Generalization bounds for learning under graph-dependence

Rui-Ray Zhang
rui.zhang@monash.edu
School of Mathematics, Monash University

Background on machine learning

- Given some input x, choose $f: x \mapsto y$ that performs well on unknown new data.

Background on machine learning

- Given some input x, choose $f : x \mapsto y$ that performs well on unknown new data.
- A training set S contains n data points $(x_i, y_i) \sim D$ (unknown).

Background on machine learning

- Given some input x, choose $f : x \mapsto y$ that performs well on unknown new data.
- A training set S contains n data points $(x_i, y_i) \sim D$ (unknown).
- A loss function measures error between true y and prediction $\widehat{y} = f(x)$

$$\ell : (y, f(x)) \mapsto \ell(y, f(x))$$

upper bounded by some $M \in \mathbb{R}_+$.

Background on machine learning

- Given some input x, choose $f : x \mapsto y$ that performs well on unknown new data.
- A training set S contains n data points $(x_i, y_i) \sim D$ (unknown).
- A loss function measures error between true y and prediction $\widehat{y} = f(x)$

$$\ell : (y, f(x)) \mapsto \ell(y, f(x))$$

  upper bounded by some $M \in \mathbb{R}_+$.
- Expected error: expected loss on new test data $(x, y) \sim D$ (unknown).

$$R(f) = \mathbb{E}[\ell(y, f(x))]$$

- Empirical error: average loss on given training data $(x_i, y_i)_{i=1}^n$.

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

Background on machine learning

- Given some input x, choose $f : x \mapsto y$ that performs well on unknown new data.
- A training set S contains n data points $(x_i, y_i) \sim D$ (unknown).
- A loss function measures error between true y and prediction $\widehat{y} = f(x)$

$$\ell : (y, f(x)) \mapsto \ell(y, f(x))$$

upper bounded by some $M \in \mathbb{R}_+$.

- Expected error: expected loss on new test data $(x, y) \sim D$ (unknown).

$$R(f) = \mathbb{E}[\ell(y, f(x))]$$

- Empirical error: average loss on given training data $(x_i, y_i)_{i=1}^n$.

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

- Goal is to establish generalisation error bounds

$$R(f) \leqslant \widehat{R}(f) + ?$$

Concentration inequalities

Concentration inequalities bounding the deviation of a function from its expectation

$$\mathbb{P}(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geqslant t)$$

are basic tools to establish generalization theory.

Concentration inequalities

Concentration inequalities bounding the deviation of a function from its expectation

$$\mathbb{P}\left(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geqslant t\right)$$

are basic tools to establish generalization theory. We choose

$$g = \mathbb{E}[\ell(y, f(x))] - \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$$

to be the difference of expected error and empirical error.

Bounded difference inequality

Definition (c-Lipschitz)

Given $\boldsymbol{c} = (c_1, \dots, c_n) \in \mathbb{R}_+^n$, a function g is $\boldsymbol{c}$-Lipschitz if

$$\left| g(x_1, \dots, x_i, \dots, x_n) - g(x_1, \dots, x_i', \dots, x_n) \right| \leqslant c_i.$$

Bounded difference inequality

Definition (c-Lipschitz)

Given $\boldsymbol{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n_+$, *a function* g *is* $\boldsymbol{c}$*-Lipschitz if*

$$\left| g(x_1, \ldots, x_i, \ldots, x_n) - g(x_1, \ldots, x_i', \ldots, x_n) \right| \leqslant c_i.$$

Theorem (Bounded difference inequality, McDiarmid 1989)

If we have that

1. g *is* $\boldsymbol{c}$*-Lipschitz,*
2. $\boldsymbol{X} = (X_1, \ldots, X_n)$ *is a vector of independent random variables,*

Bounded difference inequality

Definition (c-Lipschitz)

Given $\boldsymbol{c} = (c_1, \dots, c_n) \in \mathbb{R}_+^n$, a function g is $\boldsymbol{c}$-Lipschitz if

$$\left| g(x_1, \dots, x_i, \dots, x_n) - g(x_1, \dots, x_i', \dots, x_n) \right| \leq c_i.$$

Theorem (Bounded difference inequality, McDiarmid 1989)

If we have that

1. *g is $\boldsymbol{c}$-Lipschitz,*
2. *$\boldsymbol{X} = (X_1, \dots, X_n)$ is a vector of independent random variables,*

then for $t > 0$,

$$\mathbb{P}(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geq t) \leq \exp\left(-\frac{2t^2}{\|\boldsymbol{c}\|_2^2} \right).$$

This is also called Azuma-Hoeffding inequality.

Bounded difference inequality

Definition (c-Lipschitz)

Given $\boldsymbol{c} = (c_1, \dots, c_n) \in \mathbb{R}_+^n$, a function g is $\boldsymbol{c}$-Lipschitz if

$$\left| g(x_1, \dots, x_i, \dots, x_n) - g(x_1, \dots, x_i', \dots, x_n) \right| \leq c_i.$$

Theorem (Bounded difference inequality, McDiarmid 1989)

If we have that

1. *g is $\boldsymbol{c}$-Lipschitz,*
2. *$\boldsymbol{X} = (X_1, \dots, X_n)$ is a vector of independent random variables,*

then for $t > 0$,

$$\mathbb{P}(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geq t) \leq \exp\left(-\frac{2t^2}{\|\boldsymbol{c}\|_2^2} \right).$$

This is also called Azuma-Hoeffding inequality.

- If all $c_i = c$, then for $\delta \in (0,1)$, with probability at least $1 - \delta$, we have

$$f - \mathbb{E}[f] \leq \|\boldsymbol{c}\|_2 \sqrt{\frac{1}{2} \log\left(\frac{1}{\delta}\right)} = c\sqrt{\frac{n}{2} \log\left(\frac{1}{\delta}\right)}.$$

Dependent random variables

- Mixing coefficients: $\alpha/\beta/\phi$-mixing, etc.
  - quantitatively measure the dependencies, and widely used in probability, statistics, etc.

$$\alpha(s) = \sup\left\{\left|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\right| : A \in \sigma\left(\{X_i\}_{-\infty}^{t}\right), B \in \sigma\left(\{X_i\}_{t+s}^{\infty}\right)\right\}$$

Dependent random variables

- Mixing coefficients: $\alpha/\beta/\phi$-mixing, etc.
  - quantitatively measure the dependencies, and widely used in probability, statistics, etc.

$$\alpha(s) = \sup\left\{\left|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\right| : A \in \sigma\left(\{X_i\}_{-\infty}^{t}\right), B \in \sigma\left(\{X_i\}_{t+s}^{\infty}\right)\right\}$$

- Dependency graphs: combinatorial, relate to independent sets, degrees, cumulants, etc.

Dependent random variables

- Mixing coefficients: $\alpha/\beta/\phi$-mixing, etc.
  - quantitatively measure the dependencies, and widely used in probability, statistics, etc.

$$\alpha(s) = \sup\left\{\left|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\right| : A \in \sigma\left(\{X_i\}_{-\infty}^{t}\right), B \in \sigma\left(\{X_i\}_{t+s}^{\infty}\right)\right\}$$

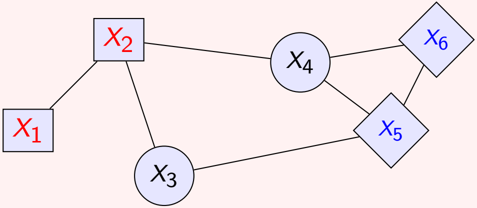
- Dependency graphs: combinatorial, relate to independent sets, degrees, cumulants, etc.
- Copula, graphical models (random field, Bayesian network, etc.), time series, etc.

Dependency Graphs

Definition (Dependency Graphs)

Graph G is a dependency graph for random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ if

- *Vertex set $V(G) = [n] = \{1, \ldots, n\}$.*

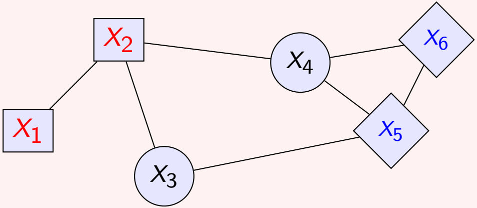


Dependency Graphs

Definition (Dependency Graphs)

Graph G is a dependency graph for random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ if

- *Vertex set $V(G) = [n] = \{1, \ldots, n\}$.*



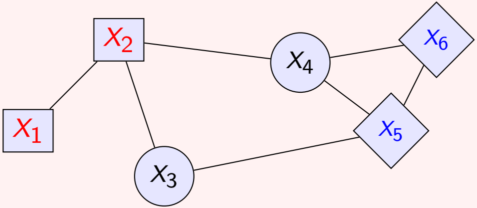
- *If disjoint subsets $I, J \subset [n]$ are non-adjacent in G, then $\{X_i\}_{i \in I}$ and $\{X_j\}_{j \in J}$ are independent.*

Dependency Graphs

Definition (Dependency Graphs)

Graph G is a dependency graph for random variables $\boldsymbol{X} = (X_1, \dots, X_n)$ if

- *Vertex set $V(G) = [n] = \{1, \dots, n\}$.*



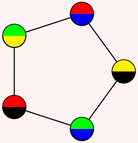
- *If disjoint subsets $I, J \subset [n]$ are non-adjacent in G, then $\{X_i\}_{i \in I}$ and $\{X_j\}_{j \in J}$ are independent.*
  - *In the above example, $\{X_1, X_2\}$ and $\{X_5, X_6\}$ are independent.*

- The dependency graph for a set of random variables is not necessarily unique.

Idea: to utilise independence among variables

Given a graph G with n vertices, a fractional vertex covering $\{(I_j, w_j)\}_j$ of G satisfies

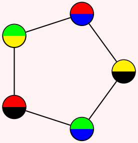
1. each $I_j \subseteq [n]$ is an independent set (no two vertices are adjacent),
2. $\sum_{j: v \in I_j} w_j = 1$ for each vertex.



Idea: to utilise independence among variables

Given a graph G with n vertices, a fractional vertex covering $\{(I_j, w_j)\}_j$ of G satisfies

1. each $I_j \subseteq [n]$ is an independent set (no two vertices are adjacent),
2. $\sum_{j:v\in I_j} w_j = 1$ for each vertex.



A function g is *decomposable* $\boldsymbol{c}$*-Lipschitz* with respect to graph G if there exist $(c_i)_{i\in I_j}$-Lipschitz functions $\{g_j\}_j$ such that

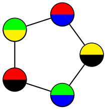
$$g(\mathrm{x}) = \sum_j w_j g_j(\mathrm{x}_{I_j}),$$

for all $\mathrm{x} = (x_1, \ldots, x_n)$, and for all fractional vertex covers $\{(I_j, w_j)\}_j$ of G.

- Summation is decomposable $\boldsymbol{c}$-Lipschitz.

Given a graph G on n vertices, a fractional vertex covering $\{(I_j, w_j)\}_j$ of G satisfies

1. each $I_j \subseteq [n]$ is an independent set (no two vertices are adjacent),
2. $\sum_{j: v \in I_j} w_j = 1$ for each vertex.



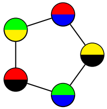
Theorem (Usunier et al. NIPS05, Z, Amini 2022+)

If we have that

1. *g is decomposable $\boldsymbol{c}$-Lipschitz,*
2. *$\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent,*

Given a graph G on n vertices, a fractional vertex covering $\{(I_j, w_j)\}_j$ of G satisfies

1. each $I_j \subseteq [n]$ is an independent set (no two vertices are adjacent),
2. $\sum_{j:v\in I_j} w_j = 1$ for each vertex.



Theorem (Usunier et al. NIPS05, Z, Amini 2022+)

If we have that

1. *g is decomposable $\boldsymbol{c}$-Lipschitz,*
2. *$\boldsymbol{X} = (X_1, \dots, X_n)$ is G-dependent,*

then for $t > 0$,

$$\mathbb{P}(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geqslant t) \leqslant \exp\left(-\frac{2t^2}{\chi^\star(G)\|\boldsymbol{c}\|_2^2}\right),$$

where $\chi^\star(G) = \sum_j w_j \leqslant \Delta(G) + 1$.

- In the above example, $\chi^\star(G) = 5/2$.
- Janson (2004) proved the case of summation.

Forest-dependent random variables

Theorem (Zhang et al., 2019)

If we have that

1. g *is* $\boldsymbol{c}$*-Lipschitz,*
2. F *is a dependency graph for* $\boldsymbol{X}$*, where* $F = \{T_i\}_{i \in [k]}$ *is a forest,*

Forest-dependent random variables

Theorem (Zhang et al., 2019)

If we have that

1. g *is* $\boldsymbol{c}$*-Lipschitz,*
2. F *is a dependency graph for* $\boldsymbol{X}$*, where* $F = \{T_i\}_{i \in [k]}$ *is a forest,*

then for $t > 0$*,*

$$\mathbb{P}(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geqslant t) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{k} c_{\min,i}^2 + \sum_{\{i,j\} \in E(F)} (c_i + c_j)^2}\right),$$

where $c_{\min,i} = \min\{c_j : j \in V(T_i)\}$ *is the minimum entry of* $\boldsymbol{c}$ *in each tree* T_i*.*

Forest-dependent random variables

Theorem (Zhang et al., 2019)

If we have that

1. *g is $\boldsymbol{c}$-Lipschitz,*
2. *F is a dependency graph for $\boldsymbol{X}$, where $F = \{T_i\}_{i \in [k]}$ is a forest,*

then for $t > 0$,

$$\mathbb{P}(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geqslant t) \leqslant \exp\left(- \frac{2t^2}{\sum_{i=1}^{k} c_{\min,i}^2 + \sum_{\{i,j\} \in E(F)} (c_i + c_j)^2} \right),$$

where $c_{\min,i} = \min\{c_j : j \in V(T_i)\}$ is the minimum entry of $\boldsymbol{c}$ in each tree T_i.

- General graphs can be handled via tree-partitions (transforming a graph to a forest by merging vertices).

Theorem (Janson, 2004)

Let random variables $\{X_i\}_{i\in V(G)}$ be G-dependent such that every X_i takes values in an interval of length $c_i \geqslant 0$. Then, for every $t > 0$,

$$\mathbb{P}\left(\sum_{i\in V(G)} X_i - \mathbb{E}\left[\sum_{i\in V(G)} X_i\right] \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\chi_f(G)\sum_{i\in V(G)} c_i^2}\right).$$

Theorem (Zhang, 2022)

Under the same setting,

$$\mathbb{P}\left(\sum_{i\in V(G)} X_i - \mathbb{E}\left[\sum_{i\in V(G)} X_i\right] \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{D}\right),$$

where D is optimised over fractional forest vertex covers of G.

- *It is no worse than Janson's, better when G is sparse.*

Theorem (Janson, 2004)

Let random variables $\{X_i\}_{i\in V(G)}$ be G-dependent such that every X_i takes values in an interval of length $c_i \geqslant 0$. Then, for every $t > 0$,

$$\mathbb{P}\left(\sum_{i\in V(G)} X_i - \mathbb{E}\left[\sum_{i\in V(G)} X_i\right] \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\chi_f(G)\sum_{i\in V(G)} c_i^2}\right).$$

Theorem (Zhang, 2022)

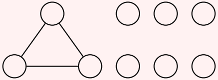
Under the same setting,

$$\mathbb{P}\left(\sum_{i\in V(G)} X_i - \mathbb{E}\left[\sum_{i\in V(G)} X_i\right] \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{D}\right),$$

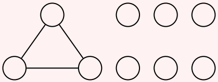
where D is optimised over fractional forest vertex covers of G.

- *It is no worse than Janson's, better when G is sparse.*
- *It generalises to certain decomposable Lipschitz functions, extending McDiarmid's.*

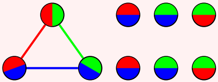
Random indicators $\{X_i\}_{i\in[9]}$ with dependency graph



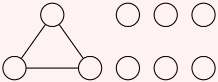
Random indicators $\{X_i\}_{i\in[9]}$ with dependency graph



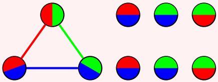
- Fractional forest vertex covers of G: covering vertices with weighted (induced) forests such that the sum of weights for each vertex equals 1 (related to fractional vertex-arboricity).



Random indicators $\{X_i\}_{i \in [9]}$ with dependency graph



- Fractional forest vertex covers of G: covering vertices with weighted (induced) forests such that the sum of weights for each vertex equals 1 (related to fractional vertex-arboricity).

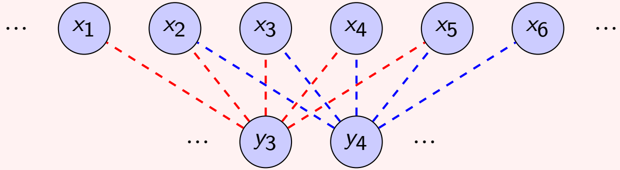


- Janson's bound vs. the new one:

$$\mathbb{P}(X - \mathbb{E}[X] \geqslant t) \leqslant \exp\left(-\frac{2t^2}{27}\right), \qquad \mathbb{P}(X - \mathbb{E}[X] \geqslant t) \leqslant \exp\left(-\frac{8t^2}{81}\right).$$

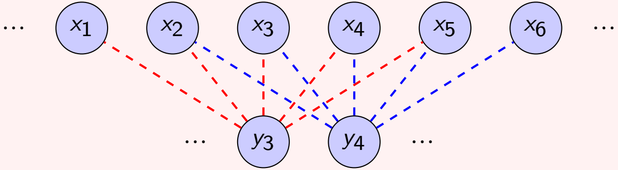
Example

- y_i: observation at location i, e.g., house price
- x_i: random variable modelling influential factors at location i



Example

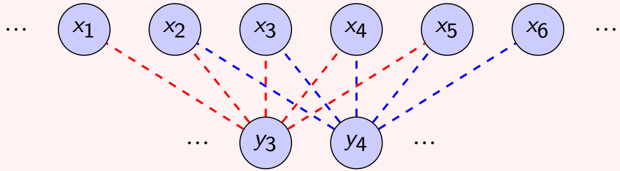
- y_i: observation at location i, e.g., house price
- x_i: random variable modelling influential factors at location i



- Given training data: $S = \{\ldots, ((x_1, x_2, x_3, x_4, x_5), y_3), ((x_2, x_3, x_4, x_5, x_6), y_4), \ldots\}$
- Find $f : (x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}) \mapsto y_i$

Example

- y_i: observation at location i, e.g., house price
- x_i: random variable modelling influential factors at location i



- Given training data: $S = \{\ldots, ((x_1, x_2, x_3, x_4, x_5), y_3), ((x_2, x_3, x_4, x_5, x_6), y_4), \ldots\}$
- Find $f : (x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}) \mapsto y_i$

Definition (Hoeffding and Robbins 1948)

A sequence of random variables $(X_i)_{i=1}^n$ is m-dependent for some $m \geqslant 1$ if $(X_j)_{j=1}^i$ and $(X_j)_{j=i+m+1}^n$ are independent for all $i > 0$.

Stability bound for learning m-dependent data

Given a sample S, a learning algorithm $\mathscr{A} : \mathsf{S} \mapsto f_{\mathsf{S}}^{\mathscr{A}}$ outputs $f_{\mathsf{S}}^{\mathscr{A}}$.

Definition (Uniform stability, Bousquet and Elisseeff 2002)

A learning algorithm $\mathscr{A}$ is β_n-uniformly stable if

$$\max_{i \in [n]} \left| \ell(y, f_{\mathsf{S}}^{\mathscr{A}}(x)) - \ell(y, f_{\mathsf{S}^{\setminus i}}^{\mathscr{A}}(x)) \right| \leqslant \beta_n,$$

where $\mathsf{S}^{\setminus i}$ denotes S with i-th data point removed.

We have

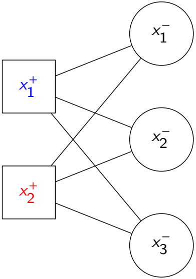
$$R(f_{\mathsf{S}}^{\mathscr{A}}) \leqslant \widehat{R}(f_{\mathsf{S}}^{\mathscr{A}}) + 2\beta_{n,2m}(2m+1) + (4n\beta_n + M)\sqrt{\frac{2m}{n} \log\left(\frac{1}{\delta}\right)},$$

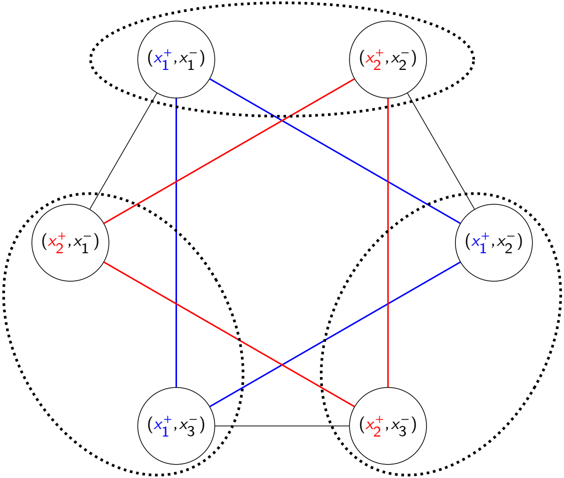
which introduces a factor $4m$ comparing with the independent case (Bousquet and Elisseeff 2002)

$$R(f_{\mathsf{S}}^{\mathscr{A}}) \leqslant \widehat{R}(f_{\mathsf{S}}^{\mathscr{A}}) + 2\beta_n + (4n\beta_n + M)\sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}.$$

Bipartite ranking

- Training set: $T = (x_i, y_i)_{1 \leqslant i \leqslant m}$ with $y_i \in \{-1, +1\}$.
- The goal: to find a scoring function h that gives higher scores to instances of the positive class than the ones of the negative class.
- For $(x, y), (x', y')$ with $y \neq y'$, we consider the unordered pairs of examples (x, x').





Bipartite ranking

- Let

$$\mathsf{S} = \{(x, x') \in T \times T \mid y \neq y'\}$$

denote the unordered pairs of examples from different classes in T.

- The empirical loss of a scoring function h over T can be written as a sum over S:

$$\widehat{R}(h) = \frac{1}{|\mathsf{S}|} \sum_{(x,x') \in \mathsf{S}} \mathbb{1}_{\{z_{x,x'}(h(x)-h(x')) \leqslant 0\}},$$

where $z_{x,x'} = 2\mathbb{1}_{\{y-y'>0\}} - 1$.

  - If $y = 1$ and $y' = -1$, then $z_{x,x'}(h(x) - h(x')) = h(x) - h(x')$.

An approach based on fractional Rademacher complexity gives the following.

Corollary

Let T be a training set composed of m_+ positive instances and m_- negative ones. Then for any scoring functions in $\{h : (x, x') \mapsto \langle \mathrm{w}, \phi(x) - \phi(x') \rangle ; \|\mathrm{w}\| \leqslant B\}$, where ϕ is a feature mapping with bounded norm, such that $\forall (x, x'), \|\phi(x) - \phi(x')\| \leqslant \Gamma$, and for any $\delta \in (0,1)$ with probability at least $1-\delta$, we have

$$R(f) \leqslant \widehat{R}(f) + \frac{4B\Gamma}{\sqrt{m}} + 3\sqrt{\frac{1}{2m}\log\left(\frac{2}{\delta}\right)},$$

where $m = \min(m_-, m_+)$.

The content is based upon

1. *McDiarmid-type Inequalities for Graph-dependent Variables and Stability Bounds*
   (with Xingwu Liu, Yuyi Wang, Liwei Wang)
   Spotlight in Advances in Neural Information Processing Systems 32 (NeurIPS 2019)
2. *When Janson meets McDiarmid: Bounded difference inequalities under graph-dependence*
   Statistics & Probability Letters, 2022
3. *Generalization bounds for learning under graph-dependence: A survey*
   (with Massih-Reza Amini, arXiv:2203.13534)

Thanks for your attention!