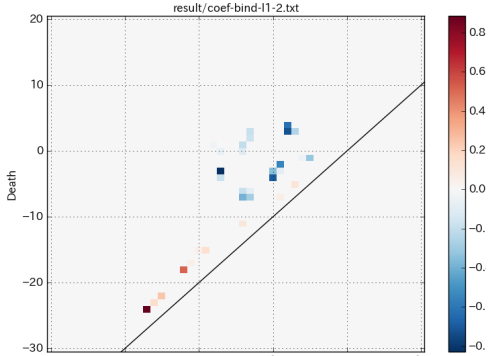
PI-LASSO
(score=0.86)







Both-RIDGE
(score=0.93)



Both-LASSO
(score=0.94)

THANK YOU