McDiarmid-Type Inequalities for Graph-Dependent Variables and Stability Bounds

Rui (Ray) Zhang[1]
rui.zhang@monash.edu
School of Mathematics, Monash University

Joint work with Xingwu Liu, Yuyi Wang, Liwei Wang

September 21, 2019

[1] This work was done when this author was a master student at the Institute of Computing Technology, Chinese Academy of Sciences and University of Chinese Academy of Sciences.

MONASH University

Contents

MONASH University

Contents

Concentration of Measure

- Concentration iequalities

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

- common assumption: random variables are independent
- what if r.v. are not independent
  - mixing coefficients: α-mixing [Rosenblatt, 1956], β-mixing [Volkonskii and Rozanov, 1959], ϕ-mixing [Ibragimov, 1962], η-mixing [Kontorovich, 2007], etc.
  - dependency graph: Local Lemma [Erdos and Lovász, 1975], Normal/Poisson Approximation [Chen, 1978, Janson et al., 1988, Baldi et al., 1989]
- Goal: McDiarmid-type concentration inequality for graph-dependent random variables

MONASH University

Janson's Hoeffding-type inequality

Definition (Dependency Graphs)

An undirected graph G is called a dependency graph of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ if

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

Theorem (Janson's inequality [Janson, 2004])

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

- $\chi^*(G)$: fractional coloring number of a dependency graph G of random variables $\mathbf{X}$
- idea: decomposition of summation to summation over independent set

McDiarmid's inequality

Definition (c-Lipschitz, bounded differences condition)

Given a vector $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}_+^n$*, a function* $f : \boldsymbol{\Omega} \to \mathbb{R}$ *is said to be* $\mathbf{c}$*-Lipschitz if for any* $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{x}' = (x_1', \ldots, x_n') \in \boldsymbol{\Omega}$*, it satisfies*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq \sum_{i=1}^{n} c_i \mathbf{1}_{\{x_i \neq x_i'\}}$$

where c_i *is called the* i*-th Lipschitz coefficient of* f*.*

originated from Hoeffding-Azuma [Hoeffding et al., 1948, Azuma, 1967]

Theorem (McDiarmid's inequality [McDiarmid, 1989])

Suppose $f : \boldsymbol{\Omega} \to \mathbb{R}$ *is* $\mathbf{c}$*-Lipschitz, and* $\mathbf{X} = (X_1, \ldots, X_n)$ *is a vector of independent r.v. with each* X_i *taking values in* Ω_i*. Then for any* $t > 0$*,*

$$\mathbf{Pr}(f(\mathbf{X}) - \mathbf{E}[f(\mathbf{X})] \geq t) \leq \exp\left(-\frac{2t^2}{\|\mathbf{c}\|_2^2}\right) \tag{1}$$

Contents

MONASH University

McDiarmid-type inequality for dependency tree

Theorem

Suppose that $f : \boldsymbol{\Omega} \to \mathbb{R}$ is a $\mathbf{c}$-Lipschitz function and T is a dependency tree of a random vector $\mathbf{X}$ that takes values in $\boldsymbol{\Omega}$. Then for any $t > 0$,

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

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

- idea: vertex exposure ordering + Doob martingale + conditional probability coupling

McDiarmid-type inequality for dependency forest

Theorem

Suppose that $f : \mathbf{\Omega} \to \mathbb{R}$ *is a* $\mathbf{c}$*-Lipschitz function and* G *is a dependency graph of a random vector* $\mathbf{X}$ *that takes values in* $\mathbf{\Omega}$*. If* G *is a forest consisting of trees* $\{T_i\}_{i \in [k]}$*, then for any* $t > 0$,

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

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

- strict generalization of the McDiarmid's inequality for i.i.d. random variables

McDiarmid-type inequality for general dependency graph

idea: transform graph to forest via merging vertices

$$\lambda_{(\phi,F)} = \sum_{\langle u,v\rangle \in E(F)} \left(|\phi^{-1}(u)| + |\phi^{-1}(v)|\right)^2 + \sum_{i=1}^{k} \min_{u \in V(T_i)} |\phi^{-1}(u)|^2$$

We call

$$\Lambda(G) = \min_{(\phi,F) \in \Phi(G)} \lambda_{(\phi,F)}$$

the Forest Complexity of the graph G

Theorem

$$\mathbf{Pr}(f(\mathbf{X}) - \mathbf{E}[f(\mathbf{X})] \geq t) \leq \exp\left(-\frac{2t^2}{\Lambda(G)\|\mathbf{c}\|_\infty^2}\right)$$

independent case: $\Lambda(G) = n$, complete graph: $\Lambda(G) = n^2$

Examples

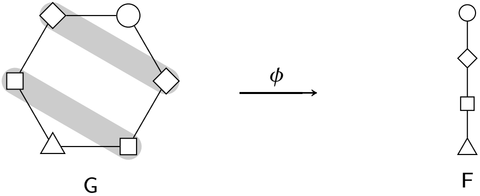


Figure: C_6: $\Lambda(G) \leq 8n - 13 = O(n)$

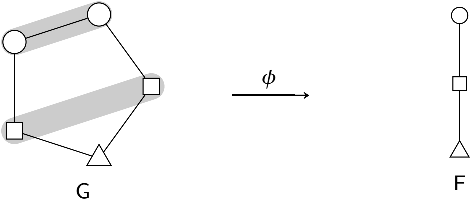


Figure: C_5: $\Lambda(G) \leq 8n - 14 = O(n)$

MONASH University

Examples

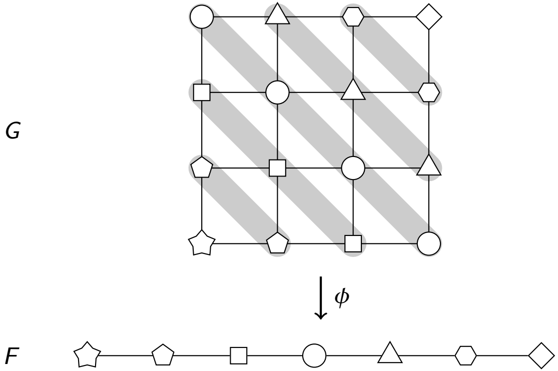


Figure: 4×4-grid $\Lambda(G) = O(m^3) = O(n^{\frac{3}{2}})$

Contents

MONASH University

Uniform stability

Define $f_{\mathbf{S}}^{\mathscr{A}} : \mathscr{X} \to \mathscr{Y}$ to be the the hypothesis that $\mathscr{A}$ has learned from the sample $\mathbf{S}$

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

Given integer $n > 0$, the learning algorithm $\mathscr{A}$ is called β_n-uniformly stable with respect to the loss function ℓ, if for any $i \in [n]$, $\mathbf{S} \in (\mathscr{X} \times \mathscr{Y})^n$, and $(x, y) \in \mathscr{X} \times \mathscr{Y}$, it holds that

$$|\ell(y, f_{\mathbf{S}}^{\mathscr{A}}(x)) - \ell(y, f_{\mathbf{S}^{\setminus i}}^{\mathscr{A}}(x))| \le \beta_n.$$

define $\Phi_{\mathscr{A}}(\mathbf{S}) = R(f_{\mathbf{S}}^{\mathscr{A}}) - \widehat{R}(f_{\mathbf{S}}^{\mathscr{A}})$

Lemma

$$\mathbf{E}[\Phi_{\mathscr{A}}(\mathbf{S})] \le 2\beta_{n,\Delta}(\Delta + 1).$$

Stability Bound for Learning Graph-dependent Data

Theorem

Given a sample $\mathbf{S}$ *of size* n *with dependency graph* G*, assume that the learning algorithm* $\mathscr{A}$ *is* β_i*-uniformly stable for any* $i \leq n$*. Suppose the maximum degree* G *is* Δ*, and the loss function* ℓ *is bounded by* M*. Let* $\beta_{n,\Delta} = \max_{i \in [0,\Delta]} \beta_{n-i}$*. For any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, it holds that*

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

Applications

Example (Spatial Poisson point process)

Consider a Poisson point process on $\mathbb{R}^2$. The number of points in each finite region follows a Poisson distribution, and the number of points in disjoint regions are independent. Given a finite set $\mathscr{I} = \{I_i\}_{i=1}^n$ of regions in $\mathbb{R}^2$, let X_i be the number of points in region I_i, $1 \le i \le n$. Then the graph

$$G\left([n], \left\{\langle i,j\rangle : I_i \cap I_j \neq \emptyset\right\}\right)$$

is a dependency graph of the random variables $\{X_i\}_{i=1}^n$.

Example (m-dependence [Hoeffding et al., 1948])

For some $m, n \in \mathbb{N}_+$, a sequence of 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 $\{X_j\}_{j=i+m+1}^n$.

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

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

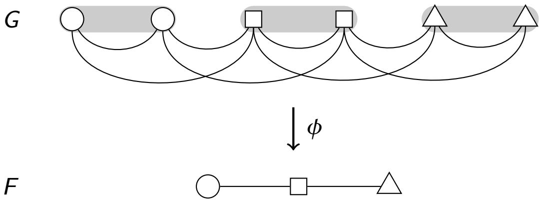
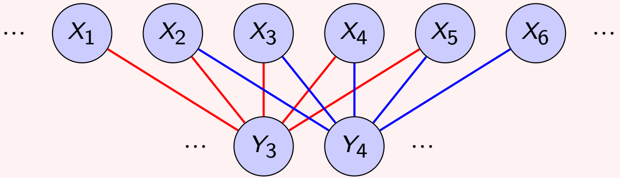


Figure: 2-dependent sequence

Example

Let y_i be the observation at location i, e.g., the house price, and x_i stand for the random variable modeling geographical effect at location i.

Example



$(\mathbf{X}_i, Y_i)$: geographical effect, house price;
$\{((X_1, X_2, X_3, X_4, X_5), Y_3), ((X_2, X_3, X_4, X_5, X_6), Y_4)\}$

MONASH University

References I

Azuma, K. (1967).
Weighted sums of certain dependent random variables.
Tohoku Mathematical Journal, Second Series, 19(3):357–367.

Baldi, P., Rinott, Y., et al. (1989).
On normal approximations of distributions in terms of dependency graphs.
The Annals of Probability, 17(4):1646–1650.

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

Chen, L. H. (1978).
Two central limit problems for dependent random variables.
Probability Theory and Related Fields, 43(3):223–243.

Erdos, P. and Lovász, L. (1975).
Problems and results on 3-chromatic hypergraphs and some related questions.
Infinite and finite sets, 10(2):609–627.

Hoeffding, W., Robbins, H., et al. (1948).
The central limit theorem for dependent random variables.
Duke Mathematical Journal, 15(3):773–780.

Ibragimov, I. A. (1962).
Some limit theorems for stationary processes.
Theory of Probability & Its Applications, 7(4):349–382.

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

References II

Janson, S., Luczak, T., and Rucinski, A. (1988).
An exponential bound for the probability of nonexistence of a specified subgraph in a random graph.
Institute for Mathematics and its Applications (USA).

Kontorovich, L. (2007).
Measure concentration of strongly mixing processes with applications.
Carnegie Mellon University.

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

Rosenblatt, M. (1956).
A central limit theorem and a strong mixing condition.
Proceedings of the National Academy of Sciences of the United States of America, 42(1):43.

Volkonskii, V. and Rozanov, Y. A. (1959).
Some limit theorems for random functions. i.
Theory of Probability & Its Applications, 4(2):178–197.

MONASH University