AMC TALK: BERNSTEIN’S INEQUALITY AND JOHNSON-LINDENSTRAUSS RANDOM PROJECTIONS

YIMING XU

ABSTRACT. In this talk I will introduce how to use an old technique in large deviation theory to derive the famous Bernstein’s inequality. Based on that, we will see how the renowned Johnson-Lindenstrauss random projection lemma in data compression comes into display. I will explain how this lemma can also be seen from the perspective of geometric functional analysis as well as the possible extension to a more general setting.

1. DERIVATION OF BERNSTEIN’ S INEQUALITY

To begin with, let us go back to the original proof of the weak law of large numbers for $L^2$ i.i.d. sequence $\{X_n\}$: For any $t > 0$, and $S_n = X_1 + \ldots + X_n$,

\[ P\left\{ \frac{S_n}{n} - E[X] \geq t \right\} = \frac{Var[S_n]}{n^2 t^2} = \frac{Var[X]}{nt^2} \to 0 \quad (1.1) \]

as $n \to \infty$. In fact, (1.1) can be essentially improved if higher moments exist. Suppose that a random variable $Z$ is not only in $L^2$ but also in $L^p$, $p > 2$. Then,

\[ P\{ |Z - E[Z]| > t \} \leq \min_p \frac{E|Z - E[Z]|^p}{t^p} \quad (1.2) \]

(1.2) is theoretically better than (1.1), but it is hard to see where improvement takes place. The following result which is due to Chernoff also uses a mixture of moments but gives a more tangible bound.

**Theorem 1.1.** Let $Z \in L^1$ and $\phi(t)$ be the logarithmic moment generating function (LMGF) of $Z - E[Z]$, that is, $\phi(\lambda) = \log E e^{\lambda (Z - E[Z])} = \log E e^{\lambda Z} - \lambda E[Z]$. Then

\[ P\{ Z - E[Z] \leq -t \} \leq e^{-\phi^*(-t)} \quad (1.3) \]

\[ P\{ Z - E[Z] \geq t \} \leq e^{-\phi^*(t)} \quad (1.4) \]

where $\phi^*$ is the Fenchel-Legendre transform of $\phi$. Particularly,

\[ P\{ |Z - E[Z]| \geq t \} \leq e^{-\phi^*(-t)} + e^{-\phi^*(t)} \quad (1.5) \]

**Proof.** For simplicity we assume $E[Z] = 0$. For $\lambda < 0$,

\[ P\{ Z \leq -t \} = P\{ e^{\lambda Z} \geq e^{-\lambda t} \} \leq e^{\lambda t + \phi(\lambda)}. \]

We can choose $\lambda \in (-\infty, 0)$ such that the right-hand side is minimized. This is equivalent to finding $\lambda \in (-\infty, 0)$ such that $(\lambda t + \phi(\lambda))$ is maximized. Using Jensen’s inequality it easy to see that $\lambda \in (-\infty, 0)$ can be replaced by $\lambda \in \mathbb{R}$. Hence, $\sup_{\lambda < 0}(\lambda t + \phi(\lambda)) = \sup_{\lambda \in \mathbb{R}}(\lambda t + \phi(\lambda)) = \phi^*(-t)$. This proves (1.3). (1.4) follows similarly. \qed

In order to apply Chernoff’s inequality one usually needs to know some information on the convex dual of the LMGF $\phi$ of $Z - E[Z]$. This gives the motivation for the following definition:
Definition 1.2. A centered random variable $Z$ is said to be sub-gaussian with variance factor $v$ if its LMGF $\phi$ satisfies $\phi(\lambda) \leq \lambda^2 v/2$ for all $\lambda \in \mathbb{R}$. We denote the set of such random variables by $\mathcal{G}(v)$.

It follows from definition that sum of independent sub-gaussian random variables is sub-gaussian. It is also easy to check that for $Z \in \mathcal{G}(v)$, $\text{Var}[Z] \leq v$.

Observe that if $f \leq g$, then $f^* \geq g^*$. Thus for sub-gaussian random variables $Z$ we have the following concentration inequality based on (1.5):

\begin{equation}
\mathbb{P}\{|Z| \geq t\} \leq 2e^{-\frac{t^2}{2v}}.
\end{equation}

(1.6) turns out to be an equivalent characterization of sub-gaussian random variables.

Theorem 1.3. The implication chain holds: $Z \in \mathcal{G}(v) \Rightarrow \mathbb{P}\{Z \geq t\} \wedge \mathbb{P}\{-Z \leq t\} \leq e^{-\frac{t^2}{2v}} \Rightarrow \mathbb{E}[Z^2n] \leq 2q!(2v)^q \leq q!(4v)^q \Rightarrow \mathbb{E}e^{\frac{t^2}{v}} \leq 1 \Rightarrow Z \in \mathcal{G}(16v)$.

Bounded random variables are an important class of sub-gaussian random variables and their corresponding concentration phenomenon is given by the following lemma.

Lemma 1.4. Suppose that $Y \in [a, b]$, where $a, b \in \mathbb{R}$. Then $Y - \mathbb{E}[Y] \in \mathcal{G}((b - a)^2/4)$, i.e., $\mathbb{P}\{|Y - \mathbb{E}[Y]| \geq t\} \leq 2e^{-\frac{2t^2}{(b-a)^2}}$.

Proof. Taylor expanding the LMGF $\phi(\lambda)$ of $Y - \mathbb{E}[Y]$ at $\lambda = 0$,

$$\phi(\lambda) = \phi(0) + \phi'(0)\lambda + \frac{1}{2}\phi''(\xi(\lambda))\lambda^2,$$

where $\xi$ is between 0 and $\lambda$. The first two terms on the right-hand side are 0. $\phi''(\xi(\lambda))$ in the last term is the variance of $Y$ under a new probability measure $\tilde{\mathbb{P}}$ such that $d\tilde{\mathbb{P}}_Y/d\mathbb{P}_Y \propto e^{\xi x}$, thus $\phi(\lambda) \leq \lambda^2 (b - a)^2/8$, as desired. \hfill $\square$

Applying Lemma 1.4 to the sum of independent bounded random variables $S_n = X_1 + \ldots + X_n$, where $X_i \in [a_i, b_i]$, we have

\begin{equation}
\psi_{S_n - \mathbb{E}[S_n]}(\lambda) \leq \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2,
\end{equation}

(1.7)

Hence,

\begin{equation}
\mathbb{P}\{|S_n - \mathbb{E}[S_n]| \geq t\} \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2}}.
\end{equation}

(1.8)

Inequality (1.8) is the known as Hoeffding’s inequality. In general, the variance of $S_n$ may be much smaller than the cheap bound $\sum_{i=1}^{n} (b_i - a_i)^2$, therefore some sharper estimates are called for.

Theorem 1.5. Let $X_1, \ldots, X_n$ be independent random variables with finite variance such that $X_i \leq b$ for some $b > 0$ almost surely for all $i \leq n$. Let $S_n = \sum_{i=1}^{n} X_i$. Then for $\varepsilon > 0$, $\varepsilon > 0$,

\begin{equation}
\mathbb{P}\{S_n - \mathbb{E}[S_n] \geq t\} \leq e^{-\frac{t^2}{2h(t/v)}},
\end{equation}

(1.9)

where $v = \sum_{i=1}^{n} \mathbb{E}[X_i^2]$ and $h(u) = (1 + u)\log(1 + u) - u$ for $u > 0$.

(1.9) is known as Bennett’s inequality. Note that $h(u)$ in the above statement is the convex dual of the LMGF of a standard Poisson distribution. For the proof one actually shows that the LMGF of $S_n - \mathbb{E}[S_n]$ can be bounded by $\frac{1}{2v}\phi(b\lambda)$, where $\phi(\lambda)$ is the LMGF of a standard Poisson distribution.
If we use an elementary inequality for \( h(u) \):

\[
(1.10) \quad h(u) \geq \frac{u^2}{2(1 + u/3)},
\]

we can get

\[
(1.11) \quad \mathbb{P}\{S_n - \mathbb{E}[S_n] \geq t\} \leq e^{-\frac{t^2}{2(v + ct)}}.
\]

(1.11) is the famous Bernstein’s inequality. In fact, Bernstein’s inequality holds in a more general setting than the bounded case, if assuming appropriate control of moments.

Before stating the theorem, we need one more definition.

**Definition 1.6.** A centered random variable \( X \) is said to be sub-gamma on the right tail with variance factor \( v \) and scale parameter \( c \) if its LMGF \( \phi \) satisfies

\[
\phi(\lambda) \leq \frac{\lambda^2 v}{2(1 - c\lambda)}
\]

for \( 0 < \lambda < 1/c \). The set of such random variables is denoted by \( \Gamma_+(v, c) \). A centered random variable \( X \) is said to be sub-gamma if \( X \in \Gamma(v, c) := \Gamma_+(v, c) \cap (-\Gamma_+(v, c)) \).

It can be checked that for a Gamma random variable \( Y \) with parameters \( (a, b) \), its centered version \( Y = Y - \mathbb{E}Y \in \Gamma(ab^2, b) \). One should note that \( Y \) is not symmetric around its mean; particularly, it is more concentrated on the left side of its mean, and the corresponding tail probability decays as sub-gaussian.

Analogous to the case of sub-gaussian random variables, the following theorem gives some equivalent characterizations of the sub-gamma property.

**Theorem 1.7.** The implication chain holds: \( X \in \Gamma(v, c) \Rightarrow \mathbb{P}\{X \geq \sqrt{2vt + ct}\} \land \mathbb{P}\{X \leq -\sqrt{2vl - ct}\} \leq e^{-t} \Rightarrow \mathbb{E}[Z^2] \leq q!(8v)^q + (2q)!(4c)^{2q} \Rightarrow Z \in \Gamma(4(8v + 16c^2), 8c) \).

A nice consequence of sub-gamma property is Bernstein’s inequality. Let \( X \in \Gamma(v, c) \) and \( \phi(t) \) be its LMGF. Then,

\[
(1.12) \quad \phi^*(-t) = \left(\frac{v \lambda^2}{2(1 - c\lambda)}\right)^*(t) = \frac{v}{c^2} r\left(\frac{ct}{v}\right)
\]

where \( r(u) = 1 + u - \sqrt{1 + 2u} \). Using the elementary inequality that

\[
(1.13) \quad r(u) \geq \frac{u^2}{2(1 + u)}, \quad u > 0
\]

therefore

\[
(1.14) \quad \phi^*(-t) \geq \frac{t^2}{2(v + ct)}.
\]

Since \( X \) is sub-gamma on both sides, \( \phi^*(-t) = \phi^*_{-X}(t) \leq \frac{t^2}{2(v + ct)} \). It follows from (1.5) that

\[
(1.15) \quad \mathbb{P}\{|X| \geq t\} \leq 2e^{-\frac{t^2}{2(v + ct)}}.
\]

It should be noted that (1.15) looks almost like Bernstein’s inequality, except that \( X \) may not be a sum of independent random variables. The following theorem gives a sufficient condition for independent random variables such that their centered sum is sub-gamma.
Theorem 1.8. Let $X_1, \ldots, X_n$ be independent random variables. Suppose that there exist $c$ and $v$ such that $\sum_{i=1}^n \mathbb{E}[X_i^2] \leq v$ and for $q \geq 3$,

\begin{align}
\sum_{i=1}^n \mathbb{E}[(X_i)^q] &\leq \frac{q!}{2} v c^{q-2} \\
\sum_{i=1}^n \mathbb{E}[(X_i)^2] &\leq \frac{q!}{2} v c^{q-2}
\end{align}

If $S_n = \sum_{i=1}^n X_i$, then for $t > 0$,

\begin{equation}
\mathbb{P}\{|S_n - \mathbb{E}S_n| \geq \sqrt{2vt} + ct\} \leq 2e^{-t} \tag{1.18}
\end{equation}

Proof. (1.16) and (1.17) together imply that $S_n - \mathbb{E}S_n \in \Gamma(v, c)$. Taylor expansion of the moment generating function of $S_n - \mathbb{E}[S_n]$ gives details. \hfill \Box

A concept related to sub-gaussian and sub-gamma is the sub-exponential property. A random variable $X$ is said to have sub-exponential distribution if $\mathbb{P}\{|X| \geq t\} \leq 2e^{-t/K_1}$ for some $K_1 > 0$. Two equivalent characterizations of sub-exponential property are that $X$ has bounded 1-Orlicz norm, that is, $\|X\|_{\Phi_1} = \inf\{t : \mathbb{E}e^{\|X\|/t} \leq 2\}$, or $\sup_{p \in \mathbb{Z}^+} \|X\|_p < \infty$. It follows from definition that if a random variable is sub-exponential, then it is sub-gamma. Since, it is easy to check that sum of independent sub-exponential random variables is sub-exponential, Bernstein’s inequality holds in this case. However, it is worth pointing out that due to the structure of $S_n$, a universal sub-gaussian tail will finally appear as $n$ becomes large. Thus the Bernstein’s inequality in such case looks like

\begin{equation}
\mathbb{P}\{|S_n| \geq t\} \leq \exp\left\{-c \min\left\{\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\Psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\Psi_1}}\right\}\right\} \tag{1.19}
\end{equation}

for some $c > 0$. The sub-gaussian tail can be explained by central limit theorem.

Bernstein’s inequality also holds in higher dimensions. Let $X_1, \ldots, X_n$ be mean-zero, symmetric and independent $n \times n$ matrices, such that $\|X_i\|_2 \leq K$ for all $i = 1, \ldots, n$. Then,

\begin{equation}
\mathbb{P}\{\|\sum_{i=1}^N X_i\|_2 \geq t\} \leq 2ne^{-\frac{\sigma^2 t^2}{2\mathbb{E}[\sum_{i=1}^N X_i^2]}} \tag{1.20}
\end{equation}

where $\sigma^2 = \|\sum_{i=1}^N \mathbb{E}[X_i^2]\|_2$. No independence between entries of $X_i$ is assumed here. The proof is based on a series of matrix inequalities and omitted here. As a remark, we note that an immediate consequence of (1.20) is that

\begin{equation}
\mathbb{E}[\|\sum_{i=1}^N X_i\|_2] = \int_0^\infty \mathbb{P}\{\|\sum_{i=1}^N X_i\|_2 \geq t\} dt \lesssim \sigma \sqrt{\log n + K \log n}. \tag{1.21}
\end{equation}
2. JOHNSON-LINDENSTRAUSS RANDOM PROJECTIONS

**Theorem 2.1** (Johnson-Lindenstrauss random projections). Let \( A \) be a set containing \( n \) distinct points in \( \mathbb{R}^D \). Let \( W \) be a \( d \times D \) random matrix with independent entries. For every \( i, j \), assume that \( \mathbb{E}[W_{ij}] = 0, \mathbb{E}[W_{ij}^2] = 1 \) and \( W_{ij} \in \mathcal{G}(v), v \geq 1 \). Then, for every \( \varepsilon, \delta > 0 \), if \( d \) is chosen such that \( d \geq (2\varepsilon^2)^{-1}(8v^2 + v\varepsilon) \log \frac{n(n-1)}{\delta} \), then \( \tilde{W} := W/\sqrt{d} \) is an \( \varepsilon \)-isometry on \( A \) with probability at least \( 1 - \delta \).

**Proof.** We first note that for any fixed unit vector \( v \in \mathbb{R}^D \), \( \mathbb{E}[||\tilde{W}v||^2] = 1 \), because of independence between entries of \( W \). The desired result would come easily supposing measure concentrates around the mean of the random variable \( ||\tilde{W}v||^2 \). Note that \( ||\tilde{W}v||^2 \) is essentially the sum of independent square of sub-gaussian random variables, therefore sub-exponential. This means that Bernstein’s inequality is available for help in this case.

To complete the details, we write down

\[
||\tilde{W}v||^2 = \frac{1}{d} \sum_{i=1}^{d} \left( \sum_{j=1}^{D} W_{ij} v_j \right)^2 = \sum_{i=1}^{d} \left( \frac{Z_i}{\sqrt{d}} \right)^2,
\]

where \( Z_i = \sum_{j=1}^{D} W_{ij} v_j \). It is easy to see from the LMGF that \( Z_i \in \mathcal{G}(v) \), therefore \( Z_i/\sqrt{d} \in \mathcal{G}(v/d) \). Hence,

\[
\sum_{i=1}^{d} \mathbb{E}\left[ \frac{Z_i^4}{d^2} \right] = \frac{16v^2}{d},
\]

\[
\sum_{i=1}^{d} \mathbb{E}\left[ \frac{Z_i^{2q}}{d^q} \right] \leq \frac{2q!(2v)^q}{d^{q-1}}, \quad q \geq 3.
\]

Applying Theorem 1.8 to \( ||\tilde{W}v||^2 - 1 \) with \( v = 16v^2/d, c = 2v/d \), and using (1.15),

\[
P\{||\tilde{W}|| - 1| \geq \varepsilon\} \leq P\{||\tilde{W}||^2 - 1| \geq (2 + o(1))\varepsilon\} \leq 2 \exp\left( -\frac{2\varepsilon^2 d}{8v^2 + v\varepsilon}\right).
\]

Using a union bound for all the \( v \)'s we need to take care of, i.e., all the elements of the set \( A - A \), the upper bound for not preserving the distance within error \( \varepsilon \) is given by

\[
\kappa = n(n-1) \exp\left( -\frac{2\varepsilon^2 d}{8v^2 + v\varepsilon}\right),
\]

which is required to be less than \( \delta \). Solving this inequality gives

\[
d \geq \frac{8v^2 + v\varepsilon}{2\varepsilon^2} \log \frac{n(n-1)}{\delta}.
\]

\( \square \)

**Remark 2.2.** (2.1) behaves as \( \sim O(\log n/\varepsilon^2) \) and is a dimension-free result which does not depend on the dimension of the original space.
Remark 2.3. Let us write $W = (W_i)_{1 \leq i \leq d}$. Suppose that the entries of $W$ are i.i.d. Then, by the law of large numbers, for each $1 \leq i \neq j \leq d$,

$$\cos \langle W^i, W^j \rangle = \frac{\langle W^i, W^j \rangle}{\|W^i\| \|W^j\|}$$

$$= \frac{1}{\sqrt{d}} \sum_{r=1}^{d} W^i_r W^j_r$$

This suggests that we can roughly view $W$ as $\sqrt{d}$ times an orthogonal projection matrix that projects a vector onto a $d$-dimensional random subspace in $\mathbb{R}^d$ under the gaussian measure. With proper scaling, this is the same as projecting the set onto a uniformly-chosen $d$-dimensional subspace of $\mathbb{R}^d$.

Remark 2.4. Note that in this case all the vectors we need to take care of are exactly the elements of the set $A - A$, which is a finite bounded symmetric set around the origin. In fact, the Johnson-Lindenstrauss random projection lemma can be generalized to the case where $A$ is not necessarily finite. The corresponding cardinality in our bound will be replaced by the Gaussian complexity, which is defined as $\gamma(A - A) = \mathbb{E}[-\sup_{x, y \in A - A} \langle Z_n, x \rangle] = \mathbb{E}[-\sup_{x \in A - A} \langle Z_n, x \rangle]$, where $Z \sim \mathcal{N}(0, I_n)$. The corresponding concentration result becomes

$$\mathbb{P}\{ \sup_{x \in A - A} ||W x||_2 - ||x||_2 \leq C[\gamma(A - A) + u \cdot \text{rad}(A - A)] \} \leq 1 - 2 \exp(-u^2).$$

for $u > 0$ and universal constant $C$ depending only on $W$. Using the fact that $\text{rad}(A - A) \leq \gamma(A - A)$, the right-hand side in the bracket is $\leq u \cdot \gamma(A - A)$.

In fact, Gaussian complexity appears in the beautiful (dual version of) Dvoretzky theorem, which states that for any symmetric convex body $K$ in $\mathbb{R}^D$, with $\text{diam}(K) = 1$, the diameter of its random projection $PK$ onto a uniformly chosen $k$-dimensional subspace shrinks by $\sqrt{n/k}$ when $k \geq k^*$ and stabilizes to a ball of radius $O(M^*(K))$ when $k \leq k^*$, where $k^* = M^*(K)^2 n$, $M^*(K) = \int_{S^{n-1}} \max_{y \in K} |\langle x, y \rangle| \sigma(x)$. In words, a phase transition occurs at the critical dimension $k^*$. Let $z = (z_1, ..., z_n)$ be the $n$-dimensional canonical standard Gaussian distribution in $\mathbb{R}^n$,

$$M^*(K) = \int_{S^{n-1}} \max_{y \in K} |\langle x, y \rangle| \sigma(x)$$

$$= \int_{\mathbb{R}^n} \max_{y \in K} |\langle \frac{z}{||z||_2}, y \rangle| d\gamma_n(z)$$

$$\approx \int_{\mathbb{R}^n} \max_{y \in K} |\langle \frac{z}{\sqrt{n}}, y \rangle| d\gamma_n(z)$$

$$= \frac{1}{\sqrt{n}} \gamma(K),$$

for large $n$. Therefore, $k^* = \gamma(K)^2$. This result, in some sense, suggests that in order have almost isometric embedding structure, the dimension of the embedding space cannot be too small, otherwise the projected set looks just like a ball regardless of its original shape. In the case of Johnson-Lindenstrauss random projections, the Gaussian complexity of $A - A$ is $\sim \mathcal{O}(\sqrt{\log n})$, the heuristic value $k^* \sim \mathcal{O}(\log n)$ ($A - A$ is not a symmetric convex body so it is only in the heuristic sense), which is close the bound $\mathcal{O}(\log n / \varepsilon^2)$. This suggests
that the bound obtained in Johnson-Lindenstrauss lemma is close to optimal. Indeed, Alon in [7] demonstrated that any embedding with $1 + \varepsilon$ distortion requires dimension at least $O(\log n/(\varepsilon^2 \log 1/\varepsilon))$, thus showing that the Johnson-Lindenstrauss lemma is nearly tight.

**Remark 2.5.** There is no similar result holding in the case of $l_1$, meaning that it is impossible to do dimension reduction in $l_1$. We give a brief explanation based on [4]. Note that any recursive diamond graph $G_k$ is $O(1)$-isomorphic to a subset $X \subseteq L_1$ of size $n$, where $k = O(\log n)$. In [4], it was shown that for $1 < p \leq 2$, any embedding of $G_k$ into $L_p$ incurs distortion at least $\sqrt{1 + (p - 1)k}$. We also know that $l_1^d$ is $O(1)$-isomorphic to $l_{p(d)}^d$ with $p(d) = 1 + 1/\log n$. Therefore, any embedding of $G_k$ into $l_1^d$ incurs distortion at least $\sqrt{1 + (p(d) - 1)k} = \sqrt{1 + k/\log d}$. If we require $\sqrt{1 + (p(d) - 1)k} = \sqrt{1 + k/\log d} \leq 1 + \varepsilon$ for small $\varepsilon > 0$, then $d \sim O(n^{O(1/\varepsilon^2)})$. A more detailed treatment for $l_p(p > 2)$ can be found in [6].

**REFERENCES**


