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

Rui (Ray) Zhang\textsuperscript{1}
rui.zhang@monash.edu
School of Mathematics, Monash University

Joint work with Xingwu Liu, Yuyi Wang, Liwei Wang

September 21, 2019

\textsuperscript{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.
Contents

1 Introduction

2 Concentration Results for Graph-Dependent Random Variables

3 Stability Bound for Learning Graph-dependent Data
Concentration of Measure

- Concentration inequalities
  \[ \Pr(f(X) - E[f(X)] \geq t) \leq ? \]

- common assumption: random variables are independent

- what if r.v. are not independent
  - mixing coefficients: \( \alpha \)-mixing [Rosenblatt, 1956], \( \beta \)-mixing [Volkonskii and Rozanov, 1959], \( \phi \)-mixing [Ibragimov, 1962], \( \eta \)-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
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])**

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

- $\chi^*(G)$: 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 \( c = (c_1, \ldots, c_n) \in \mathbb{R}^n_+ \), a function \( f : \Omega \to \mathbb{R} \) is said to be \( c \)-Lipschitz if for any \( x = (x_1, \ldots, x_n), x' = (x'_1, \ldots, x'_n) \in \Omega \), it satisfies

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

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

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

(1)
1 Introduction

2 Concentration Results for Graph-Dependent Random Variables

3 Stability Bound for Learning Graph-dependent Data
Theorem

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

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

where \( c_{\text{min}} \) is the minimum entry in \( c \).

- idea: vertex exposure ordering + Doob martingale + conditional probability coupling
McDiarmid-type inequality for dependency forest

**Theorem**

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

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

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
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**

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

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

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

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

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

1 Introduction

2 Concentration Results for Graph-Dependent Random Variables

3 Stability Bound for Learning Graph-dependent Data
Define $f_S^A : \mathcal{X} \to \mathcal{Y}$ to be the hypothesis that $A$ has learned from the sample $S$.

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

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

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

Define $\Phi_A(S) = R(f_S^A) - \hat{R}(f_S^A)$.

**Lemma**

$$E[\Phi_A(S)] \leq 2\beta_{n,\Delta}(\Delta + 1).$$
Theorem

Given a sample \( S \) of size \( n \) with dependency graph \( G \), assume that the learning algorithm \( \mathcal{A} \) is \( \beta_i \)-uniformly stable for any \( i \leq n \). Suppose the maximum degree \( G \) is \( \Delta \), and the loss function \( \ell \) 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

\[
R(f_{\mathcal{A}S}) - \hat{R}(f_{\mathcal{A}S}) \leq 2\beta_{n,\Delta}(\Delta + 1) + \frac{4n\beta_n + M}{n} \sqrt{\frac{\Lambda(G)\ln(1/\delta)}{2}}.
\]
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 $\mathcal{I} = \{I_i\}_{i=1}^n$ of regions in $\mathbb{R}^2$, let $X_i$ be the number of points in region $I_i$, $1 \leq i \leq n$. Then the graph

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

is a dependency graph of the random variables $\{X_i\}_{i=1}^n$. 
\[ \Lambda(G) \leq \left( \left\lceil \frac{n}{m} \right\rceil - 1 \right) (m + m)^2 + m^2 \leq 4mn = O(mn) \]

\[ R(f_S^G) \leq \hat{R}(f_S^G) + 2\beta_{n,2m}(2m+1) + (4n\beta_n + M)\sqrt{\frac{2m\ln(1/\delta)}{n}}. \]

**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 \(\{X_j\}_{j=i+m+1}^n\).

**Figure:** 2-dependent sequence
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 $i$.

$$(X_i, Y_i): \text{geographical effect, house price;}$$

$$\{(X_1, X_2, X_3, X_4, X_5), Y_3\}, \{(X_2, X_3, X_4, X_5, X_6), Y_4\}$$
Weighted sums of certain dependent random variables.

On normal approximations of distributions in terms of dependency graphs.

Stability and generalization.

Two central limit problems for dependent random variables.

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.

Some limit theorems for stationary processes.

Large deviations for sums of partly dependent random variables.
An exponential bound for the probability of nonexistence of a specified subgraph in a random graph.
Institute for Mathematics and its Applications (USA).

Measure concentration of strongly mixing processes with applications.
Carnegie Mellon University.

On the method of bounded differences.

A central limit theorem and a strong mixing condition.

Some limit theorems for random functions. i.