Concentration Inequalities & Stein’s Method (Fall 2019)  Lecture: 02

Concentration function and Cramér-Chernoff bound

Lecturer: Partha S. Dey  Date: Aug 29, 2019  Scribe: K.S. Krishnan <kesavsk2@illinois.edu>

In this lecture, two methods of finding sub-Gaussian tail bounds are highlighted. In particular, the concepts introduced are the use of Talagrand’s concentration function and the Cramér-Chernoff bound. In order to sketch a brief heuristic, if we intend to summarize the behavior of a random variable by a few estimates, then broadly speaking, the mean or median highlight the central behavior, the fluctuations about the central values are broadly governed by the variance. Concentration refers to studying and trying to bound the tail/extremal behavior.

2.1 Methods for Sub-Gaussian Random Variables

2.1.1 Talagrand’s concentration function

In the last class, Levy’s concentration function was defined on the real line. However, as defined, it is not suitable for the settings of higher dimensional or general metric spaces.

Definition 2.1 (Talagrand’s concentration function). Let $(\mathcal{X}, \mathcal{F}, d, \mu)$ be a metric space equipped with metric $d$ and probability measure $\mu$. Let $A \in \mathcal{F}$ be such that $\mu(A) \geq \frac{1}{2}$ and $A_t := \{x \mid \exists y \in A, d(x, y) \leq t\}$ be the $t$-fattening of $A$. Talagrand’s concentration function $\alpha(t)$ is defined as

$$\alpha(t) := \sup_{A : \mu(A) \geq \frac{1}{2}} \mu((A_t)^c). \quad (2.1)$$

The following example illustrates the use of the Talagrand concentration function to obtain tail bounds when our metric space valued random variable is mapped to the real line via a Lipschitz function. It is possible to obtain a bound on the upper tail in terms of deviation from the median.

Example 2.2. Let $f : \mathcal{X} \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant $k$. Let $Y \sim \mu$ and $X = f(Y)$. Let $A := \{y \mid f(y) \leq MX\}$. First, note that $A^t \subseteq \{y \mid f(y) \leq MX + kt\}$. This follows because if $y \in A^t$, there exists $z \in A$ such that $d(z, y) \leq t$ and thus

$$f(y) = f(z) + (f(y) - f(z)) \leq f(z) + kt \leq MX + kt$$

$$\Rightarrow \{y \mid f(y) > MX + kt\} \subseteq (A^t)^c$$

$$\Rightarrow \mu\{X > MX + kt\} \leq \mu((A^t)^c) \leq \alpha(t).$$

It is possible to explicitly evaluate the Talagrand function in certain cases, such as the case of the Gaussian distribution on $\mathbb{R}^N$. This is aided by the Gaussian isoperimetric Inequality.

Theorem 2.3 (Gaussian Isoperimetric Inequality). Let $\mathcal{X} = \mathbb{R}^N$ where $1 \leq N < \infty$. Let $\mu = \bigotimes_{n=1}^{N} \gamma$ where $\gamma$ is the standard Gaussian measure on $\mathbb{R}$. Then for all $A \in \mathcal{B}_{\mathbb{R}^N}$ and any half plane $H$ with $\mu(A) = \mu(H)$, we have

$$\mu(A^t) \leq \mu(H^c). \quad (2.2)$$
If equality holds for some $t > 0$, then $A$ is a half space.

Notes to keep in mind: The Gaussian isoperimetric inequality enables us to deal exclusively with half spaces. Let $H$ denote a half space. Note that for any $H$, there is a rotation matrix $R \in SO_n$ such that $R(H) = \{ x \mid x_1 < z \}$ for some $z \in \mathbb{R}$. Now note that the Gaussian density is invariant under rotations and thus $\mu(H) = \mu(R(H)) = \Phi(z)$. Further it is clear that $R(H_t) = \{ x \mid x_1 < t + z \}$.

**Corollary 2.4.** For the Gaussian measure, $\alpha(t) = P(Z > t)$ for all $t > 0$ where $Z \sim N(0, 1)$.

We may now combine the results from the theorem above and the example prior to obtain the following

**Theorem 2.5.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a 1-Lipschitz function and let $Z_1, Z_2, \ldots, Z_N$ be iid standard Gaussian random variables. Then, for $Y = f(Z_1, Z_2, \ldots, Z_N)$ and any $t > 0$ we have

$$P(Y \geq MY + t) \quad P(Y \leq MY - t) \leq e^{-\frac{t^2}{2}}$$

Over here we have used the tail bound on the normal CDF, namely $P(Z < t) \leq e^{-\frac{t^2}{2}}$.

### 2.1.2 Moment Generating Function Methods

In the last lecture, the moment generating function was defined as $m_X(t) := E[e^{tX}]$. We use the MGF to obtain a tail bound through the exponential Markov inequality. Since the exponential function is monotone, we have that $P[X \geq x] = P[e^{tX} \geq e^{tx}]$. Applying Markov’s Inequality to the latter term, we obtain

$$P[X \geq x] \leq m_X(t)e^{-tx}$$

For the subsequent parts we assume that the MGF exists and is finite for $t \in [-\delta, \delta]$ for some $\delta > 0$. We can optimise the inequality above to get a sharper bound for the tail, and for those purposes, we define the log moment generating function or cumulant generating function as follows

**Definition 2.6** (Log-moment generating function). Denoted $\varphi_X(t) = \log m_X(t)$

Before proceeding further note that the log moment generating function behaves nicely for sums of independent random variables. This is because of the product structure of the moment generating function, as discussed in the previous lecture. Thus, if $X_1, X_2, \ldots, X_N$ are independent such that $\varphi_{X_i}(t)$ exists on some common interval, then $\varphi_{S_N}(t) = \sum_{i=1}^N \varphi_{X_i}(t)$. Further, note that through Holder’s inequality, we can establish that $\varphi_X(t)$ is convex in $t$. It is convenient now to read the MGF bound as

$$P[X \geq x] \leq e^{-(tx - \varphi_X(t))}$$

Finding $\sup_x (tx - \varphi_X(t))$ yields the sharpest inequality of the form above. We denote the supremum as $\varphi_X^*(x)$. Note that this new function is also convex in $x$, and is known as the Legendre transform, or convex dual of $\varphi_X$.

**Theorem 2.7** (Cramér-Chernoff Bound). For $x \geq E[X]$, we have $P[X \geq x] \leq e^{-\varphi_X^*(x)}$
In particular, if we have \( Z \sim N(0, \sigma^2) \), it is easy to calculate the log moment generating function \( \varphi_Z(t) = \frac{\sigma^2 t^2}{2} \), and therefore the Legendre dual which turns out to be \( \varphi^*_Z(x) = \frac{\sigma^2 x^2}{2} \). Thus we have obtained a tail bound identical to the approach prior.

**Definition 2.8** (Sub Gaussian random variables). A random variable \( X \) is said to be sub-Gaussian with variance \( v \) if \( \varphi_X(t) \leq \frac{vt^2}{2} \) for all \( t \in \mathbb{R} \). We say that \( X \in \mathcal{G}(v) \).

**Theorem 2.9.** Let \( X_i \in \mathcal{G}(v_i) \) with \( i = 1, 2, \ldots, N \) be independent. Then \( S_N = \sum_{i=1}^{N} X_i \) is sub Gaussian with \( S_N \in \mathcal{G}(\sum_{i=1}^{N} v_i) \).

This theorem is an immediate consequence of independent sums and the Chernoff bound.

**Example 2.10.** Let \( Z \sim N(0, 1) \) and let \( X = Z^2 \). We can evaluate \( m_X(t) = \frac{1}{\sqrt{1-2t}} \) for \( t < \frac{1}{2} \) and \( \infty \) otherwise. Now note that \( \mathbb{P}[Z \geq x] = \frac{1}{2} \mathbb{P}[X \geq x^2] \). We can apply Chernoff’s bound to the right side to obtain that \( \mathbb{P}[X \geq x] \leq \frac{1}{2} e^{-\frac{1}{2}(2\log(x)+x^2-1)} \), which yields a sharper tail bound for the normal distribution. If we were to apply the Chernoff bound to \( \mathbb{P}[X \geq x] \), we can show that \( X \) has sub exponential tail.

**Exercise 2.1.** Let \((X_1 \ldots X_n)\) be a vector whose components are iid Gaussian. Let \( A \) be a symmetric matrix. Find an upper bound for \( \mathbb{P}(Y \geq x) \) where \( Y = \langle X, AX \rangle \). (Hint: Realize that \( X \) and \( OX \) have the same distribution where \( O \in O_n(\mathbb{R}) \))

**Example 2.11.** Let \( X \sim \text{Poi}(\lambda) \), that is \( \mathbb{P}[X = k] = \frac{e^{-\lambda} \lambda^k}{k!} \). By manipulating the power series we find that \( m_X(t) = e^{-\lambda+\lambda e^t} \). Let \( \bar{X} \) denote the centered version of \( X \), in which case we find that \( \varphi_X(t) = \lambda(e^t-t-1) \).

Proceeding to find the Legendre transform and thus find the Chernoff bound, we find that the optimal \( t \) is of the form \( t^* = \log(1 + \frac{z}{\lambda}) \). Let \( z = \frac{x}{\lambda} \). We find that \( \varphi_X^*(\lambda z) = \lambda(z \log(z+1) - z + \ln(z)) = \lambda((1+z) \log(1+z) - z) \).

In particular, through Taylor expansion, it can be seen that in the regime of \( \lambda \) large, the behavior is approximately Gaussian, as we lose the term linear in \( z \) and neglect higher order terms.

The observation above leads immediately to

**Exercise 2.2.** Show that for \( X \sim \text{Poi}(\lambda) \), \( \lim_{\lambda \to \infty} \frac{X-\lambda}{\sqrt{\lambda}} = Z \) in distribution, where \( Z \sim N(0, 1) \).