ORF 550 KIRAN VODRAHALLI

MATRIX CONCENTRATION AND APPLICATIONS TO MACHINE LEARNING JANUARY 17, 2017

Contents

1 Introduction

In ORF 550, we have mostly studied the extensions of concentration inequalities via an approach like tensorization to achieve concentration bounds for high-dimensional vectors. In this short review, we will be interested in the background and application of deriving concentration inequalities in the spectral norm for sums of independent random matrices.

There are many applications of such bounds for random matrices. Often, randomized and approximation algorithms can greatly speed up the compututation of matrix properties used in algorithms. The analysis of such algorithms requires error bounds on the amount of deviation, and matrix concentration inequalities can be used to give guarantees of accuracy. Randomized approximation of data is one such application. Randomized dimension reduction is one such setting [Tropp15]. Recent work in optimization theory has also made use of matrix concentration, in particular to give approximation guarantees to an estimate of the Hessian matrix, used in a second-order gradient method [Agarwal16].

First, we will summarize some key definitions and properties from matrix concentration. Then, we will present the Matrix Bernstein Inequality, which is useful in cases where a random matrix Z with $\mathbf{E}[Z] = 0$ can be expressed as a sum of independent random matrices S_k with bounded spectral norm $\|S_k\| \leq B$.

$$Z = \sum_{k=1}^{n} S_k \tag{1}$$

We will then present an example of its usage in machine learning and optimization, outlining the basic pattern of its usage.

Throughout this work, we will liberally use the excellent monograph of [Tropp15]. We primarily draw from Chapters $1, 2, 3$,and 6.

2 Background on Matrix Concentration

First, we provide some definitions from matrix analysis and probability.

Definition 2.1. Matrix Exponential.
We can define $f(x) = e^x$. Then, letting $A = Q\Sigma Q^*$,

$$e^A = Qf(\Sigma)Q^* = f(A) \tag{2}$$

where f is applied element-wise to diagonal matrix Σ. Note that the this matrix is positive definite if A is Hermitian. Note also that if λ is an eigenvalue of A, then e^λ is an eigenvalue of e^A. We define the matrix logarithm as the inverse operator.

Definition 2.2. Matrix Variance Statistic.
Suppose Y is a random Hermitian matrix with zero mean. Then, the matrix variance statistic $v(Y)$ is given by

$$v(Y) = \|\mathrm{Var}(Y)\| = \|\mathbf{E}[(Y - \mathbf{E}[Y])^2]\| = \|\mathbf{E}[Y^2]\| \tag{3}$$

In the case where Y is not Hermitian and can be rectangular, we define

$$v(Y) = \max\{\|\mathbf{E}[YY^*]\|, \|\mathbf{E}[Y^*Y]\|\} \tag{4}$$

Now, we would like to generalize Chernoff bounds from the scalar setting to the matrix setting. To do so, we define the **matrix cumulant generating function**:

Definition 2.3. Matrix Cumulant Generating Function.
Let X be a random Hermitian matrix with zero mean and let $\theta \in \mathbb{R}$. Then,

$$\psi_X(\theta) = \log \mathbf{E}[e^{\theta X}] \tag{5}$$

We would like to be able to get Chernoff bounds for the extreme eigenvalues of X. It turns out we can get analagous tail bounds for the eigenvalues:

Lemma 2.4. *Tail Bounds for Eigenvalues.*
For all $t \in \mathbb{R}$, Y a random Hermitian matrix with zero mean, eigenvalues λ,

$$\begin{aligned} \boldsymbol{P}\{\lambda_{max} \geq t\} &\leq \inf_{\theta > 0} e^{-\theta t} \boldsymbol{E}[\mathrm{Tr}(e^{\theta Y})] \\ \boldsymbol{P}\{\lambda_{min} \leq t\} &\leq \inf_{\theta < 0} e^{-\theta t} \boldsymbol{E}[\mathrm{Tr}(e^{\theta Y})] \end{aligned} \tag{6}$$

Proof. The proof closely follows the standard method. We replace the terms in the probability with their exponentiated versions and apply Markov's inequality, just like in usual Chernoff bounds. Then, by the definition of matrix exponential and singular value decomposition, we have

$$e^{\lambda_{max}(\theta Y)} = \lambda_{max}(e^{\theta Y}) \tag{7}$$

Since we have that the matrix exponential of a Hermitian matrix is positive definite, we know the spectrum is non-zero and thus the top eigenvalue is bounded by the trace of $e^{\theta Y}$, which is the sum of the eigenvalues in the Hermitian case. The proof of the λ_{min} case is analagous. This result is valid for all positive θ, so we minimize accordingly.Thus, we conclude the desired result. □

Unlike the scalar setting, we can also get bounds on the expected values of the maximum and minimum eigenvalues:

Lemma 2.5. *Expectation Bounds for Eigenvalues.*
Let Y be a random Hermitian matrix with zero mean, eigenvalues λ,

$$\begin{aligned} \boldsymbol{E}[\lambda_{max}] &\leq \inf_{\theta > 0} \frac{1}{\theta} \log \boldsymbol{E}[\mathrm{Tr}(e^{\theta Y})] \\ \boldsymbol{E}[\lambda_{min}] &\geq \sup_{\theta < 0} \frac{1}{\theta} \log \boldsymbol{E}[\mathrm{Tr}(e^{\theta Y})] \end{aligned} \tag{8}$$

Proof. First, we write

$$\begin{aligned}\mathbf{E}[\lambda_{max}(Y)] &= \frac{\theta}{\theta}\mathbf{E}[\log e^{\lambda_{max}(Y)}] \\ &= \frac{1}{\theta}\mathbf{E}[\log e^{\theta\lambda_{max}(Y)}] \\ &= \frac{1}{\theta}\mathbf{E}[\log e^{\lambda_{max}(\theta Y)}]\end{aligned} \tag{9}$$

Then, as in the tail bounds, Jensen's inequality, the definition of matrix exponential, and the positive definiteness of a Hermitian matrix exponential give the desired result. □

Our goal is to now take the case where we have a random matrix expressed as a sum of independent random matrices $Z = \sum_{k=1}^{n} S_k$, and specialize the previous inequalities so that the expectation is over individual summands. This approach will allow us to exploit specific properties of the independent sum terms S_k. Tropp notes that this approach leads to matrix concentration inequalities sharp for specific examples, but it is perhaps the case that one can do better in other settings [Tropp15].

In the scalar setting, there are no problems directly generalizing the individual random variable to a sum of independent random variables since the moment generating functions "factorize" without any inequalities. We mean this in the sense that for scalar random variables $(X_1, \cdots, X_n)$,

$$M_{\sum_{k=1}^{n} X_k}(\theta) = \sum_{k=1}^{n} M_{X_k}(\theta) \tag{10}$$

This fails to be true in the matrix case, and taking the trace of the matrix moment generating functions only yields a subadditive inequality for the case where there are two summands. However, the matrix cumulant generating function is subadditive:

Lemma 2.6. *Subadditivity of Matrix Cumulant Generating Function.*
Consider a finite sequence $\{Y_k\}_{k\in[n]}$ of independent random Hermitian matrices of the same dimension. Then

$$\boldsymbol{E}[\mathrm{Tr} e^{\sum_k \theta Y_k}] \leq \mathrm{Tr} e^{\sum_k \log \boldsymbol{E}[e^{\theta Y_k}]} \tag{11}$$

Proof. This proof follows from repeated application of the tower property of conditional probability in addition to a theorem from Lieb:

Theorem 2.7. *Lieb's Concavity Theorem.*
Let H be a deterministic Hermitian matrix. Then the map

$$A \to \mathrm{Tr} e^{H+\log A} \tag{12}$$

is concave on the positive definite cone.

Jensen's inequality gives us from this theorem that

$$\mathbf{E}[\mathrm{Tr} e^{H+X}] \leq \mathrm{Tr} e^{H+\log \mathbf{E}[e^X]} \tag{13}$$

Then we just expand out the lefthand term in the lemma statement by noting that

$$\mathbf{E}[\mathrm{Tr} e^{\sum_k \theta Y_k}] = \mathbf{E}[\mathbf{E}_n[\mathrm{Tr} e^{\sum_{k=1}^{n-1} \theta Y_k + \theta Y_n}]] \tag{14}$$

Applying the previous inequality gives

$$\mathbf{E}[\mathrm{Tr}e^{\sum_k \theta Y_k}] \leq \mathbf{E}[\mathrm{Tr}e^{\sum_{k=1}^{n-1} \theta Y_k + \theta \log \mathbf{E}[e^{Y_n}]}] \tag{15}$$

where we have chosen $H = \sum_{k=1}^{n-1} \theta Y_k$. This selection is valid since we are only taking expectation over Y_n, and crucially, the rest of the terms are **independent** from Y_n. We iterate this process for all n terms, each time converting a Y_k to its matrix cumulant. At step m, H_m is chosen to be all terms (both cumulants and Ys) that never had index m. These choices of H_m are valid for the same reason as before. After applying these steps for all n of the Y_ks, we complete the proof. □

Now we get our matrix concentration result: By substituting in our subadditivity bound into the Chernoff-like bounds we derived previously, we get concentration bounds which control the tightness of the bound via the information of the independent summand matrices.

Theorem 2.8. *Bounds for Sums of Independent Random Matrices.*
Consider $\{Y_k\}_k$, a finite sequence of independent, random, Hermitian matrices. Let $Z = \sum_{k=1}^n Y_k$. Then for eigenvalues λ, $\theta \in \mathbb{R}$, for all $t \in \mathbb{R}$,

$$\begin{aligned} \boldsymbol{P}\{\lambda_{max}(Z) \geq t\} &\leq \inf_{\theta > 0} e^{-\theta t} \mathrm{Tr} e^{\sum_k \log \boldsymbol{E}[e^{\theta Y_k}]} \\ \boldsymbol{P}\{\lambda_{min}(Z) \leq t\} &\leq \inf_{\theta < 0} e^{-\theta t} \mathrm{Tr} e^{\sum_k \log \boldsymbol{E}[e^{\theta Y_k}]} \end{aligned} \tag{16}$$

and

$$\begin{aligned} \boldsymbol{E}[\lambda_{max}(Z)] &\leq \inf_{\theta > 0} \frac{1}{\theta} \log \mathrm{Tr} e^{\sum_k \log \boldsymbol{E}[e^{\theta Y_k}]} \\ \boldsymbol{E}[\lambda_{min}(Z)] &\geq \sup_{\theta < 0} \frac{1}{\theta} \log \mathrm{Tr} e^{\sum_k \log \boldsymbol{E}[e^{\theta Y_k}]} \end{aligned} \tag{17}$$

Proof. Direct from applying subadditivity of matrix cumulants and the previous Chernoff-like bounds. □

3 Matrix Bernstein Inequality

We now introduce a case where we are in a position to exploit the bounds for sums of independent random matrices derived in the previous section with specific knowledge about the independent summand matrices. The **Matrix Bernstein** inequality is the case where we bound the spectral norm of each of the independent summands. We write

$$Z = \sum_{k=1}^{n} S_k,\ \mathbf{E}[S_k] = 0,\ \|S_k\| \leq B \tag{18}$$

Our inequality will control the expectatation and tail behavior of Z in terms of the matrix variance statistic $v(Z)$ as well as the uniform bounds on the spectral norm of the summands, B. We will prove the Hermitian case.

Theorem 3.1. *Matrix Bernstein Inequality.*
Consider a finite sequence $\{S_k\}_k$ of independent random Hermitian matrices of dimension d. Assume that $\boldsymbol{E}[S_k] = 0$ and $\|S_k\| \leq B$ for all k. Let $Z = \sum_{k=1}^n S_k$. Then we can bound the expectation

$$\boldsymbol{E}[\|Z\|] \leq \sqrt{2v(Z)\log(d)} + \frac{B}{3}\log(d) \tag{19}$$

and get a tail bound on the spectral norm for all $t \geq 0$*,*

$$\boldsymbol{P}\{\|Z\| \geq t\} \leq d e^{\frac{-t^2/2}{v(Z)+Bt/3}} \tag{20}$$

These results naturally extend to uncentered random matrices.

Proof. First we introduce a lemma giving a bound on the cumulant generating function after making the Bernstein assumption:

Lemma 3.2. *Matrix Cumulant Generating Function Bound.*
Let X *be a random Hermitian matrix satisfying* $\boldsymbol{E}[X] = 0$ *and* $\lambda_{max}(X) \leq B$*. Then for all* $0 < \theta < 3/B$*,*

$$\log \boldsymbol{E}[e^{\theta X}] \preceq g(\theta)\boldsymbol{E}[X^2] \tag{21}$$

Let $g(\theta) = \frac{\theta^2/2}{1-\theta B/3}$*.*

Proof. We can write $e^{\theta X} = I + \theta X + Xf(X)X$, where $f(x) = \frac{e^{\theta x} - \theta x - 1}{x^2}$ for $x \neq 0$, and $f(0) = \frac{\theta^2}{2}$. f is increasing since it has a positive derivative, and thus $f(x) \leq f(B)$ for $x \leq B$. Using the definition of matrix exponential and the fact that X is Hermitian with positive eigenvalues and that B is an upper bound on eigenvalues of X, we have that $f(X) \preceq f(B) * I$. Then the semidefinite relation is preserved under conjugation, so

$$e^{\theta X} \preceq I + \theta X + f(B)X^2 \tag{22}$$

Let's fill in $f(B)$ with something concrete. Writing out the Taylor series for $f(B)$, we get

$$\begin{aligned} f(B) = \frac{e^{\theta B} - \theta B - 1}{B^2} &= \frac{1}{B^2}\sum_{q=2}^{\infty}\frac{(\theta B)^q}{q!} \\ &\leq \frac{\theta^2}{2}\sum_{q=2}^{\infty}\frac{(\theta B)^{q-2}}{3^{q-2}} \\ &= \frac{\theta^2}{2}\frac{1}{1-\theta B/3} \end{aligned} \tag{23}$$

since $q! \geq 2 * 3^{q-2}$ for $q \geq 2$ and by infinite geometric series identity. Therefore, we complete a bound on the matrix moment generating function, since taking expectations preserves semidefinite ordering (and recall we assume $\mathbf{E}[X] = 0$):

$$\begin{aligned} e^{\theta X} &\preceq I + \theta X + \frac{\theta^2/2}{1-\theta B/3}X^2 \\ \mathbf{E}[e^{\theta X}] &\preceq I + \frac{\theta^2/2}{1-\theta B/3}\mathbf{E}[X^2] \\ &\preceq \exp\left(\frac{\theta^2/2}{1-\theta B/3}\mathbf{E}[X^2]\right) \end{aligned} \tag{24}$$

where we apply the definition of matrix exponential to the inequality $1 + x \leq e^x$ for all $x \in \mathbb{R}$ in order to lift this inequality to a semidefinite ordering on matrices. Taking logarithms yields the desired bound. □

Now, we can use this lemma inside the Sum of Independent Random Matrices Bound from the previous section. We have by combining these two lemmas that

$$\begin{aligned}\mathbf{E}[\lambda_{max}(Z)] &\leq \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{Tr} e^{\sum_k \log \mathbf{E}[e^{\theta S_k}]} \\ &\leq \inf_{0<\theta<3/L} \frac{1}{\theta} \log \operatorname{Tr} e^{g(\theta)\sum_k \mathbf{E}[S_k^2]} \\ &= \inf_{0<\theta<3/L} \frac{1}{\theta} \log \operatorname{Tr} e^{g(\theta)\sum_k \mathbf{E}[Z^2]}\end{aligned} \tag{25}$$

where the trace exponential monotonicity and the additivity of variance under independent variables are used.

Then, we can bound the trace by d times the maximum eigenvalue and apply the definition of matrix exponential:

$$\begin{aligned}\mathbf{E}[\lambda_{max}(Z)] &\leq \inf_{0<\theta<3/L} \frac{1}{\theta} \log \left[d\lambda_{max}\left(e^{g(\theta)\mathbf{E}[Z^2]}\right)\right] \\ &= \inf_{0<\theta<3/L} \frac{1}{\theta} \log \left[d\left(e^{g(\theta)\lambda_{max}(\mathbf{E}[Z^2])}\right)\right] \\ &\leq \inf_{0<\theta<3/L} \frac{1}{\theta} \log \left[d\left(e^{g(\theta)v(Z)}\right)\right] \\ &= \inf_{0<\theta<3/L} \left[\frac{\log d}{\theta} + \frac{\theta^2/2}{1-\theta B/3} \cdot v(Z)\right]\end{aligned} \tag{26}$$

Choosing an appropriate θ yields the result.

The method to get the tail bound is nearly identical, except we use the tail bound inequality instead of the expectation inequality, and choose $\theta = t/(v(Z) + Bt/3)$ which yields the desired bound.

□

4 Application to Second-Order Optimization

We now introduce a setting where matrix concentration can be used to give bounds proving the efficient of a stochastic second-order optimization method which can run in linear time [Agarwal16].

4.1 Setup

Our setting is the following optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \left\{\frac{1}{m}\sum_{k=1}^{m} f_k(x) + R(x)\right\} \tag{27}$$

where each $f_k(x)$ is a convex function and $R(x)$ is a convex regularizer. One method of solving this problem is via Newton's method, which is a second-order gradient approach to continuous optimization. The Newton update is given by

$$x_{t+1} = x_t - \nabla^{-2} f(x_t)\nabla f(x_t) \tag{28}$$

This update is computationally challenging to perform, since it requires that we find the inverse Hessian matrix. In this paper, the authors instead derive a stochastic approximation scheme for the

Hessian, and calculate matrix-vector products directly instead of performing matrix multiplication between the Hessian and $f(x_t)$.

In the analysis, we assume that each f_k is α-strongly convex and β-smooth. For simplicity in this summary, we define the condition number to be $\kappa = \frac{\beta}{\alpha}$, which corresponds to

$$\kappa = \max_x \frac{\lambda_{max}(\nabla^2 f(x))}{\lambda_{min}(\nabla^2 f(x))} \tag{29}$$

We further assume that $\frac{I}{\kappa} \preceq \nabla^2 f_k \preceq I$, by definition of strong convexity and condition numbers.

4.2 Hessian Inverse Sampling Estimator

To estimate the Hessian inverse, we define the following estimator. We use a truncated Taylor approximation of the matrix inverse, up to j:

$$\begin{aligned} A_j^{-1} &= \sum_{i=0}^{j} (I - A)^i \\ A_j^{-1} &= I + (I - A)A_{j-1}^{-1} \end{aligned} \tag{30}$$

We simply take $A_0 = I$, and view $A_j^{-1} = \tilde{\nabla}^{-2} f_j$. We use two parameters to fully define our approximation: First, we take S_1 different sampled independent estimates of the the Hessian inverse from the m functions which f is comprised of ($f = \sum_{k=1}^{m} f_k$), expanding each estimate to S_2 terms of the Taylor expansion. Then, we average the S_1 estimates to get our final estimator. Thus, we see that our inverse Hessian matrix can be expressed as a sum of independent random matrices. Moreover, due to our condition number assumptions, each random matrix has bounded spectral norm. Thus, this is the correct setting for a matrix Bernstein inequality. We write the sum here:

$$\tilde{\nabla}^{-2} f = \sum_{k=1}^{S_1} \frac{1}{S_1} \tilde{\nabla}^{-2} f_{S_2}^{k} \tag{31}$$

where each summand is independently drawn.

Our goal is to bound the deviations of the inverse Hessian estimator from the average. Thus, we must bound the spectral norm of each $\tilde{\nabla}^{-2} f_{S_2}$ before we apply matrix Bernstein.

Lemma 4.1. *Uniform Bound on Spectral Norm.*
For any $f_{S_2} \in \{f_k\}_{k \in [m]}$ we have

$$\| \frac{1}{S_1} \tilde{\nabla}^{-2} f_{S_2} \| \leq \frac{\kappa}{S_1} \tag{32}$$

Proof. First note that we have by definition that

$$\tilde{\nabla}^{-2} f_{S_2}(x_t) = \sum_{i=0}^{S_2} (I - \nabla^2 f(x_t))^i \tag{33}$$

Since $\frac{I}{\kappa} \preceq \nabla^2$, we have $\|I - \nabla^2 f(x_t)\| \leq 1 - \frac{1}{\kappa}$ by the fact that semidefinite orderings are preserved under the norms. Therefore,

$$\begin{aligned}\|\frac{1}{S_1}\tilde{\nabla}^{-2}f_{S_2}\| &\leq \frac{1}{S_1}\sum_{i=0}^{S_2}(1-\frac{1}{\kappa})^i \\ &\leq \frac{1}{S_1}\sum_{i=0}^{\infty}(1-\frac{1}{\kappa})^i \\ &= \frac{1}{S_1}\frac{1}{1-(1-\frac{1}{\kappa})} \\ &= \frac{\kappa}{S_1}\end{aligned} \tag{34}$$

□

Now that we have the uniform spectral summand bound, we need to bound $v(\tilde{\nabla}^{-2}f)$, the matrix variance statistic for the whole estimate. Since the terms of the sum are independent and bounded uniformly, we can use the additivity of variance to get that

$$\begin{aligned}v(\tilde{\nabla}^{-2}f) &\leq \sum_{k=1}^{S_1}\left\|\frac{1}{S_1^2}\tilde{\nabla}^{-2}f_{S_2}^{k}\right\|^2 \\ &\leq \frac{1}{S_1}\|\tilde{\nabla}^{-2}f_{S_2}\|^2 = \frac{\kappa^2}{S_1}\end{aligned} \tag{35}$$

Finally, note that $\left\|\tilde{\nabla}^{-2}f\right\| \leq 1$ since $\tilde{\nabla}^{-2}f \preceq I$.

We are finally ready to apply matrix Bernstein. We have to mean center in order to apply the theorem; for this, we pay a factor of 2 in the probability bound. From the theorem, we have

$$\mathbf{P}\{\|\tilde{\nabla}^{-2}f - \mathbf{E}[\tilde{\nabla}^{-2}f]\| \geq t\} \leq 2d\cdot\exp\left\{\frac{-t^2/2}{2v(\tilde{\nabla}^{-2}f - \mathbf{E}[\tilde{\nabla}^{-2}f]) + Bt/3}\right\} \tag{36}$$

where $B = 2\frac{\kappa}{S_1}$. Then, we note that this probability is only sensical if $t \leq \|\tilde{\nabla}^{-2}f - \mathbf{E}[\tilde{\nabla}^{-2}f]\| \leq 1$. Thus, we can set the t in the denominator to 1 to simplify the expression. Overall, we get

$$\begin{aligned}\mathbf{P}\{\|\tilde{\nabla}^{-2}f - \mathbf{E}[\tilde{\nabla}^{-2}f]\| \geq t\} &\leq 2d\cdot\exp\left\{\frac{-t^2/2}{2\frac{\kappa^2}{S_1} + 2\frac{\kappa/3}{S_1}}\right\} \\ &= 2d\cdot\exp\left\{\frac{-t^2S_1}{4\kappa^2 + 4\kappa/3}\right\} \\ &\leq 2d\cdot\exp\left\{\frac{-t^2S_1}{\mathcal{O}(1)\cdot\kappa^2}\right\}\end{aligned} \tag{37}$$

noting that $\kappa \geq 1$ and thus that $\kappa \leq \kappa^2 \implies -\frac{1}{\kappa} \leq -\frac{1}{\kappa^2}$.

This mean deviation bound on the inverse Hessian estimator can then be applied to give a δ-bound on the spectral norm of the difference between the estimate and the true inverse Hessian, which is then used in the paper to prove runtime and sample complexity guarantees.

References

[Tropp15] Tropp, Joel. Matrix Concentration Inequalities (2015).

[Agarwal16] Agarwal, Naman, Bullins, Brian, and Hazan, Elad. Second Order Stochastic Optimization for Machine Learning in Linear Time (2016).