Bounded Differences Inequalities for Graph-dependent Random Variables and Stability Bounds

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

Joint work with Xingwu Liu, Yuyi Wang, Liwei Wang

MONASH University

Bounded Differences Inequality

$$\mathbf{P}(f(\mathbf{X}) - \mathbf{E}[f(\mathbf{X})] \geq t) \leq ?$$

Definition (c-Lipschitz, Bounded Differences Condition)

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

$$\left| f(x_1, \ldots, \mathbf{x_i}, \ldots, x_n) - f(x_1, \ldots, \mathbf{x_i'}, \ldots, x_n) \right| \leq \mathbf{c_i}$$

Theorem (Bounded Differences Inequality [McDiarmid, 1989])

1. f *is* $\mathbf{c}$*-Lipschitz*
2. $\mathbf{X} = (X_1, \ldots, X_n)$ *are independent random variables*

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

also called Azuma-Hoeffding inequality.

Dependent Random Variables

- Mixing coefficients: $\alpha/\beta/\phi/\Phi/\eta$-mixing, etc.
  - quantitively measure the dependence among random variables.
  - widely used in probability theory, statistical theory.
- Dependency graphs: Lovász Local Lemma, Normal/Poisson approximation, Janson's/Suen's Inequality, etc.
  - combinatorial, independent set, max degree, cumulant, spanning tree, etc.
- Copula, graphical models (random field, Bayesian network, etc.), statistical physics, time series, etc.

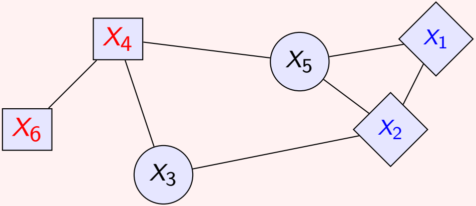
Dependency Graphs

Definition (Dependency Graphs)

G is called a dependency graph for random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ if

- *Vertices $V(G) = [n] = \{1, \ldots, n\}$ represent random variables $X_1, \ldots, X_n$*
- *If disjoint $I, J \subset [n]$ are non-adjacent in G, $\{X_i\}_{i \in I}$ and $\{X_j\}_{j \in J}$ are independent.*

Example



$\{X_1, X_2\}$ and $\{X_4, X_6\}$ are independent.

- The dependency graph for a set of random variables is not necessarily unique.
- There are weaker versions of dependency graphs, e.g. the one used in LLL.

Janson's Hoeffding-type inequality

Theorem ([Hoeffding, 1963])

X: *independent random variables*

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geq t\right) \leq \exp\left(-\frac{2t^2}{\|\mathbf{c}\|_2^2}\right)$$

Theorem ([Janson, 2004])

X: *graph-dependent random variables*

$$\mathbf{P}\left(\sum_{i\in V(G)} X_i - \mathbf{E}\left[\sum_{i\in V(G)} X_i\right] \geq t\right) \leq \exp\left(-\frac{2t^2}{\chi^*(G)\|\mathbf{c}\|_2^2}\right)$$

- $\chi^*(G)$: fractional chromatic number of dependency graph G for random variables **X**.
- idea: decomposition of summation to summations over independent set.
- Janson has another well-known inequality for dependency graphs.

Tree-Dependent Random Variables

Theorem ([Zhang et al., 2019])

1. *f is $\mathbf{c}$-Lipschitz*
2. *T is a dependency graph for $\mathbf{X}$; T is a tree*

$$\mathbf{P}(f(\mathbf{X}) - \mathbf{E}[f(\mathbf{X})] \geq t) \leq \exp\left(-\frac{2t^2}{c_{\min}^2 + \sum_{\{i,j\}\in E(T)}(c_i + c_j)^2}\right)$$

where $c_{\min}$ is the minimum entry of $\mathbf{c}$.

MONASH University

Forest-Dependent Random Variables

Theorem ([Zhang et al., 2019])

1. *f is $\mathbf{c}$-Lipschitz*
2. *F is a dependency graph for $\mathbf{X}$; $F = \{T_i\}_{i \in [k]}$ is a forest*

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

where $c_{\min,i} = \min\{c_j : j \in V(T_i)\}$

- strict generalisation of the McDiarmid's inequality for independent random variables
- By transforming graph to forest via merging vertices, using the notion of Forest Complexity $\Lambda(G)$, we can handle general graph G

$$\exp\left(-\frac{2t^2}{\Lambda(G)\mathbf{c}_{\max}^2}\right)$$

Examples

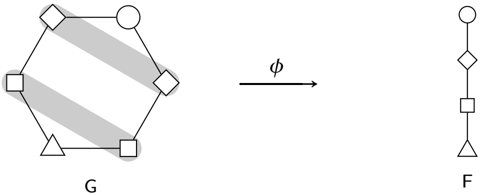


Figure: C_6: $\Lambda(G) \leq 8n - 13 = O(n)$

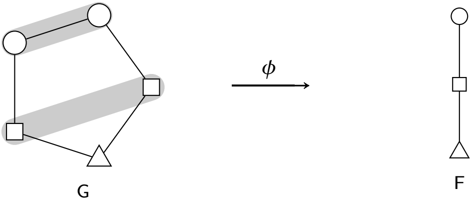


Figure: C_5: $\Lambda(G) \leq 8n - 14 = O(n)$

MONASH University

Examples

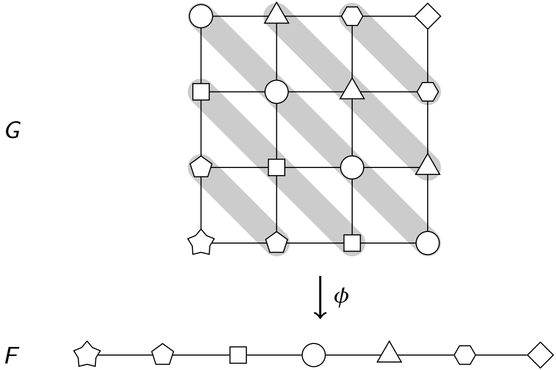


Figure: 4×4-grid $\Lambda(G) = O(n^{\frac{3}{2}})$

m-dependence

Corollary (m-dependence [Zhang et al., 2019])

Random variables $\{X_i\}_{i=1}^n$ *is called m-dependent if for any* $i \in [n-m-1]$, $\{X_j\}_{j=1}^{i}$ *is independent of* $\{X_j\}_{j=i+m+1}^{n}$.

$$\Lambda(G) \le \left(\left\lceil \frac{n}{m} \right\rceil - 1\right)(m+m)^2 + m^2 \le 4mn = O(mn)$$

$$\mathbf{P}\left(f(\mathbf{X}) - \mathbf{E}\left[f(\mathbf{X})\right] \ge t\right) \le \exp\left(-\frac{2t^2}{4mn\mathbf{c}_{\max}^2}\right)$$

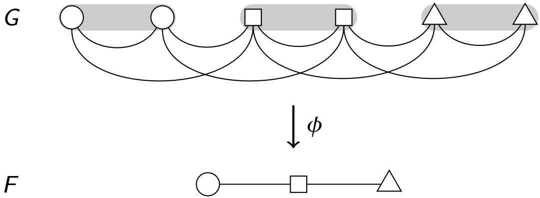


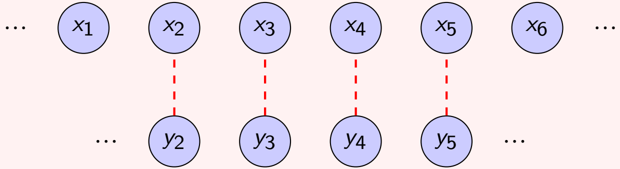
Figure: 2-dependent sequence

MONASH University

Applications in Machine Learning

Example

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



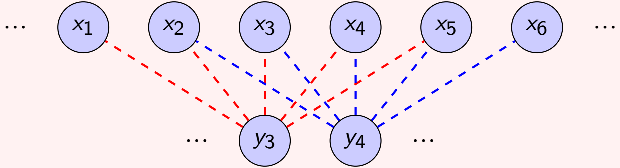
- Given training data: $\mathbf{S} = \{\ldots, (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5), \ldots\}$
- Find predictive function $f : x_i \mapsto y_i$

MONASH University

Applications in Machine Learning

Example

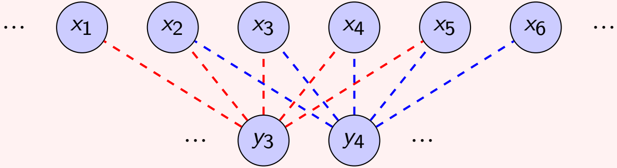
- y_i: the observation at location i, e.g., the house price
- x_i: the random variable modelling influential factors at location i



Applications in Machine Learning

Example

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



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

Background on Machine Learning

- Given input x, choose $f : x \mapsto y$ that perform well on unknown new data.
- A training data set $\mathbf{S}$ contains n samples $(x_i, y_i) \sim D$ (unknown)
- Loss function measures error between true y and predicted $f(x)$

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

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

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

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

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

- Goal is to establish generalisation error bounds

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

MONASH University

Stability Bound for Learning Graph-Dependent Data

A learning algorithm $\mathscr{A} : \mathbf{S} \mapsto f_{\mathbf{S}}^{\mathscr{A}}$ outputs $f_{\mathbf{S}}^{\mathscr{A}}$ given samples $\mathbf{S}$

Definition (Uniform Stability [Bousquet and Elisseeff, 2002])

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

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

Lemma

- $R(f_{\mathbf{S}}^{\mathscr{A}}) - \widehat{R}(f_{\mathbf{S}}^{\mathscr{A}})$ *is* $(4\beta_n + M/n)$*-Lipschitz*
- $\mathbf{E}\left[R(f_{\mathbf{S}}^{\mathscr{A}}) - \widehat{R}(f_{\mathbf{S}}^{\mathscr{A}})\right] \leq 2\beta_{n,\Delta}(\Delta + 1)$, Δ*: max degree*

Theorem ([Zhang et al., 2019])

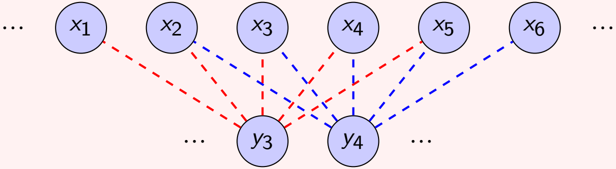
Let $\beta_{n,\Delta} = \max_{i \in [0,\Delta]} \beta_{n-i}$. *For* $\delta \in (0,1)$, *with probability at least* $1 - \delta$,

$$R(f_{\mathbf{S}}^{\mathscr{A}}) \leq \widehat{R}(f_{\mathbf{S}}^{\mathscr{A}}) + 2\beta_{n,\Delta}(\Delta + 1) + (4\beta_n + M/n)\sqrt{\frac{\Lambda(G)\ln(1/\delta)}{2}}$$

Stability Bound for Learning m-dependent Data

Example

- y_i: the observation at location i, e.g., the house price
- x_i: the random variable modelling geographical effect at location i



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

Corollary ([Zhang et al., 2019])

$$R(f_{\mathbf{S}}^{\mathscr{A}}) \leq \widehat{R}(f_{\mathbf{S}}^{\mathscr{A}}) + 2\beta_{n,2m}(2m+1) + (4n\beta_n + M)\sqrt{\frac{2m\ln(1/\delta)}{n}}$$

Future Work

- Improve the results to match the summation bound by Janson.
- Results using other dependency graph models
  - Weighted dependency graphs [Féray et al., 2018]
  - Threshold dependency graphs [Lampert et al., 2018]
- Results for weaker versions of dependency graphs
  - Weak dependency graph: $\mathbf{X}_i$ is independent of $\mathbf{X}_{[n]\setminus N^+(i)}$
  - Pairwise independence: $\mathbf{X}_i$ is independent of $\mathbf{X}_u : u \notin N^+(i)$
- Results for dependency hypergraphs
  - Dependent random variables are generated by independent ones by sharing variables (similar to the variable version Lovász Local Lemma)

MONASH University

- *McDiarmid-type Inequalities for Graph-dependent Variables and Stability Bounds*
  Spotlight in Advances in Neural Information Processing Systems 32 (NeurIPS 2019)
- Thanks for your time and attention!

MONASH University

References I

Bousquet, O. and Elisseeff, A. (2002).
Stability and generalization.
Journal of machine learning research, 2(Mar):499–526.

Féray, V. et al. (2018).
Weighted dependency graphs.
Electronic Journal of Probability, 23.

Hoeffding, W. (1963).
Probability inequalities for sums of bounded random variables.
Journal of the American statistical association, 58(301):13–30.

Janson, S. (2004).
Large deviations for sums of partly dependent random variables.
Random Structures & Algorithms, 24(3):234–248.

Lampert, C. H., Ralaivola, L., and Zimin, A. (2018).
Dependency-dependent bounds for sums of dependent random variables.
arXiv preprint arXiv:1811.01404.

McDiarmid, C. (1989).
On the method of bounded differences.
Surveys in combinatorics, 141(1):148–188.

Zhang, R.-R., Liu, X., Wang, Y., and Wang, L. (2019).
Mcdiarmid-type inequalities for graph-dependent variables and stability bounds.
In *Advances in Neural Information Processing Systems 32*, pages 10889–10899.

MONASH University