Summary Part 3: Applications of Concentration Inequalities

Tianpei Xie

Jan. 26th., 2023

Contents

1 Applications

1.1 U-Statistics

- **Example** (***U-Statistics***) [Wainwright, 2019]
  Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a ***symmetric function*** of its arguments. Given an i.i.d. sequence $\{X_k, k \geq 1\}$, of random variables, the quantity

$$U := \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k) \tag{1}$$

  is known as a ***pairwise U-statistic***. For instance, if $g(s,t) = |s-t|$, then U is an *unbiased estimator* of *the mean absolute pairwise deviation* $\mathbb{E}[|X_1 - X_2|]$. Note that, while U is ***not*** *a sum of independent random variables*, the *dependence is relatively weak*, and this fact can be revealed by a martingale analysis.

  If g is bounded (say $\|g\|_\infty \leq b$), then *the Bounded Difference Inequality* can be used to establish the concentration of U around its mean. Viewing U as a function $f(x) = f(x_1, \ldots, x_n)$, for any given coordinate k, we have

$$\begin{aligned} \left| f(x) - f(x^{(-k)}) \right| &\leq \frac{1}{\binom{n}{2}} \sum_{j \neq k} \left| g(x_j, x_k) - g(x_j, x'_k) \right| \\ &\leq \frac{(n-1)2b}{\binom{n}{2}} = \frac{4b}{n}, \end{aligned}$$

  so that the bounded differences property holds with parameter $L_k = \frac{4b}{n}$ in each coordinate. Thus, we conclude that

$$\mathbb{P}\{|U - \mathbb{E}[U]| \geq t\} \leq 2 \exp\left(-\frac{n\, t^2}{8b^2} \right),$$

  This tail inequality implies that U is a consistent estimate of $\mathbb{E}[U]$, and also yields *finite sample bounds* on its quality as an estimator. Similar techniques can be used to obtain *tail bounds on U-statistics of higher order*, involving sums over k-tuples of variables. ∎

1.2 Jackknife Estimation and Boostrapping

1.3 Kernel-Density Estimation

- **Example** (***Kernel Density Estimation***)
  Let $Z_1, \ldots, Z_n$ be i.i.d. samples drawn according to some (unknown) density ϕ on the real line. The density is estimated by the kernel estimate

$$\phi_n(z) = \frac{1}{n\, h_n} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{h_n} \right),$$

  where $h_n > 0$ is a *smoothing parameter*, and K is a nonnegative function with $\int K(z) = 1$. The performance of the estimate is typically measured by ***the*** L_1 ***error***:

$$X(n) := f(Z_1, \ldots, Z_n) = \int |\phi(z) - \phi_n(z)|\, dz.$$

It is easy to see that

$$\begin{aligned}\left|f(z_1,\ldots,z_n)-f_i(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n)\right| &\le \frac{1}{nh_n}\int\left|K\left(\frac{z-z_i}{h_n}\right)-K\left(\frac{z-z_i'}{h_n}\right)\right|dz\\ &\le \frac{2}{n},\end{aligned}$$

so without further work we obtain

$$\mathrm{Var}\left(X(n)\right)\le\frac{1}{n}$$

It is known that for every ϕ, $\sqrt{n}\mathbb{E}\left[X(n)\right]\to\infty$, which implies, by *Chebyshev's inequality*, that for every $\epsilon>0$

$$\mathbb{P}\left\{\left|\frac{X(n)}{\mathbb{E}\left[X(n)\right]}-1\right|>\epsilon\right\}=\mathbb{P}\left\{\left|X(n)-\mathbb{E}\left[X(n)\right]\right|>\epsilon\mathbb{E}\left[X(n)\right]\right\}\le\frac{\mathrm{Var}(X(n))}{\epsilon^2(\mathbb{E}\left[X(n)\right])^2}\to 0$$

as $n\to\infty$. That is, $\frac{X(n)}{\mathbb{E}[X(n)]}\to 1$ ***in probability***, or in other words, $X(n)$ is ***relatively stable***. This means that ***the random L_1-error*** essentially *behaves like* ***its expected value***.

By bounded difference inequality, we have

$$\mathbb{P}\left\{\left|X(n)-\mathbb{E}\left[X(n)\right]\right|\ge t\right\}\le 2\exp\left(-\frac{nt^2}{2}\right)\quad\blacksquare$$

1.4 Random Graph

- **Example** (***Clique Number in Erdös-Rényi Random Graphs***) [Wainwright, 2019]
  Let $\mathcal{G}=(\mathcal{V},\mathcal{E})$ be an *undirected graph*, where $\mathcal{V}=\{1,\ldots,d\}$ is the vertex set and $\mathcal{E}=\{(i,j),i,j\in\mathcal{V}\}$ is the undirected edge set. *A* ***graph clique*** C is a subset of vertices such that $(i,j)\in\mathcal{E}$ *for all* $i,j\in C$. ***The clique number*** $C(\mathcal{G})$ of the graph is ***the cardinality of the largest clique***. Note that $C(\mathcal{G})\in[1,d]$. When the edges $\mathcal{E}$ of the graph are drawn according to some random process, then *the clique number* $C(\mathcal{G})$ is a *random variable*, and we can study its concentration around its mean $\mathbb{E}\left[C(\mathcal{G})\right]$.

  The Erdös-Rényi ensemble *of* ***random graphs*** is one of the most well-studied models: it is defined by a parameter $p\in(0,1)$ that specifies the probability with which each edge (i,j) is *included* in the graph, ***independently*** *across all* $\binom{d}{2}$ *edges*. More formally, for each $i<j$, let us introduce a ***Bernoulli edge-indicator*** *variable* $X_{i,j}$ with parameter p, where $X_{i,j}=1$ means that edge (i,j) is *included* in the graph, and $X_{i,j}=0$ means that it is *not included*.

  Note that the $\binom{d}{2}$-dimensional random vector $Z:=\{X_{i,j}\}_{i<j}$ specifies the edge set; thus, we may view *the clique number* $C(\mathcal{G})$ as a function $Z\to f(Z)$. Based on definition in Section **??**, we see that $f(Z)$ is a ***configuration function*** with property of "*being in a clique*".

  Let Z' denote a vector in which a *single coordinate* of Z has been changed, and let $\mathcal{G}'$ and $\mathcal{G}$ be the associated graphs. It is easy to see that $C(\mathcal{G}')$ can differ from $C(\mathcal{G})$ by at most 1, so that

$$\left|f(Z)-f(Z')\right|\le 1,$$

Thus, the function $C(\mathcal{G}) = f(Z)$ satisfies *the bounded difference property* in each coordinate with parameter $L = 1$, so that

$$\mathbb{P}\left\{\frac{1}{n}\left|C(\mathcal{G}) - \mathbb{E}\left[C(\mathcal{G})\right]\right| \geq \delta\right\} \leq 2\exp\left(-2n\delta^2\right).$$

Consequently, we see that the clique number of an *Erdös-Rényi random graph* is *very sharply concentrated around its expectation.* ∎

1.5 Minimum Weight Spanning Tree

1.6 Rademacher Complexity

- **Example** (***Rademacher Complexity***) [Wainwright, 2019]
Let $\{\epsilon_k\}_{k=1}^n$ be an i.i.d. sequence of *Rademacher variables* (i.e., taking the values $\{-1, +1\}$ *equiprobably*). Given a collection of vectors $\mathcal{A} \subset \mathbb{R}^n$, define the random variable

$$Z := \sup_{a\in\mathcal{A}} \sum_{k=1}^{n} \epsilon_k a_k = \sup_{a\in\mathcal{A}} \langle a\,,\, \epsilon\rangle\,. \tag{2}$$

The random variable Z measures the ***size*** *of $\mathcal{A}$* in a certain sense, and its expectation

$$\mathfrak{R}(\mathcal{A}) := \mathbb{E}\left[Z(\mathcal{A})\right] \tag{3}$$

is known as ***the Rademacher complexity*** *of the set $\mathcal{A}$.*

Let us now show how the bounded difference inequality can be used to establish that $Z(\mathcal{A})$ is ***sub-Gaussian***. Viewing $Z(\mathcal{A})$ as a function $(\epsilon_1, \ldots, \epsilon_n) \to f(\epsilon_1, \ldots, \epsilon_n)$, we need to *bound the maximum change* when coordinate k is changed. Given two Rademacher vectors $\epsilon, \epsilon' \in \{-1, +1\}^n$, recall our definition of the modified vector $\epsilon^{(-k)}$. Since

$$f(\epsilon^{(-k)}) \geq \left\langle a\,,\, \epsilon^{(-k)}\right\rangle, \quad \text{for any } a \in \mathcal{A},$$

we have

$$\langle a\,,\, \epsilon\rangle - f(\epsilon^{(-k)}) \leq \left\langle a\,,\, \epsilon - \epsilon^{(-k)}\right\rangle = a_k(\epsilon_k - \epsilon'_k) \leq 2\left|a_k\right|.$$

Taking *the supremum over $\mathcal{A}$ on both sides*, we obtain the inequality

$$f(\epsilon) - f(\epsilon^{(-k)}) \leq 2\sup_{a\in\mathcal{A}} \left|a_k\right|.$$

Since the same argument applies with the roles of ϵ and $\epsilon^{(-k)}$ *reversed*, we conclude that f satisfies *the bounded difference inequality* in coordinate k with parameter $L_k := 2\sup_{a\in\mathcal{A}} |a_k|$.

Consequently, the bounded difference inequality implies that the random variable $Z(\mathcal{A})$ is *sub-Gaussian* with parameter at most $2\sqrt{\sum_{k=1}^n \sup_{a\in\mathcal{A}} a_k^2}$. This sub-Gaussian parameter can be reduced to the (potentially much) smaller quantity $\sqrt{\sup_{a\in\mathcal{A}} \sum_{k=1}^n a_k^2}$ using alternative techniques. ∎

1.7 Dimensionality Reduction

2 Self-Bounding Functions

2.1 Definitions, Variance Bounds and Concentration

- Another simple property which is satisfied for many important examples is the so-called *self-bounding property.*

  Definition (***Self-Bounding Property***)
  A ***nonnegative*** *function* $f : \mathcal{X}^n \to [0, \infty)$ has the ***self-bounding property*** if *there exist* functions $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ such that for all $z_1, \ldots, z_n \in \mathcal{X}$ and all $i = 1, \ldots, n$,

$$0 \leq f(z_1, \ldots, z_n) - f_i(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n) \leq 1 \tag{4}$$

  and also

$$\sum_{i=1}^{n} \left(f(z_1, \ldots, z_n) - f_i(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)\right) \leq f(z_1, \ldots, z_n). \tag{5}$$

- **Remark** Clearly if f has the ***self-bounding property***,

$$\sum_{i=1}^{n} \left(f(z_1, \ldots, z_n) - f_i(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)\right)^2 \leq f(z_1, \ldots, z_n) \tag{6}$$

  Taking expectation on both sides, we have the following inequality

- **Corollary 2.1** *[Boucheron et al., 2013]*
  If f has the ***self-bounding property****, then*

$$Var(f(Z)) \leq \mathbb{E}\left[f(Z)\right].$$

- **Remark** (***Relative Stability***) [Boucheron et al., 2013]
  A sequence of nonnegative random variables $(Z_n)_{n\in\mathbb{N}}$ is said to be ***relatively stable*** if

$$\frac{Z_n}{\mathbb{E}\left[Z_n\right]} \xrightarrow{\mathbb{P}} 1.$$

  This property guarantees that ***the random fluctuations*** *of Z_n around its* ***expectation*** *are* ***of negligible size*** *when compared to the expectation*, and therefore ***most information about the size of Z_n is given by*** $\mathbb{E}\left[Z_n\right]$.

  Bounding the variance of Z_n by its expected value *implies, in many cases,* ***the relative stability*** *of $(Z_n)_{n\in\mathbb{N}}$.* If Z_n has the ***self-bounding property***, then, by *Chebyshev's inequality*, for all $\epsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{Z_n}{\mathbb{E}\left[Z_n\right]} - 1\right| > \epsilon\right\} \leq \frac{\mathrm{Var}(Z_n)}{\epsilon^2(\mathbb{E}\left[Z_n\right])^2} \leq \frac{1}{\epsilon^2\mathbb{E}\left[Z_n\right]}.$$

  Thus, for relative stability, it suffices to have $\mathbb{E}\left[Z_n\right] \to \infty$.

2.2 Configuration Function

- An important class of functions satisfying *the self-bounding property* consists of the so-called ***configuration functions***.

  Definition (***Configuration Function***)
  Assume that we have a property Π ***defined over the union of finite products*** of a set $\mathcal{X}$, that is, a sequence of sets

  $$\Pi_1 \subset \mathcal{X},\ \Pi_2 \subset \mathcal{X} \times \mathcal{X},\ \ \ldots,\ \ \Pi_n \subset \mathcal{X}^n.$$

  We say that $(z_1, \ldots, z_m) \in \mathcal{X}^m$ ***satisfies the property*** Π if $(z_1, \ldots, z_m) \in \Pi_m$.

  We assume that Π is ***hereditary*** in the sense that if $(z_1, \ldots, z_m)$ satisfies Π then so does ***any sub-sequence*** $\{z_{i_1}, \ldots, z_{i_k}\}$ of $(z_1, \ldots, z_m)$.

  The function f that maps any vector $z = (z_1, \ldots, z_n)$ to ***the size*** *of a* ***largest sub-sequence*** *satisfying* Π is ***the configuration function*** *associated with property* Π.

- **Corollary 2.2** *[Boucheron et al., 2013]*
  Let f be a ***configuration function****, and let $X = f(Z_1, \ldots, Z_n)$, where $Z_1, \ldots, Z_n$ are* ***independent*** *random variables. Then*

  $$Var(f(Z)) \leq \mathbb{E}\left[f(Z)\right].$$

2.3 VC-Dimension and Growth Function

- **Example** (***VC Dimension***)
  Let $\mathcal{H}$ be an arbitrary collection of subsets of $\mathcal{X}$, and let $x = (x_1, \ldots, x_n)$ be a vector of n points of $\mathcal{X}$. Define the ***trace*** of $\mathcal{H}$ on x by

  $$\mathrm{tr}\,(x) = \{A \cap \{x_1, \ldots, x_n\} : A \in \mathcal{H}\}.$$

  The shatter coefficient, (or *Vapnik-Chervonenkis* ***growth function***) of $\mathcal{H}$ in x is $\tau_{\mathcal{H}}(x) = |\mathrm{tr}\,(x)|$, *the size of the trace.* $\tau_{\mathcal{H}}(x)$ is the number of different subsets of the n-point set $\{x_1, \ldots, x_n\}$ generated by intersecting it with elements of $\mathcal{H}$. A subset $\{x_{i_1}, \ldots, x_{i_k}\}$ of $\{x_1, \ldots, x_n\}$ is said to be ***shattered*** if $2^k = T(x_{i_1}, \ldots, x_{i_k})$.

  The VC dimension $D(x)$ *of* $\mathcal{H}$ (with respect to x) is the *cardinality* k of *the largest shattered subset of* x. From the definition it is obvious that $f(x) = D(x)$ is a ***configuration function*** (associated with the property of "***shatteredness***") and therefore if $X_1, \ldots, X_n$ are *independent random variables*, then

  $$\mathrm{Var}(D(X)) \leq \mathbb{E}\left[D(X)\right].$$

References

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.