Concentration and generalization for learning under graph-dependence

Rui-Ray Zhang
rui.zhang@bse.eu
Barcelona School of Economics

Background on machine learning

- The goal is, for some input (x, y), to choose $f : x \mapsto \widehat{y}$ such that $\widehat{y}$ is closed to y.

Background on machine learning

- The goal is, for some input (x, y), to choose $f : x \mapsto \widehat{y}$ such that $\widehat{y}$ is closed to y.
- A training set $\boldsymbol{S}$ contains n data points $(x_i, y_i) \sim D$ (unknown).

Background on machine learning

- The goal is, for some input (x,y), to choose $f : x \mapsto \hat{y}$ such that $\hat{y}$ is closed to y.
- A training set $\boldsymbol{S}$ contains n data points $(x_i, y_i) \sim D$ (unknown).
- A loss function measures error between true y and prediction $\hat{y} = f(x)$

$$\ell : (y, \hat{y}) \mapsto \ell(y, f(x)),$$

Background on machine learning

- The goal is, for some input (x, y), to choose $f : x \mapsto \widehat{y}$ such that $\widehat{y}$ is closed to y.
- A training set $\boldsymbol{S}$ contains n data points $(x_i, y_i) \sim D$ (unknown).
- A loss function measures error between true y and prediction $\widehat{y} = f(x)$

$$\ell : (y, \widehat{y}) \mapsto \ell(y, f(x)),$$

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

Background on machine learning

- The goal is, for some input (x, y), to choose $f : x \mapsto \widehat{y}$ such that $\widehat{y}$ is closed to y.
- A training set $\boldsymbol{S}$ contains n data points $(x_i, y_i) \sim D$ (unknown).
- A loss function measures error between true y and prediction $\widehat{y} = f(x)$

$$\ell : (y, \widehat{y}) \mapsto \ell(y, f(x)),$$

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

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

Background on machine learning

- The goal is, for some input (x,y), to choose $f : x \mapsto \widehat{y}$ such that $\widehat{y}$ is closed to y.
- A training set $\boldsymbol{S}$ contains n data points $(x_i, y_i) \sim D$ (unknown).
- A loss function measures error between true y and prediction $\widehat{y} = f(x)$

$$\ell : (y, \widehat{y}) \mapsto \ell(y, f(x)),$$

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

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

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

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

Background on machine learning

- The goal is, for some input (x,y), to choose $f : x \mapsto \widehat{y}$ such that $\widehat{y}$ is closed to y.
- A training set $\boldsymbol{S}$ contains n data points $(x_i, y_i) \sim D$ (unknown).
- A loss function measures error between true y and prediction $\widehat{y} = f(x)$

$$\ell : (y, \widehat{y}) \mapsto \ell(y, f(x)),$$

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

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

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

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

- Generalisation error bounds bound expected error using empirical error:

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

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

- measuring of the complexity of the output hypothesis space,

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

- measuring of the complexity of the output hypothesis space,
  - VC theory (Vapnik and Chervonenkis 1971).
  - Rademacher complexity (Bartlett and Mendelson 2002).

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

- measuring of the complexity of the output hypothesis space,
  - VC theory (Vapnik and Chervonenkis 1971).
  - Rademacher complexity (Bartlett and Mendelson 2002).
- exploiting properties of the learning algorithm,

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

- measuring of the complexity of the output hypothesis space,
  - VC theory (Vapnik and Chervonenkis 1971).
  - Rademacher complexity (Bartlett and Mendelson 2002).
- exploiting properties of the learning algorithm,
  - Algorithmic stability (Bousquet and Elisseeff 2002).
  - PAC-Bayesian bounds (McAllester, 1999).

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

- measuring of the complexity of the output hypothesis space,
  - VC theory (Vapnik and Chervonenkis 1971).
  - Rademacher complexity (Bartlett and Mendelson 2002).
- exploiting properties of the learning algorithm,
  - Algorithmic stability (Bousquet and Elisseeff 2002).
  - PAC-Bayesian bounds (McAllester, 1999).
- considering mutual information between training sample and the output hypothesis of the learning algorithm,
  - Mutual information bound (Russo and Zou 2016).

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

- measuring of the complexity of the output hypothesis space,
  - VC theory (Vapnik and Chervonenkis 1971).
  - Rademacher complexity (Bartlett and Mendelson 2002).
- exploiting properties of the learning algorithm,
  - Algorithmic stability (Bousquet and Elisseeff 2002).
  - PAC-Bayesian bounds (McAllester, 1999).
- considering mutual information between training sample and the output hypothesis of the learning algorithm,
  - Mutual information bound (Russo and Zou 2016).
- or convex analysis: (Lugosi and Neu 2022).

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

- measuring of the complexity of the output hypothesis space,
  - VC theory (Vapnik and Chervonenkis 1971).
  - Rademacher complexity (Bartlett and Mendelson 2002).
- exploiting properties of the learning algorithm,
  - Algorithmic stability (Bousquet and Elisseeff 2002).
  - PAC-Bayesian bounds (McAllester, 1999).
- considering mutual information between training sample and the output hypothesis of the learning algorithm,
  - Mutual information bound (Russo and Zou 2016).
- or convex analysis: (Lugosi and Neu 2022).

Most of them assume that samples are i.i.d.

Background on machine learning

The ways to establish generalisation error (also called generalization gap) bounds are by

- measuring of the complexity of the output hypothesis space,
  - VC theory (Vapnik and Chervonenkis 1971).
  - Rademacher complexity (Bartlett and Mendelson 2002).
- exploiting properties of the learning algorithm,
  - Algorithmic stability (Bousquet and Elisseeff 2002).
  - PAC-Bayesian bounds (McAllester, 1999).
- considering mutual information between training sample and the output hypothesis of the learning algorithm,
  - Mutual information bound (Russo and Zou 2016).
- or convex analysis: (Lugosi and Neu 2022).

Most of them assume that samples are i.i.d., which is false in many (if not all) settings.

Concentration inequalities

Let $\boldsymbol{X}$ be a vector of i.i.d. random variables.

Concentration inequalities

Let $\boldsymbol{X}$ be a vector of i.i.d. random variables. Concentration inequalities bound the probability of deviation

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

Concentration inequalities

Let $\boldsymbol{X}$ be a vector of i.i.d. random variables. Concentration inequalities bound the probability of deviation

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

and are the basic tools to establish generalization theory

Concentration inequalities

Let $\boldsymbol{X}$ be a vector of i.i.d. random variables. Concentration inequalities bound the probability of deviation

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

and are the basic tools to establish generalization theory, in which

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

is the difference of expected error and empirical error.

Bounded difference inequality

c-Lipschitz

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

$$\left| g(x_1, \ldots, \mathrm{x_i}, \ldots, x_n) - g(x_1, \ldots, \mathrm{x_i'}, \ldots, x_n) \right| \leq \mathrm{c_i}.$$

Bounded difference inequality

c-Lipschitz

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

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

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

c-Lipschitz

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

$$\left| g(x_1, \ldots, \mathrm{x_i}, \ldots, x_n) - g(x_1, \ldots, \mathrm{x_i'}, \ldots, x_n) \right| \leq \mathrm{c_i}.$$

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,

then for $t > 0$,

$$\mathbb{P}\left(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i \in [n]} c_i^2}\right).$$

This is also called Azuma-Hoeffding inequality.

Bounded difference inequality

c-Lipschitz

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

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

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,

then for $t > 0$,

$$\mathbb{P}\left(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})] \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i \in [n]} c_i^2}\right).$$

This is also called Azuma-Hoeffding inequality.

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

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

Dependent random variables

- Mixing coefficients: α, β, ϕ-mixing, etc.
  - quantitatively measure the dependencies, and widely used in probability, statistics, e.g.,

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

Dependent random variables

- Mixing coefficients: α, β, ϕ-mixing, etc.
  - quantitatively measure the dependencies, and widely used in probability, statistics, e.g.,

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

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

Dependent random variables

- Mixing coefficients: α, β, ϕ-mixing, etc.
  - quantitatively measure the dependencies, and widely used in probability, statistics, e.g.,

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

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

Dependency Graphs

Definition (G-dependent variables)

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

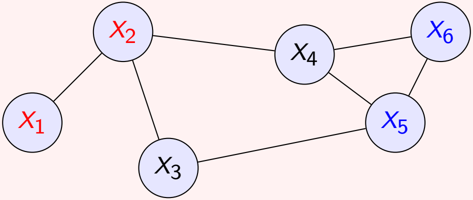
1. *Vertex set $V(G) = [n] = \{1, \ldots, n\}$.*
2. *If disjoint subsets $I, J \subset [n]$ are non-adjacent in G, then $\{X_i\}_{i \in I}$ and $\{X_j\}_{j \in J}$ are independent.*

Dependency Graphs

Definition (G-dependent variables)

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

1. *Vertex set $V(G) = [n] = \{1, \ldots, n\}$.*
2. *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.*



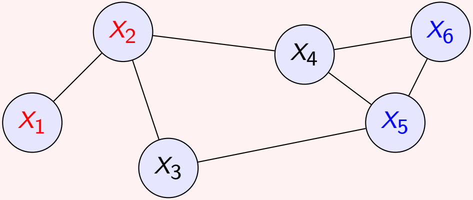
In the above example, $\{X_1, X_2\}$ and $\{X_5, X_6\}$ are independent.

Dependency Graphs

Definition (G-dependent variables)

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

1. *Vertex set $V(G) = [n] = \{1, \ldots, n\}$.*
2. *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.

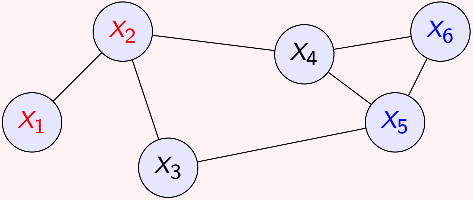
- The dependency graph for a set of random variables is not necessarily unique.

Dependency Graphs

Definition (G-dependent variables)

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

1. *Vertex set $V(G) = [n] = \{1, \dots, n\}$.*
2. *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.

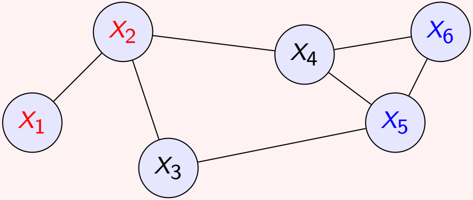
Question: is this a good model?

Dependency Graphs

Definition (G-dependent variables)

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

1. *Vertex set $V(G) = [n] = \{1, \ldots, n\}$.*
2. *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.

Question: is this a good model?

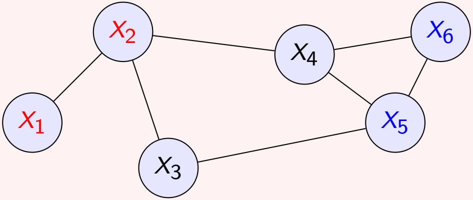
"All models are wrong, but some are useful". – George Box.

Dependency Graphs

Definition (G-dependent variables)

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

1. *Vertex set $V(G) = [n] = \{1, \dots, n\}$.*
2. *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.

Question: is this a good model?

"All models are wrong, but some are useful". – George Box.
This model has deep connections to cumulant, cluster expansion, Stein's method, etc.

Concentration under dependence

Janson, 2004

If we have that

1. $\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent, and every X_i is in an interval of length $c_i \geqslant 0$,

Concentration under dependence

Janson, 2004

If we have that

1. $\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent, and every X_i is in an interval of length $c_i \geqslant 0$,
2. the function is a summation,

then, for every $t > 0$,

Concentration under dependence

Janson, 2004

If we have that

1. $\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent, and every X_i is in an interval of length $c_i \geqslant 0$,
2. the function is a summation,

then, for every $t > 0$,

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

Concentration under dependence

Janson, 2004

If we have that

1. $\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent, and every X_i is in an interval of length $c_i \geqslant 0$,
2. the function is a summation,

then, for every $t > 0$,

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

where $\chi_f(G)$ is the fractional chromatic number of G, and $\chi_f(G) \leqslant \Delta(G) + 1$.

Concentration under dependence

Janson, 2004

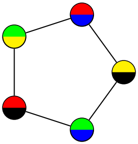
If we have that

1. $\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent, and every X_i is in an interval of length $c_i \geqslant 0$,
2. the function is a summation,

then, for every $t > 0$,

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

where $\chi_f(G)$ is the fractional chromatic number of G, and $\chi_f(G) \leqslant \Delta(G) + 1$.



Janson's idea: to utilise independence among variables

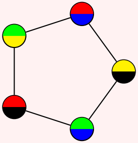
Given a graph G with n vertices, a fractional vertex covering $\{(I_j, w_j)\}_j$ of G satisfies

1. each $I_j \subseteq [n]$ is an independent set (no two vertices are adjacent),

Janson's 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.

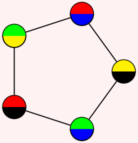


- In the above example, $\chi_f(C_5) = 5/2 \leqslant \Delta(C_5) + 1 = 3$.

Janson's 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.



- In the above example, $\chi_f(C_5) = 5/2 \leqslant \Delta(C_5) + 1 = 3$.

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

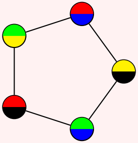
$$g(\boldsymbol{x}) = \sum_j w_j g_j(\mathrm{x}_{I_j}),$$

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

Janson's 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.



- In the above example, $\chi_f(C_5) = 5/2 \leqslant \Delta(C_5) + 1 = 3$.

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

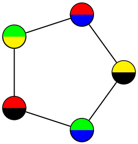
$$g(\boldsymbol{x}) = \sum_j w_j g_j(\mathrm{x}_{I_j}),$$

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 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.



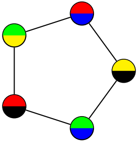
Theorem (Usunier-Amini-Gallinari 2005; Z. 2022)

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 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.



Theorem (Usunier-Amini-Gallinari 2005; Z. 2022)

If we have that

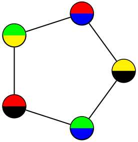
1. *g is decomposable $\boldsymbol{c}$-Lipschitz,*
2. *$\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent,*

then for $t > 0$,

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

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.



Theorem (Usunier-Amini-Gallinari 2005; Z. 2022)

If we have that

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

then for $t > 0$,

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

Forest-dependent random variables

Theorem (Z. Liu, Wang, Wang 2019)

If we have that

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

Forest-dependent random variables

Theorem (Z. Liu, Wang, Wang 2019)

If we have that

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

then for $t > 0$,

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

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 (Z. Liu, Wang, Wang 2019)

If we have that

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

then for $t > 0$,

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

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 are handled via tree-partitions (transforming a graph to a forest by merging vertices).

Cramér-Chernoff method

c-Lipschitz + independence

Cramér-Chernoff method

$$\begin{aligned} & \boldsymbol{c}\text{-Lipschitz} + \text{independence} \\ \Rightarrow \quad & \sup_{\alpha \in \Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha \right] - \inf_{\beta \in \Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta \right] \leqslant c_i \end{aligned}$$

Cramér-Chernoff method

$\boldsymbol{c}$-Lipschitz + independence

$$\Rightarrow \quad \sup_{\alpha \in \Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha\right] - \inf_{\beta \in \Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta\right] \leqslant c_i$$

$$\Rightarrow \quad M_i = \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i]} = \mathrm{x}_{[i]}\right] - \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}\right] \leqslant c_i$$

Cramér-Chernoff method

$$
\begin{aligned}
& \boldsymbol{c}\text{-Lipschitz} + \text{independence} \\
\Rightarrow\quad & \sup_{\alpha\in\Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha\right] - \inf_{\beta\in\Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta\right] \leqslant c_i \\
\Rightarrow\quad & M_i = \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i]} = \mathrm{x}_{[i]}\right] - \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}\right] \leqslant c_i \\
\Rightarrow\quad & \mathbb{E}\left[\exp\left(s(g - \mathbb{E}[g])\right)\right] = \mathbb{E}\left[\exp\left(s \sum_{i\in[n]} M_i\right)\right] \leqslant \exp\left(\frac{s^2}{8}\sum_{i=1}^{n} c_i^2\right)
\end{aligned}
$$

Cramér-Chernoff method

$$
\begin{aligned}
& \boldsymbol{c}\text{-Lipschitz} + \text{independence} \\
\Rightarrow \quad & \sup_{\alpha\in\Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha\right] - \inf_{\beta\in\Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta\right] \leqslant c_i \\
\Rightarrow \quad & M_i = \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i]} = \mathrm{x}_{[i]}\right] - \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}\right] \leqslant c_i \\
\Rightarrow \quad & \mathbb{E}\left[\exp\left(s(g - \mathbb{E}[g])\right)\right] = \mathbb{E}\left[\exp\left(s \sum_{i\in[n]} M_i\right)\right] \leqslant \exp\left(\frac{s^2}{8}\sum_{i=1}^{n} c_i^2\right) \\
\Rightarrow \quad & \mathbb{P}\left(g - \mathbb{E}[g] \geqslant t\right) \leqslant \inf_{s>0}\left(\frac{\mathbb{E}[\exp(s(g-\mathbb{E}[g]))]}{e^{st}}\right) = \exp\left(-\frac{2t^2}{\|\boldsymbol{c}\|_2^2}\right)
\end{aligned}
$$

Proof Sketch

- Choose c_{min} as root, expose vertices via topological ordering, i.e., child i before parent $p(i)$. We will show that

$$\sup_{\alpha \in \Omega_i} \mathbb{E}\left[g(\boldsymbol{X}) \middle| \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha\right] - \inf_{\beta \in \Omega_i} \mathbb{E}\left[g(\boldsymbol{X}) \middle| \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta\right] \leqslant c_i + c_{p(i)},$$

Proof Sketch

- Choose c_{min} as root, expose vertices via topological ordering, i.e., child i before parent $p(i)$. We will show that

$$\sup_{\alpha \in \Omega_i} \mathbb{E}\left[g(\boldsymbol{X}) \middle| \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha \right] - \inf_{\beta \in \Omega_i} \mathbb{E}\left[g(\boldsymbol{X}) \middle| \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta \right] \leqslant c_i + c_{p(i)},$$

- by choosing a suitable coupling of conditional distributions

Proof Sketch

- Choose $c_{\min}$ as root, expose vertices via topological ordering, i.e., child i before parent $p(i)$. We will show that

$$\sup_{\alpha \in \Omega_i} \mathbb{E}\left[g(\boldsymbol{X}) \middle| \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha\right] - \inf_{\beta \in \Omega_i} \mathbb{E}\left[g(\boldsymbol{X}) \middle| \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta\right] \leqslant c_i + c_{p(i)},$$

- by choosing a suitable coupling of conditional distributions , i.e., a joint distribution $\mathbb{P}$ of $(\boldsymbol{X}_{[i+1:n]}, \widetilde{\boldsymbol{X}}_{[i+1:n]})$ with desirable marginal distributions and with few different bits

$$\mathbb{P}(\boldsymbol{X}_{[i+1:n]}) = \mathbb{P}\left(\boldsymbol{X}_{[i+1:n]} \middle| \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha\right)$$
$$\mathbb{P}(\widetilde{\boldsymbol{X}}_{[i+1:n]}) = \mathbb{P}\left(\boldsymbol{X}_{[i+1:n]} \middle| \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta\right)$$

Proof Sketch

- $T(i)$: subtree rooted at vertex i

Proof Sketch

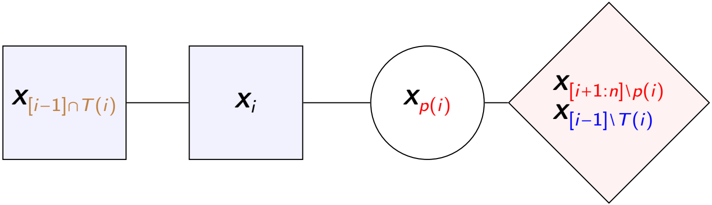
- $T(i)$: subtree rooted at vertex i

$$\mathbb{P}\left(\boldsymbol{X}_{[i+1:n]} \middle| \boldsymbol{X}_{[i-1]}, \boldsymbol{X}_i\right) = \mathbb{P}\left(\boldsymbol{X}_{p(i)}, \boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]\cap T(i)}, \boldsymbol{X}_{[i-1]\setminus T(i)}, \boldsymbol{X}_i\right)$$

Proof Sketch

- $T(i)$: subtree rooted at vertex i

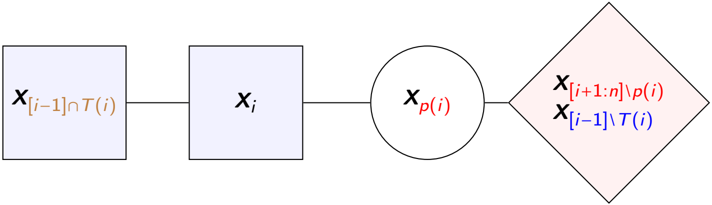
$$\mathbb{P}\left(\boldsymbol{X}_{[i+1:n]} \middle| \boldsymbol{X}_{[i-1]}, \boldsymbol{X}_i\right) = \mathbb{P}\left(\boldsymbol{X}_{p(i)}, \boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]\cap T(i)}, \boldsymbol{X}_{[i-1]\setminus T(i)}, \boldsymbol{X}_i\right)$$



Proof Sketch

- $T(i)$: subtree rooted at vertex i

$$\mathbb{P}\left(\boldsymbol{X}_{[i+1:n]}\middle|\boldsymbol{X}_{[i-1]},\boldsymbol{X}_i\right)=\mathbb{P}\left(\boldsymbol{X}_{p(i)},\boldsymbol{X}_{[i+1:n]\setminus p(i)}\middle|\boldsymbol{X}_{[i-1]\cap T(i)},\boldsymbol{X}_{[i-1]\setminus T(i)},\boldsymbol{X}_i\right)$$

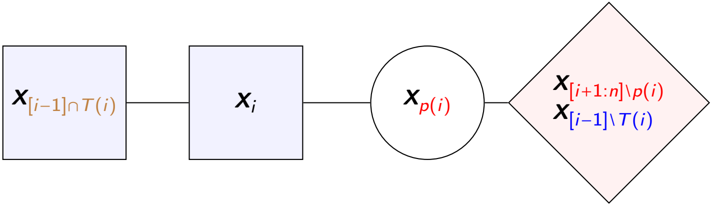


$$\begin{aligned}\mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)}\middle|\boldsymbol{X}_{[i-1]},\boldsymbol{X}_i\right) &= \mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)}\middle|\boldsymbol{X}_{[i-1]\setminus T(i)},\boldsymbol{X}_{[i-1]\cap T(i)},\boldsymbol{X}_i\right)\\ &= \mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)}\middle|\boldsymbol{X}_{[i-1]\setminus T(i)}\right)\end{aligned}$$

Proof Sketch

- $T(i)$: subtree rooted at vertex i

$$\mathbb{P}\left(\boldsymbol{X}_{[i+1:n]} \middle| \boldsymbol{X}_{[i-1]}, \boldsymbol{X}_i\right) = \mathbb{P}\left(\boldsymbol{X}_{p(i)}, \boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]\cap T(i)}, \boldsymbol{X}_{[i-1]\setminus T(i)}, \boldsymbol{X}_i\right)$$



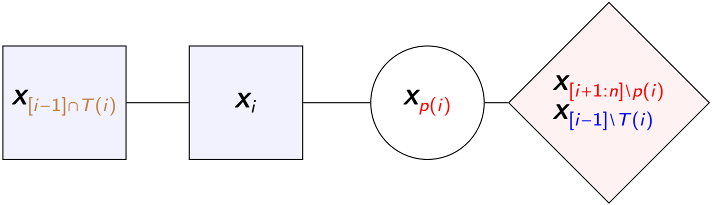
$$\begin{aligned}\mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]}, \boldsymbol{X}_i\right) &= \mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]\setminus T(i)}, \boldsymbol{X}_{[i-1]\cap T(i)}, \boldsymbol{X}_i\right) \\ &= \mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]\setminus T(i)}\right)\end{aligned}$$

- If we set the conditional distribution of $\boldsymbol{X}_{[i+1:n]\setminus p(i)}$ to be the same,

Proof Sketch

- $T(i)$: subtree rooted at vertex i

$$\mathbb{P}\left(\boldsymbol{X}_{[i+1:n]} \middle| \boldsymbol{X}_{[i-1]}, \boldsymbol{X}_i\right) = \mathbb{P}\left(\boldsymbol{X}_{p(i)}, \boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]\cap T(i)}, \boldsymbol{X}_{[i-1]\setminus T(i)}, \boldsymbol{X}_i\right)$$



$$\begin{aligned}\mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]}, \boldsymbol{X}_i\right) &= \mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]\setminus T(i)}, \boldsymbol{X}_{[i-1]\cap T(i)}, \boldsymbol{X}_i\right) \\ &= \mathbb{P}\left(\boldsymbol{X}_{[i+1:n]\setminus p(i)} \middle| \boldsymbol{X}_{[i-1]\setminus T(i)}\right)\end{aligned}$$

- If we set the conditional distribution of $\boldsymbol{X}_{[i+1:n]\setminus p(i)}$ to be the same,
- then the change of $\boldsymbol{X}_i$ only influences $\left\{\boldsymbol{X}_i, \boldsymbol{X}_{p(i)}\right\}$, which is bounded by $c_i + c_{p(i)}$.

Proof Sketch

$\boldsymbol{c}$-Lipschitz + tree-dependence

$$\Rightarrow \quad \sup_{\alpha \in \Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha\right] - \inf_{\beta \in \Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta\right] \leqslant c_i + c_{p(i)}$$

Proof Sketch

$$
\begin{aligned}
& \boldsymbol{c}\text{-Lipschitz} + \text{tree-dependence} \\
\Rightarrow \quad & \sup_{\alpha \in \Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \alpha \right] - \inf_{\beta \in \Omega_i} \mathbb{E}\left[g \,\middle|\, \boldsymbol{X}_{[i-1]} = \mathrm{x}_{[i-1]}, \boldsymbol{X}_i = \beta \right] \leqslant c_i + c_{p(i)} \\
\Rightarrow \quad & \mathbb{E}\left[\exp\left(s(g - \mathbb{E}[g])\right)\right] \leqslant \exp\left(\frac{s^2}{8} \left(c_n^2 + \sum_{i \in V(G) \setminus n} \left(c_i + c_{p(i)}\right)^2 \right) \right) \\
\Rightarrow \quad & \mathbb{P}\left(g - \mathbb{E}[g] \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{c_{\min}^2 + \sum_{\{i,j\} \in E(T)} (c_i + c_j)^2} \right)
\end{aligned}
$$

Janson, 2004

If we have that

1. $\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent, and every X_i is in an interval of length $c_i \geqslant 0$,
2. and the function is a summation,

then, for every $t > 0$,

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

where $\chi_f(G)$ is the fractional chromatic number of G.

Z., 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.

Janson, 2004

If we have that

1. $\boldsymbol{X} = (X_1, \ldots, X_n)$ is G-dependent, and every X_i is in an interval of length $c_i \geqslant 0$,
2. and the function is a summation,

then, for every $t > 0$,

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

where $\chi_f(G)$ is the fractional chromatic number of G.

Z., 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.

Janson, 2004

If we have that

1. $\boldsymbol{X} = (X_1, \dots, X_n)$ is G-dependent, and every X_i is in an interval of length $c_i \geqslant 0$,
2. and the function is a summation,

then, for every $t > 0$,

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

where $\chi_f(G)$ is the fractional chromatic number of G.

Z., 2022

Under the same setting,

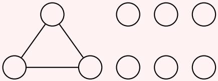
$$\mathbb{P}\left(\sum_{i \in V(G)} X_i - \mathbb{E}\left[\sum_{i \in V(G)} X_i\right] \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{D}\right),$$

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 forest-decomposable functions, extending McDiarmid's.

Example

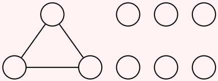
Let $\{X_i\}_{i\in[9]}$ be random indicators with the dependency graph G, and $X = \sum_{i\in[9]} X_i$.



$$\mathbb{P}(X - \mathbb{E}[X] \geqslant t) \leqslant ?$$

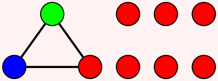
Example

Let $\{X_i\}_{i\in[9]}$ be random indicators with the dependency graph G, and $X = \sum_{i\in[9]} X_i$.



$$\mathbb{P}(X - \mathbb{E}[X] \geqslant t) \leqslant ?$$

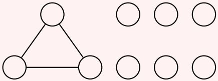
- Janson's idea: to fractionally cover vertices with weighted independent sets such that the sum of weights for each vertex equals 1



$$\mathbb{P}(X - \mathbb{E}[X] \geqslant t) \leqslant \exp\left(-\frac{2t^2}{3\times 9}\right)$$

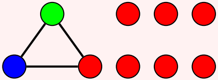
Example

Let $\{X_i\}_{i\in[9]}$ be random indicators with the dependency graph G, and $X = \sum_{i\in[9]} X_i$.



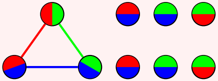
$$\mathbb{P}(X - \mathbb{E}[X] \geqslant t) \leqslant ?$$

- Janson's idea: to fractionally cover vertices with weighted independent sets such that the sum of weights for each vertex equals 1



$$\mathbb{P}(X - \mathbb{E}[X] \geqslant t) \leqslant \exp\left(-\frac{2t^2}{3\times 9}\right)$$

- New idea: to fractionally cover vertices with weighted induced forests such that the sum of weights for each vertex equals 1

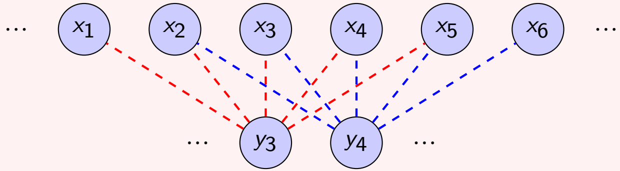


$$\mathbb{P}(X - \mathbb{E}[X] \geqslant t) \leqslant \exp\left(-\frac{8t^2}{81}\right)$$

Applications in statistics

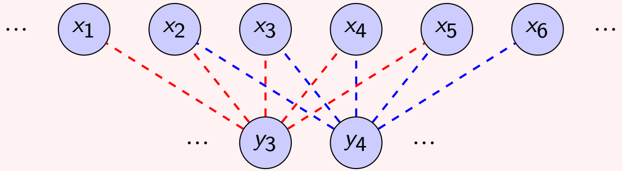
Example

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



Example

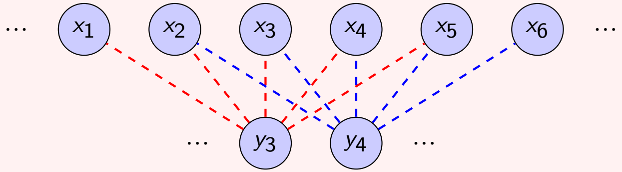
- y_i: observation at location i, e.g., house price
- x_i: random variable modelling influential factors at location i



- Given training data: $\boldsymbol{S} = \{\ldots,((x_1,x_2,x_3,x_4,x_5),y_3),((x_2,x_3,x_4,x_5,x_6),y_4),\ldots\}$.
- To 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

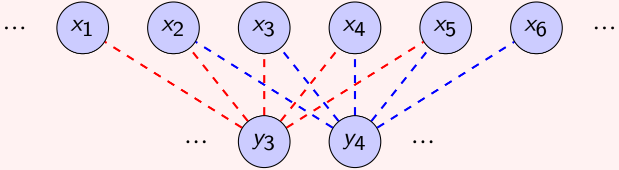


- Given training data: $\boldsymbol{S} = \{\ldots, ((x_1, x_2, x_3, x_4, x_5), y_3), ((x_2, x_3, x_4, x_5, x_6), y_4), \ldots\}$.
- To find $f : (x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}) \mapsto y_i$.

Question: is this realistic?

Example

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

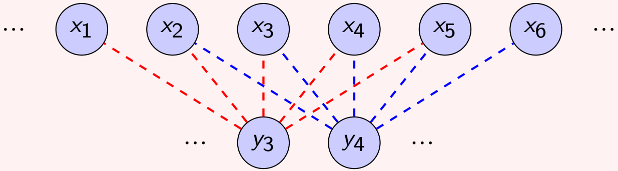


- Given training data: $\boldsymbol{S} = \{\ldots, ((x_1, x_2, x_3, x_4, x_5), y_3), ((x_2, x_3, x_4, x_5, x_6), y_4), \ldots\}$.
- To find $f : (x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}) \mapsto y_i$.

Question: is this realistic? Ask property agents

Example

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



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

Question: is this realistic? Ask property agents or statisticians!

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

Stability bound for learning m-dependent data

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

Uniform stability (Bousquet and Elisseeff 2002)

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

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

where $\boldsymbol{S}^{\setminus i}$ is by deleting i-th data point from $\boldsymbol{S}$.

Stability bound for learning m-dependent data

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

Uniform stability (Bousquet and Elisseeff 2002)

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

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

where $\boldsymbol{S}^{\setminus i}$ is by deleting i-th data point from $\boldsymbol{S}$.

Z. Liu, Wang, Wang 2019

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

which introduces some multiplicative factor of order m,
comparing with the independent case (Bousquet and Elisseeff 2002):

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

Stability bound for learning m-dependent data

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

Uniform stability (Bousquet and Elisseeff 2002)

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

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

where $\boldsymbol{S}^{\setminus i}$ is by deleting i-th data point from $\boldsymbol{S}$.

Z. Liu, Wang, Wang 2019

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

which introduces some multiplicative factor of order m,
comparing with the independent case (Bousquet and Elisseeff 2002):

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

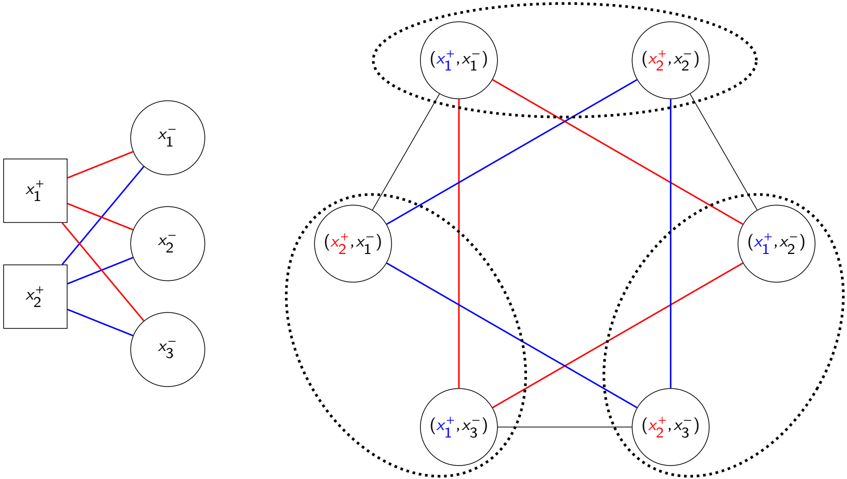
Z. 2022 contains slightly improved concentration results for m-dependent case.

Bipartite ranking

- Training set: $T = (x_i, y_i)_{i=1}^{m}$ 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.

Bipartite ranking

- Training set: $T = (x_i, y_i)_{i=1}^m$ 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 unordered pairs of examples (x,x').



Bipartite ranking

- Let $\boldsymbol{S} = \{(x, x') \in T \times T : y \neq y'\}$ be the set of unordered pairs of examples from different classes in T.

Bipartite ranking

- Let $\boldsymbol{S} = \{(x, x') \in T \times T : y \neq y'\}$ be the set of unordered pairs of examples from different classes in T.
- The empirical loss of a scoring function h over T can be written as a summation over the pairs of instances of different classes:

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

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

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

Bipartite ranking

An approach based on fractional Rademacher complexity gives the following.

Corollary (Z. and Amini 2023+)

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

- **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),
- **When Janson meets McDiarmid: Bounded difference inequalities under graph-dependence**
  Statistics & Probability Letters, 2022,
- **Generalization bounds for learning under graph-dependence: A survey**
  (with Massih-Reza Amini, arXiv:2203.13534).

The content is based upon

- **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),
- **When Janson meets McDiarmid: Bounded difference inequalities under graph-dependence**
  Statistics & Probability Letters, 2022,
- **Generalization bounds for learning under graph-dependence: A survey** (with Massih-Reza Amini, arXiv:2203.13534).

Thanks for your attention!

The content is based upon

- **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),
- **When Janson meets McDiarmid: Bounded difference inequalities under graph-dependence**
  Statistics & Probability Letters, 2022,
- **Generalization bounds for learning under graph-dependence: A survey** (with Massih-Reza Amini, arXiv:2203.13534).

Thanks for your attention!
Question: how to compress it into a 5-min talk at ICSDS?