Quadratic forms and Symmetrization, Chap 6

High Dim Probability & Linear Algebra for ML and Sig Proc

Namrata Vaswani

Iowa State University

General

1. Chaos: $X^\top \mathbf{A} X$ with X being a r. vector with independent, zero-mean, coordinates.
2. Clearly $\mathbb{E}[Chaos] = trace(\mathbf{A})$ if $\mathbb{E}[X_i^2] = 1$ (unit variance also). Without this, $\mathbb{E}[Chaos] = \sum_i a_i i \mathbb{E}[X_i^2]$
3. Concentration bounds not so easy; use the "decoupling trick": replace Chaos by $X^\top A X'$ where X' is an indep copy of X.
4. Jensen's inequality: for convex F, $F(\mathbb{E}[X]) \le \mathbb{E}[F(X)]$ (recall)

Main results

1. Theorem 6.1.1 / Remark 6.1.3: Decoupling
   Let X be an n-length vector with independent zero-mean coordinates and $\mathbf{A}$ a matrix with ZEROS on DIAGONAL. Then for every convex func F,

   $$\mathbb{E}[F(X^\top \mathbf{A} X)] \le \mathbb{E}[F(4X^\top \mathbf{A} X')]$$

   (NOTE: no subG or other distribution assumption needed)
   (NOTE 2: I had a MAJOR MISTAKE in DECOUPLING RESULT – NOW FIXED - RHS expression is also $\mathbb{E}[F(.)]$: the expectation is outside)
   More generally, for any $\mathbf{A}$,

   $$\mathbb{E}[F(\sum_{i \ne j} a_{ij} X_i X_j)] \le \mathbb{E}[F(\sum_{ij} a_{ij} X_i X_j')] = \mathbb{E}[F(4X^\top \mathbf{A} X')]$$

   Do Ex 6.1.4, 6.1.5: easy modifications of above proof.

2. Hanson-Wright inequality: concen bound for chaos: **this requires subGaussian distrib**
Let X be a n-length vector with indep zero-mean, subG-K coordinates. Then

$$Pr(|X^\top \mathbf{A} X - \mathbb{E}[X^\top \mathbf{A} X]| \geq t) \leq 2 \exp\left(-min\left(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t^2}{K^2 \|A\|}\right)\right)$$

K: max of all subG norms of all vectors.
W.l.o.g. can assume $K \geq 1$: reason is simpler than the one I earlier gave: for subG, we always use an upper bound on subG norm, so even the true max subG norm is 0.2, it is upper bounded by 1. We use $K \geq 1$ in the last step to argue that $K^4 \geq K^2$

   1. Application: Bound $\|\mathbf{B}X\|$ for a given matrix $\mathbf{B}$ and for a r vector X having independent, zero-mean, unit variance *subG*(K) entries (Theorem 6.3.2). Idea: $\|\mathbf{B}X\|^2 = X^\top(\mathbf{B}^\top \mathbf{B})X$ = chaos with $\mathbf{A} \equiv \mathbf{B}^\top \mathbf{B}$.

3. Lemma 6.1.2: Let Y, Z indep and $\mathbb{E}[Z] = 0$. Then for every convex $F(.)$,

$$F(Y) \leq \mathbb{E}[F(Y+Z)|Y]$$

and so

$$\mathbb{E}[F(Y)] \leq \mathbb{E}[F(Y+Z)]$$

Proof: $F(Y) = F(Y + \mathbb{E}[Z]) = F(Y + \mathbb{E}[Z|Y]) = F(\mathbb{E}[Y+Z|Y]) \leq \mathbb{E}[F(Y+Z)|Y]$
Use EZ=0; indep of Y, Z; cond on Y, Y is constant; Jensen.

4. Lemma 6.2.2: MGF of Gaussian chaos
   Let G, G' are independent and each is standard Gaussian vector. Then

$$\mathbb{E}[\exp(\lambda G^\top \mathbf{A} G')] \leq \exp(C\lambda^2 \|\mathbf{A}\|_F^2), \ \forall \ |\lambda| < c/\|A\|$$

   Proof: write SVD of **A**, use rotation invar of Gaussian, condition on X' and use expression for scalar Gaussians' MGF, finally use the fact that Gaussian-squared is sub-expo and use sub-expo property (MGF bound). Recall that $\|\mathbf{A}\|_F^2$ is sum of its singular values while $\|\mathbf{A}\|$ is its max singular value.
5. Lemma 6.2.3: Comparison lemma: MGF of subG chaos is upper bounded by that of Gaussian chaos
   Let X, X' independent, zero-mean, $subG(K)$ r vectors. Then,

$$\mathbb{E}[\exp(\lambda X^\top \mathbf{A} X')] \leq \mathbb{E}[\exp(\lambda (CK^2) G^\top \mathbf{A} G')]$$

   where G, G' are independent and both are standard Gaussian r. vectors.
   Proof: Recall that MGF of a standard Gaussian is $MGF(s) = \exp(s^2/2)$, and that of a zero mean variance v Gaussian is $\exp(s^2 v/2)$.

1. First condition on X', and use subG property followed by comparing with above to show that

$$\mathbb{E}_{X|X'}[\exp(\lambda X^\top \mathbf{A} X')] \leq \exp(\lambda^2 (CK^2) \|\mathbf{A}X'\|^2) = \mathbb{E}_{G|X'}[\exp((\lambda\sqrt{2C}K)(G^\top \mathbf{A} X')]$$

The last equality follows by using the fact that $(G^\top \mathbf{A} X')$ is zero-mean Gaussian with variance $v = \|\mathbf{A}X'\|^2$ and comparing second expression with its MGF

2. Thus,

$$\begin{aligned}
\mathbb{E}[\exp(\lambda X^\top \mathbf{A} X')] &= \mathbb{E}_{X'} \mathbb{E}_{X|X'}[\exp(\lambda X^\top \mathbf{A} X')] \\
&\leq \mathbb{E}_{X'} \mathbb{E}_{G|X'}[\exp((\lambda\sqrt{2C}K)(G^\top \mathbf{A} X')] \\
&= \mathbb{E}_{G} \mathbb{E}_{X'|G}[\exp((\lambda\sqrt{2C}K)(X'^\top \mathbf{A} G)] \\
&\leq \mathbb{E}_{G}[\exp((\lambda\sqrt{2C}K)^2 (CK^2) \|\mathbf{A}G\|^2] \\
&= \mathbb{E}_{G}\left[\mathbb{E}_{G'|G}\left[\exp\left(\sqrt{2(\lambda\sqrt{2C}K)^2 (CK^2)} G'^\top \mathbf{A} G\right)\right]\right] \\
&= \mathbb{E}[\exp(\lambda(\tilde{C}K^2) G'^\top \mathbf{A} G)]
\end{aligned}$$

second row used previous step, third row is Fubini, fourth row used subG property of X', fifth row compares with scalar Gaussian MGF of $G'^\top \mathbf{A} G$ given G (this is scalar Gaussian with variance $\|\mathbf{A}G\|^2$), last row simplifies

Proof of Decoupling result

1. Step 1: replace chaos by "partial chaos" (sum of disjoint sets of i,j)
   1. Let $I = \{i : \delta_i = 1\}$ and $\delta_i \stackrel{\text{iid}}{\sim} Bern(1/2)$ and indep of X. Clearly $\mathbb{E}_\delta[\delta_i(1-\delta_j)] = 1/4$ for $i \neq j$.
   2. Clearly $I^c = \{j : \delta_j = 0\} = \{j : 1-\delta_j = 1\}$ and so $\delta_i(1-\delta_j) \neq 0$ only if $i \in I, j \in I^c$.
   3. Fix X first. Then, $\sum_{i\neq j} a_{ij} X_i X_j = \sum_{i \neq j} 4\mathbb{E}[\delta_i(1-\delta_j)] a_{ij} X_i X_j = \mathbb{E}_\delta[\sum_{i \neq j} 4\delta_i(1-\delta_j) a_{ij} X_i X_j] = \mathbb{E}_I[\sum_{i\in I, j \in I^c} 4\delta_i(1-\delta_j) a_{ij} X_i X_j]$. Thus,

$$\sum_{i\neq j} a_{ij} X_i X_j = \mathbb{E}_I[\sum_{i \in I, j \in I^c} 4\delta_i(1-\delta_j) a_{ij} X_i X_j] = \mathbb{E}_I[\sum_{i\in I, j\in I^c} 4a_{ij} X_i X_j]$$

   4. Apply F, apply Jensen to get,

$$F(\sum_{i\neq j} a_{ij} X_i X_j) = F(\mathbb{E}_I[\sum_{i\in I, j \in I^c} 4a_{ij} X_i X_j]) \leq \mathbb{E}_I[F(\sum_{i\in I, j\in I^c} 4a_{ij} X_i X_j)]$$

   5. Take $\mathbb{E}[.]$ over X, then use Fubini to get

$$\mathbb{E}_X[F(\sum_{i\neq j} a_{ij} X_i X_j)] \leq \mathbb{E}_X[\mathbb{E}_I[F(\sum_{i\in I, j\in I^c} 4a_{ij} X_i X_j)]] = \mathbb{E}_I[\mathbb{E}_X[F(\sum_{i \in I, j \in I^c} 4a_{ij} X_i X_j)]]$$

6. Since *average* $\leq$ *max*, there is at least one I_0 s.t. the following is true

$$\mathbb{E}_I[\mathbb{E}_X[F(\sum_{i\in I, j\in I^c} 4a_{ij}X_iX_j)]] \leq \max_I \mathbb{E}_X[F(\sum_{i\in I, j\in I^c} 4a_{ij}X_iX_j)]]$$

$$= \mathbb{E}_X[F(\sum_{i\in I_0, j\in I_0^c} 4a_{ij}X_iX_j)]$$

Fix this I_0 for rest of the proof.

Thus, so far we have shown that

$$\mathbb{E}_X[F(\sum_{i\neq j} a_{ij}X_iX_j)] \leq \mathbb{E}_X[F(\sum_{i\in I_0, j\in I_0^c} 4a_{ij}X_iX_j)]$$

2. Replace the X_j by X_j'
   1. The RHS of above is a function of $X_{I_0}, X_{I_0^c}$, i.e. $RHS = g(X_{I_0}, X_{I_0^c})$. Since $X_{I_0}, X_{I_0^c}$ are independent of each other, we can replace the latter by $X'_{I_0^c}$ inside the expected value, i.e.,

$$\mathbb{E}_X[F(\sum_{i\in I_0, j\in I_0^c} 4a_{ij}X_iX_j)] = \mathbb{E}_X[F(\sum_{i\in I_0, j\in I_0^c} 4a_{ij}X_iX_j')]$$

3. Complete partial chaos to chaos by conditioning on $W := \{X_{I_0}, X'_{I_0^c}\}$ and then using Lemma
   1. Let $Y := \sum_{i \in I_0, j \in I_0^c} 4a_{ij} X_i X'_j$, $Z_1 := \sum_{i \in I_0, j \in I_0} 4a_{ij} X_i X'_j$, $Z_2 := \sum_{i \in I_0^c, j \in I_0} 4a_{ij} X_i X'_j$, $Z_3 := \sum_{i \in I_0^c, j \in I_0^c} 4a_{ij} X_i X'_j$ Notice that

$$\sum_{i,j} 4a_{ij} X_i X'_j = Y + Z_1 + Z_2 + Z_3$$

   2. Notice also that conditioned on W, $Y = h(W)$ is a constant, the randomness in Z_1 is due to X'_{I_0} (which is indep of W), that in Z_2 is due to $X_{I_0^c}, X'_{I_0}$ (which is indep of W), that is Z_3 is due to $X_{I_0^c}$ (which is indep of W), while $Y = h(W)$. Thus given W all the Z_i are indep of Y. And $\mathbb{E}[Z_i|W] = 0$ for all three of them.
      Thus, given W, $Z \equiv Z_1 + Z_2 + Z_3$ has zero mean and is indep of Y. This means we can apply the Lemma 6.1.2 conditioned on W

$$\mathbb{E}[F(Y)|W] \leq \mathbb{E}[F(Y+Z)|W] = \mathbb{E}[F(Y + Z_1 + Z_2 + Z_3)|W]$$

3. Now taking expectation over W,

$$\mathbb{E}[F(Y)] \leq \mathbb{E}[F(Y + Z_1 + Z_2 + Z_3)]$$

or

$$\mathbb{E}_X[F(\sum_{i \in I_0, j \in I_0^c} 4a_{ij}X_iX_j'))] \leq \mathbb{E}[F(\sum_{i,j} 4a_{ij}X_iX_j'))]$$

Combining the above three steps,

$$\mathbb{E}_X[F(\sum_{i \neq j} a_{ij}X_iX_j)] \leq \mathbb{E}[F(\sum_{i,j} 4a_{ij}X_iX_j'))]$$

Proof of Hanson-Wright

1. Split the probability into diagonal and off-diagonal (cross) terms.
2. Diagonal term: is a sum of independent sub-expo terms which we have handled before. Use sub-expo Bern inequality.
3. Off-diagonal term: bound using decoupling result, comparison lemma, MGF of Gaussian chaos lemma.

1. Basics: X is symmetric means $X, -X$ have same distribution. This is for zero-mean setting.
   More generally, we can say Y is symmetric about its mean if $X = Y - \mathbb{E}[Y]$ is a symmetric r.v.
   1. Let X be any rv and ζ is *SymBern*. Then ζX and $\zeta |X|$ have same distribution.
   2. If X is symmetric, then it has same the distrib as ζX or $\zeta |X|$
   3. For any rv X, let X' be independent copy. Then $X - X'$ is symmetric.
      1. Thus, $X - X'$ and $\zeta(X - X')$ have same distribution.
   4. Let $X = [X_1, X_2 ... X_N]'$ be a r vector and X' its indep copy. Let ζ be a vector of indep symBern rvs.
      1. By earlier claims, $X_i - X_i'$ are symmetric and have same distrib as $\zeta_i(X_i - X_i')$
      2. If the different X_is are indep, then $X_i - X_i'$s are indep and so are $\zeta_i(X_i - X_i')$. In this case, $X - X'$ has same distrib as $\zeta. * (X - X')$.

2. Lemma 6.4.2 on Symmetrization (check that it also works for sums of random matrices) Let $X_1, X_2, ..X_N$ be independent zero-mean r. vectors and $\epsilon_1, \epsilon_2, ..\epsilon_N$ be indep symBern rvs indep of the X_is.

$$0.5\mathbb{E}[\|\sum_i \epsilon_i X_i\|] \leq \mathbb{E}[\|\sum_i X_i\|] \leq 2\mathbb{E}[\|\sum_i \epsilon_i X_i\|]$$

   Proof: uses above facts and Lemma 6.1.2: $F(Y) \leq \mathbb{E}[F(Y + \mathbb{E}[Z])|Y]$ if Y, Z indep and F convex , applied for $F(.) = \|.\|$.
   1. All the exercises are interesting
3. Theorem 6.5.1: bounding norm of r. matrix with not identically distrib entries. Let B is n x n symmetric matrix with entries on and above diagonal being indep and zero mean. Then

$$\mathbb{E}[\|B\|] \leq C\sqrt{\log n}\mathbb{E}[\max_i \|B^i\|]$$

   In above B^i is i-th row of B.
   1. This is tight up to log factor since $\|B\| \geq \max_i \|B^i\|$ and so this is true for their expected values too.
   2. Compare this with Cor 4.4.8
      - ★ Cor 4.4.8 needs that the entries are subG-K. This result does not.

Symmetrization III

- ★ The above result gives a tighter bound than Cor 4.4.8 (whose bound is $CK\sqrt{n}$) for when different rows have very different norms

3. Extend to non-symmetric or rectangular matrices: uses "dilation" trick: For any matrix G, define $B = [0, G; G^\top, 0]$, can show easily that B is symmetric with eigenvalues $\pm\sigma_i(G)$.

Proof:

1. symmetrization lemma and matrix Khintchine inequality Ex 5.4.13 which states

$$\mathbb{E}[\|\sum_i \epsilon_i A_i\|] \leq C\sqrt{1+\log n}\sqrt{\|\sum_i A_i^2\|}$$

here A_i are deterministic matrices.

2. Split B as

$$B = \sum_{i \leq j} Z_{ij}$$

where $Z_{ij} = B_{ij}(e_i e_j^\top + e_j e_i^\top)$ for $i < j$ and $= B_{ii} e_i e_i^\top$ for $i = j$.

3. Clearly these matrices are independent. So by symmetrization lemma,

$$\mathbb{E}[\|B\|] = \mathbb{E}[\|\sum_{i \leq j} Z_{ij}\|] \leq 2\mathbb{E}[\|\sum_{i \leq j} \epsilon_{ij} Z_{ij}\|]$$

4. Condition on Z_{ij}, apply matrix Khintchine, then take average over Zij to conclude

$$\mathbb{E}[\|B\|] \leq 2\mathbb{E}[\|\sum_{i\leq j} \epsilon_{ij} Z_{ij}\|] \leq C\sqrt{\log n}\mathbb{E}[\sqrt{\|\sum_{ij} Z_{ij}^2\|}]$$

Simplify and argue that $\sum_{ij} Z_{ij}^2$ is a diagonal matrix, thus its norm is its max magnitude entry.

4. Matrix Khintchine proof:
   1. follows from matrix Bernstein and integral identity.
5. Matrix completion application Theorem 6.6.1 : does not assume incoherence. Let $\mathbf{X}$ be nxn with rank r.
   Let $\hat{\mathbf{X}}$ be rank r approx of $Y = P_\Omega(X)$ where Ω is the observed entries set generated using the $Bern(p)$ model. Then,

$$\mathbb{E}[\frac{1}{n}\|\hat{\mathbf{X}} - \mathbf{X}\|_F] \leq C\sqrt{\frac{r\log n}{pn^2}}\|\mathbf{X}\|_{\max}$$

   1. If we use incoherence assumption, then from standard results, $\|\mathbf{X}\|_{\max} \leq (\mu r/n)\|\mathbf{X}\|$

Proof:

1. $\mathbb{E}[Y] = pX$, add subtract Y/p, use rank r approx property of $\hat{\mathbf{X}}$,
2. then we are left to bound $2\mathbb{E}[\|Y - pX\|]$.
3. To do this, use rectangular version of previous theorem.
4. Then, for a fixed i, bound the row or column norms using scalar bounded Bernstein or Chernoff inequality, union bound for their max, then integral identity to convert high probab bound to bound on $\mathbb{E}[.]$. See Ex 6.6.2
5. Finally pass to Frob norm by using the fact that $\hat{\mathbf{X}} - \mathbf{X}$ is at most rank $2r$.