# Graphs in Machine Learning

## SSL Learnability

When Does Graph-Based SSL Provably Help?

Michal Valko

*Inria & ENS Paris-Saclay, MVA*

Partially based on material by: Branislav Kveton,
Mikhail Belkin, Jerry Zhu

# SSL with Graphs: What is behind it?

Why and when it helps?

# SSL with Graphs: What is behind it?

Why and when it helps?

Can we guarantee benefit of SSL over SL?

# SSL with Graphs: What is behind it?

Why and when it helps?

Can we guarantee benefit of SSL over SL?

Are there cases when **manifold** SSL is provably helpful?

# SSL with Graphs: What is behind it?

Why and when it helps?

Can we guarantee benefit of SSL over SL?

Are there cases when **manifold** SSL is provably helpful?

Say $\mathcal{X}$ is supported on manifold $\mathcal{M}$. Compare two cases:

- SL: does not know about $\mathcal{M}$ and only knows $(\mathbf{x}_i, y_i)$

# SSL with Graphs: What is behind it?

Why and when it helps?

Can we guarantee benefit of SSL over SL?

Are there cases when **manifold** SSL is provably helpful?

Say $\mathcal{X}$ is supported on manifold $\mathcal{M}$. Compare two cases:

- SL: does not know about $\mathcal{M}$ and only knows $(\mathbf{x}_i, y_i)$

- SSL: perfect knowledge of $\mathcal{M}$

# SSL with Graphs: What is behind it?

Why and when it helps?

Can we guarantee benefit of SSL over SL?

Are there cases when **manifold** SSL is provably helpful?

Say $\mathcal{X}$ is supported on manifold $\mathcal{M}$. Compare two cases:

- SL: does not know about $\mathcal{M}$ and only knows $(\mathbf{x}_i, y_i)$

- SSL: perfect knowledge of $\mathcal{M}$

# SSL with Graphs: What is behind it?

Why and when it helps?

Can we guarantee benefit of SSL over SL?

Are there cases when **manifold** SSL is provably helpful?

Say $\mathcal{X}$ is supported on manifold $\mathcal{M}$. Compare two cases:

- SL: does not know about $\mathcal{M}$ and only knows $(\mathbf{x}_i, y_i)$

- SSL: perfect knowledge of $\mathcal{M} \equiv$ humongous amounts of $\mathbf{x}_i$

http://people.cs.uchicago.edu/~niyogi/papersps/ssminimax2.pdf

## SSL with Graphs: What is behind it?

Set of learning problems - collections $\mathcal{P}$ of probability distributions:

$\mathcal{P}$

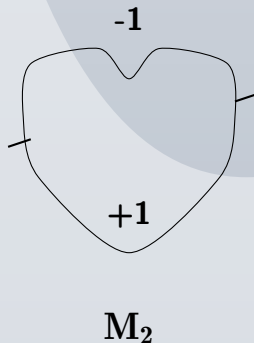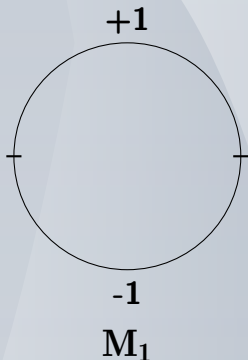## SSL with Graphs: What is behind it?

Set of learning problems - collections $\mathcal{P}$ of probability distributions:

$$\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$$

## SSL with Graphs: What is behind it?

Set of learning problems - collections $\mathcal{P}$ of probability distributions:

$$\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \cup_{\mathcal{M}} \{ p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M} \}$$

## SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M}\}$

## SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M}\}$

**Regression function** $m_p = \mathbb{E}[y|x]$ when $x \in \mathcal{M}$

# SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M}\}$

**Regression function** $m_p = \mathbb{E}[y | x]$ when $x \in \mathcal{M}$

**Algorithm** $A$ and **labeled examples** $\bar{z} = \{z_i\}_{i=1}^{n_l} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$

# SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}}$ is uniform on $\mathcal{M}\}$

**Regression function** $m_p = \mathbb{E}[y|x]$ when $x \in \mathcal{M}$

**Algorithm** $A$ and **labeled examples** $\bar{z} = \{z_i\}_{i=1}^{n_l} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$

**Minimax rate**

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{x}})} \right]$$

## SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M}\}$

**Regression function** $m_p = \mathbb{E}\left[y | x\right]$ when $x \in \mathcal{M}$

**Algorithm** $A$ and **labeled examples** $\bar{z} = \{z_i\}_{i=1}^{n_l} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$

**Minimax rate**

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

Since $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$

## SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M}\}$
**Regression function** $m_p = \mathbb{E}[y | x]$ when $x \in \mathcal{M}$
**Algorithm** $A$ and **labeled examples** $\bar{z} = \{z_i\}_{i=1}^{n_l} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$
**Minimax rate**

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

Since $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$

$$R(n_l, \mathcal{P}) = \inf_A \sup_{\mathcal{M}} \sup_{p \in \mathcal{P}_{\mathcal{M}}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

## SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M}\}$
**Regression function** $m_p = \mathbb{E}[y|x]$ when $x \in \mathcal{M}$
**Algorithm** $A$ and **labeled examples** $\bar{z} = \{z_i\}_{i=1}^{n_l} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$
**Minimax rate**

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

Since $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$

$$R(n_l, \mathcal{P}) = \inf_A \sup_{\mathcal{M}} \sup_{p \in \mathcal{P}_{\mathcal{M}}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

(SSL) When $A$ is allowed to know $\mathcal{M}$

## SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M}\}$

**Regression function** $m_p = \mathbb{E}[y|x]$ when $x \in \mathcal{M}$

**Algorithm** $A$ and **labeled examples** $\bar{z} = \{z_i\}_{i=1}^{n_l} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$

**Minimax rate**

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

Since $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$

$$R(n_l, \mathcal{P}) = \inf_A \sup_{\mathcal{M}} \sup_{p \in \mathcal{P}_{\mathcal{M}}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

(SSL) When $A$ is allowed to know $\mathcal{M}$

$$Q(n_l, \mathcal{P}) = \sup_{\mathcal{M}} \inf_A \sup_{p \in \mathcal{P}_{\mathcal{M}}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

## SSL with Graphs: What is behind it?

**Set of problems** $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_{\mathcal{X}} \text{ is uniform on } \mathcal{M}\}$
**Regression function** $m_p = \mathbb{E}[y|x]$ when $x \in \mathcal{M}$
**Algorithm** $A$ and **labeled examples** $\bar{z} = \{z_i\}_{i=1}^{n_l} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$
**Minimax rate**

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

Since $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$

$$R(n_l, \mathcal{P}) = \inf_A \sup_{\mathcal{M}} \sup_{p \in \mathcal{P}_{\mathcal{M}}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

(SSL) When $A$ is allowed to know $\mathcal{M}$

$$Q(n_l, \mathcal{P}) = \sup_{\mathcal{M}} \inf_A \sup_{p \in \mathcal{P}_{\mathcal{M}}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right]$$

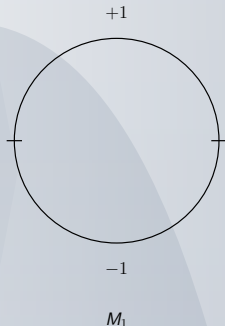In which cases there is a gap between $Q(n_l, \mathcal{P})$ and $R(n_l, \mathcal{P})$?

# SSL with Graphs: What is behind it?
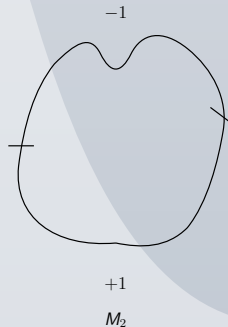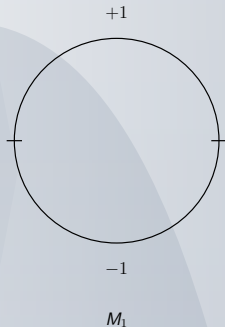
**Hypothesis space** $\mathcal{H}$: half of the circle as $+1$ and the rest as $-1$
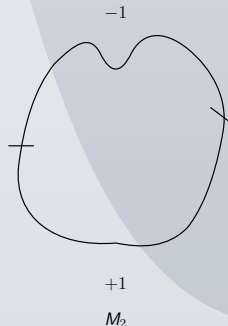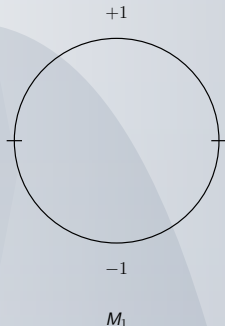
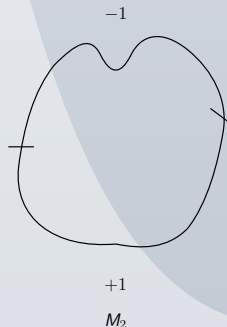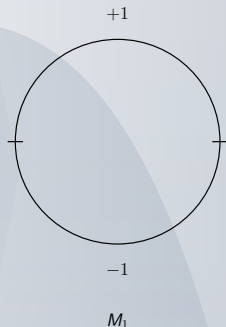# SSL with Graphs: What is behind it?

**Hypothesis space** $\mathcal{H}$: half of the circle as $+1$ and the rest as $-1$
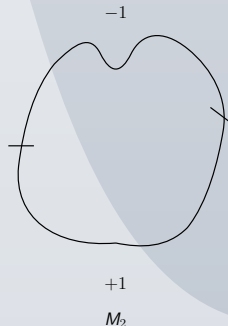


**Case 1:** $\mathcal{M}$ is known to the learner ($\mathcal{H}_{\mathcal{M}}$)

# SSL with Graphs: What is behind it?

**Hypothesis space** $\mathcal{H}$: half of the circle as $+1$ and the rest as $-1$



**Case 1:** $\mathcal{M}$ is known to the learner $(\mathcal{H}_{\mathcal{M}})$
What is a VC dimension of $\mathcal{H}_{\mathcal{M}}$?

# SSL with Graphs: What is behind it?

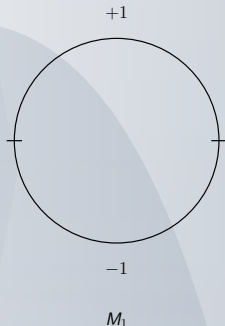**Hypothesis space** $\mathcal{H}$: half of the circle as $+1$ and the rest as $-1$



$M_1$

$M_2$

**Case 1:** $\mathcal{M}$ is known to the learner ($\mathcal{H}_{\mathcal{M}}$)
What is a VC dimension of $\mathcal{H}_{\mathcal{M}}$? 2

## SSL with Graphs: What is behind it?

**Hypothesis space** $\mathcal{H}$: half of the circle as $+1$ and the rest as $-1$



$M_1$

$M_2$
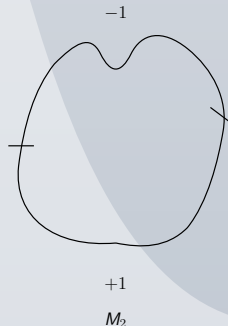
**Case 1:** $\mathcal{M}$ is known to the learner $(\mathcal{H}_\mathcal{M})$
What is a VC dimension of $\mathcal{H}_\mathcal{M}$? 2

## SSL with Graphs: What is behind it?

**Case 2:** $\mathcal{M}$ is **unknown** to the learner

## SSL with Graphs: What is behind it?

**Case 2:** $\mathcal{M}$ is **unknown** to the learner

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \| A(\bar{z}) - m_p \|_{L^2(p_{\mathbf{X}})} \right] =$$
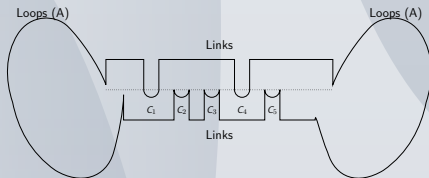
## SSL with Graphs: What is behind it?

**Case 2:** $\mathcal{M}$ is **unknown** to the learner

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right] = \Omega\left(1\right)$$
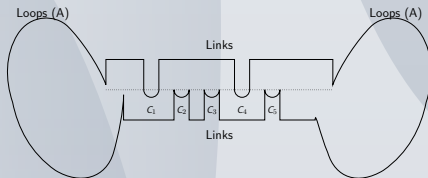
## SSL with Graphs: What is behind it?

**Case 2:** $\mathcal{M}$ is **unknown** to the learner

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_\mathbf{X})} \right] = \Omega(1)$$

We consider $2^d$ manifolds of the form

$$\mathcal{M} = \text{Loops} \cup \text{Links} \cup C \text{ where } C = \cup_{i=1}^d C_i$$

## SSL with Graphs: What is behind it?

**Case 2:** $\mathcal{M}$ is **unknown** to the learner

$$R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})} \right] = \Omega(1)$$
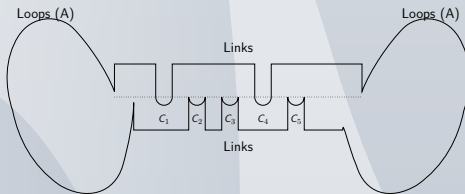
We consider $2^d$ manifolds of the form

$$\mathcal{M} = \text{Loops} \cup \text{Links} \cup C \text{ where } C = \cup_{i=1}^d C_i$$



**Main idea**: $d$ segments in $C$, $d - l$ with no data, $2^l$ possible
choices for labels, which helps us to lower bound
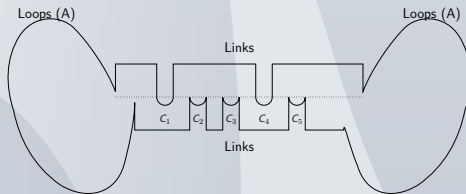$$\|A(\bar{z}) - m_p\|_{L^2(p_{\mathbf{X}})}$$

# SSL with Graphs: What is behind it?



**Knowing the manifold helps**

- $C_1$ and $C_4$ are close
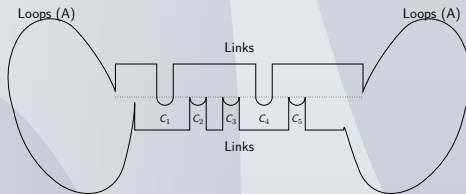
# SSL with Graphs: What is behind it?



**Knowing the manifold helps**

- $C_1$ and $C_4$ are close
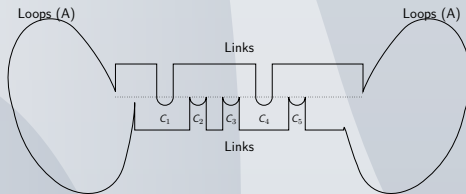- $C_1$ and $C_3$ are far

# SSL with Graphs: What is behind it?



**Knowing the manifold helps**

- $C_1$ and $C_4$ are close
- $C_1$ and $C_3$ are far
- we also need: **target function varies smoothly**

# SSL with Graphs: What is behind it?



**Knowing the manifold helps**

- $C_1$ and $C_4$ are close
- $C_1$ and $C_3$ are far
- we also need: **target function varies smoothly**
- altogether: **closeness on manifold $\rightarrow$ similarity in labels**

# SSL with Graphs: What is behind it?

What does it mean to **know** $\mathcal{M}$?

# SSL with Graphs: What is behind it?

What does it mean to **know** $\mathcal{M}$?

**Different degrees of knowing** $\mathcal{M}$

- set membership oracle: $\mathbf{x} \stackrel{?}{\in} \mathcal{M}$
- approximate oracle
- knowing the harmonic functions on $\mathcal{M}$
- knowing the Laplacian $\mathcal{L}_{\mathcal{M}}$
- knowing eigenvalues and eigen*functions*
- topological invariants, e.g., dimension
- metric information: geodesic distance

# Michal Valko

michal.valko@inria.fr

Inria & ENS Paris-Saclay, MVA

https://misovalko.github.io/mva-ml-graphs.html