Advanced Graph Algorithms and Optimization Spring 2020

More Gaussian Elimination, Introduction to Random Matrix Concentration

Rasmus Kyng *Lecture 7 — Wednesday, April 1*

1 Gaussian Elimination Recap and Structure Claim

Last time, we studied a convex function minimization problem, and saw how solve it by coordinatewise minimizaiton, which I claimed is really Gaussian elimination in disguise. Let's recap part of what we saw.

Consider a Laplacian $\boldsymbol{L}$ of a connected graph $G = (V, E, \boldsymbol{w})$, where $\boldsymbol{w} \in \mathbb{R}^E$ is a vector of positive edge weights. Let $\boldsymbol{W} \in \mathbb{R}^{E \times E}$ be the diagonal matrix with the edge weights on the diagonal, i.e. $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w})$ and $\boldsymbol{L} = \boldsymbol{B}\,\boldsymbol{W}\,\boldsymbol{B}^\top$. Let $\boldsymbol{d} \in \mathbb{R}^V$ be a demand vector s.t. $\boldsymbol{d} \perp \mathbf{1}$.

We defined an energy

$$\mathcal{E}(\boldsymbol{x}) = -\boldsymbol{d}^\top \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x}$$

Note that this function is convex and is minimized at $\boldsymbol{x}$ s.t. $\boldsymbol{L}\boldsymbol{x} = \boldsymbol{d}$.

To understand how to minimize over the first variable, we introduce some notation for the first row and column of the Laplacian:

$$\boldsymbol{L} = \begin{pmatrix} W & -\boldsymbol{a}^\top \\ -\boldsymbol{a} & \mathrm{diag}(\boldsymbol{a}) + \boldsymbol{L}_{-1} \end{pmatrix} \tag{1}$$

Note that W is the weighted degree of vertex 1, and that

$$\begin{pmatrix} W & -\boldsymbol{a}^\top \\ -\boldsymbol{a} & \mathrm{diag}(\boldsymbol{a}) \end{pmatrix} \tag{2}$$

is the Laplacian of the subgraph of G containing only the edges incident on vertex 1, while $\boldsymbol{L}_{-1}$ is the Laplacian of the subgraph of G containing all edges *not* incident on vertex 1.

Let us also write $\boldsymbol{d} = \begin{pmatrix} b \\ \boldsymbol{c} \end{pmatrix}$ where $y \in \mathbb{R}$ and $\boldsymbol{c} \in \mathbb{R}^{V \setminus \{1\}}$.

Now,

$$\mathcal{E}(\boldsymbol{x}) = -\boldsymbol{d}^\top \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} = -\begin{pmatrix} b \\ \boldsymbol{c} \end{pmatrix}^\top \begin{pmatrix} y \\ \boldsymbol{z} \end{pmatrix} + \frac{1}{2}\begin{pmatrix} y \\ \boldsymbol{z} \end{pmatrix}^\top \begin{pmatrix} W & -\boldsymbol{a}^\top \\ -\boldsymbol{a} & \mathrm{diag}(\boldsymbol{a}) + \boldsymbol{L}_{-1} \end{pmatrix} \begin{pmatrix} y \\ \boldsymbol{z} \end{pmatrix}$$

Now, to minimize over y, we set $\frac{\partial}{\partial y}\mathcal{E}(\boldsymbol{x}) = 0$ and get

$$-b + yW - \boldsymbol{a}^\top \boldsymbol{z} = 0.$$

Solving for y, we get that the minimizing y is

$$y = \frac{1}{W}(b + \boldsymbol{a}^\top \boldsymbol{z}). \tag{3}$$

By substituting in this value of y, we found

$$\min_y \mathcal{E}\begin{pmatrix} y \\ \boldsymbol{z} \end{pmatrix} = -\left(\boldsymbol{c} + b\frac{1}{W}\boldsymbol{a}\right)^\top \boldsymbol{z} - \frac{b^2}{2W} + \frac{1}{2}\boldsymbol{z}^\top \boldsymbol{S}\boldsymbol{z}.$$

where $\boldsymbol{S} = \mathrm{diag}(\boldsymbol{a}) - \frac{1}{W}\boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{L}_{-1}$. We let $\boldsymbol{d}' = \boldsymbol{c} + b\frac{1}{W}\boldsymbol{a}$, and $\mathcal{E}'(\boldsymbol{z}) = -\boldsymbol{d}^\top \boldsymbol{z} + \frac{1}{2}\boldsymbol{z}^\top \boldsymbol{S}\boldsymbol{z}$, and noted that

$$\arg\min_{\boldsymbol{z}} \min_y \mathcal{E}\begin{pmatrix} y \\ \boldsymbol{z} \end{pmatrix} = \arg\min_{\boldsymbol{z}} \mathcal{E}'(\boldsymbol{z}),$$

and we ended the lecture by stating with the following claim.

Claim 1.1.

1. $\boldsymbol{d}' \perp \mathbf{1}$ *when* $\boldsymbol{d} \perp \mathbf{1}$.
2. $\boldsymbol{S} = \mathrm{diag}(\boldsymbol{a}) - \frac{1}{W}\boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{L}_{-1}$ *is a Laplacian of a connected graph on the vertex set* $V \setminus \{1\}$.

Now we're ready to prove it.

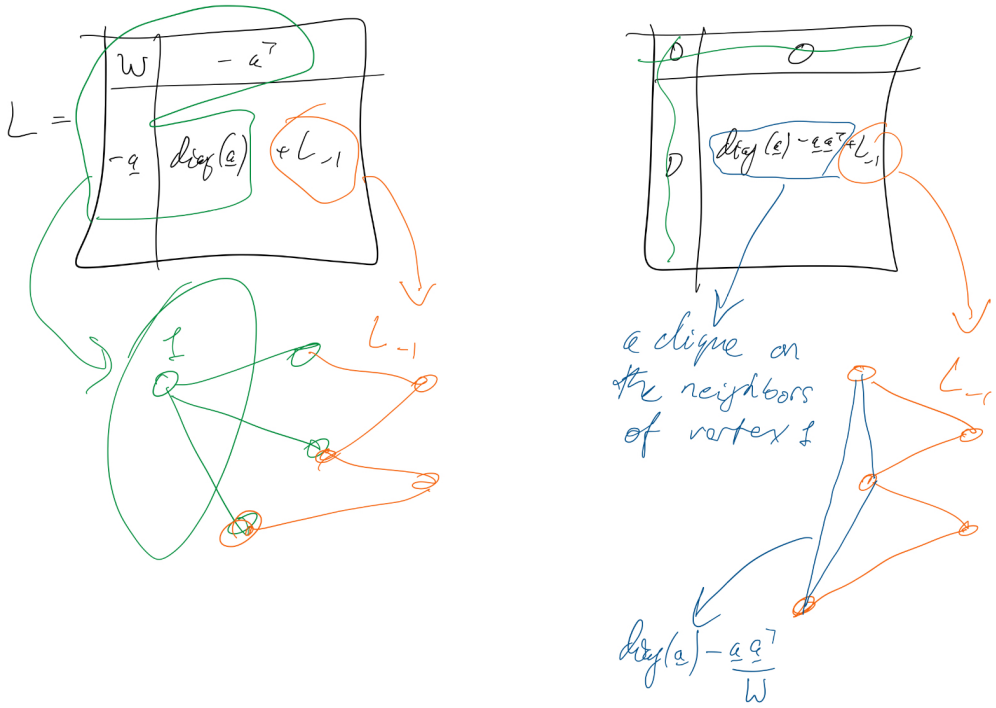
Proof. To establish the first part, we note that $\mathbf{1}^\top \boldsymbol{d}' = \mathbf{1}^\top \boldsymbol{c} + b\frac{\mathbf{1}^\top \boldsymbol{a}}{W} = \mathbf{1}^\top \boldsymbol{c} + b = \mathbf{1}^\top \boldsymbol{d} = 0$. To establish the second part, we notice that $\boldsymbol{L}_{-1}$ is a graph Laplacian by definition. Since the sum of two graph Laplacians is another graph Laplacian, it now suffices to show that $\boldsymbol{S}$ is a graph Laplacian.

Claim 1.2. *A matrix* $\boldsymbol{M}$ *is a graph Laplacian if and only it satisfies the following conditions:*

- $\boldsymbol{M}^\top = \boldsymbol{M}$.
- *The diagonal entries of* $\boldsymbol{M}$ *are non-negative, and the off-diagonal entries of* $\boldsymbol{M}$ *are non-positive.*
- $\boldsymbol{M}\mathbf{1} = \mathbf{0}$.

Let's see that Claim 1.2 is true. Firstly, when the conditions hold we can write $\boldsymbol{M} = \boldsymbol{D} - \boldsymbol{A}$ where $\boldsymbol{D}$ is diagonal and non-negative, and $\boldsymbol{A}$ is non-negative, symmetric, and zero on the diagonal, and from the last condition $\boldsymbol{D}(i,i) = \sum_{j \neq i} \boldsymbol{A}(i,j)$. Thus we can view $\boldsymbol{A}$ as a graph adjacency matrix and $\boldsymbol{D}$ as the corresponding diagonal matrix of weighted degrees. Secondly, it is easy to check that the conditions hold for any graph Laplacian, so the conditions indeed hold if and only if. Now we have to check that the claim applies to $\boldsymbol{S}$. We leave this as an exercise for the reader.

Finally, we want to argue that the graph corresponding to $\boldsymbol{S}$ is connected. Consider any $i, j \in V \setminus \{1\}$. Since G, the graph of $\boldsymbol{L}$, is connected, there exists a simple path in G connecting i and j. If this path does not use vertex 1, it is a path in the graph of $\boldsymbol{L}_{-1}$ and hence in the graph of $\boldsymbol{S}$. If the path does use vertex 1, it must do so by reaching the vertex on some edge $(v, 1)$ and leaving on a different edge $(1, u)$. Replace this pair of edges with edge (u, v), which appears in the graph of $\boldsymbol{S}$ because $\boldsymbol{S}(u, v) < 0$. Now we have a path in the graph of $\boldsymbol{S}$. □



2 An Additive View of Gaussian Elimination

Cholesky decomposition basics. Again we consider a graph Laplacian $\boldsymbol{L} \in \mathbb{R}^{n \times n}$ of a conected graph $G = (V, E, \boldsymbol{w})$, where as usual $|V| = n$ and $|E| = m$.

In this Section, we'll study how to decompose a graph Laplacian as $\boldsymbol{L} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{L}}^\top$, where $\boldsymbol{\mathcal{L}} \in \mathbb{R}^{n \times n}$ is a lower triangular matrix, i.e. $\boldsymbol{\mathcal{L}}(i, j) = 0$ for $i < j$. Such a factorization is called a Cholesky decomposition. It is essentially the result of Gaussian elimination with slight twist to ensure the matrices maintained at intermediate steps of the algorithm remain symmetric.

We use $\text{nnz}(\boldsymbol{A})$ to denote the number of non-zero entries of matrix $\boldsymbol{A}$.

Lemma 2.1. *Given an* invertible *square lower triangular matrix $\boldsymbol{\mathcal{L}}$, we can solve the linear equation $\boldsymbol{\mathcal{L}}\boldsymbol{y} = \boldsymbol{b}$ in time $O(\text{nnz}(\boldsymbol{\mathcal{L}}))$. Similarly, given an upper triangular matrix $\boldsymbol{\mathcal{U}}$, we can solve linear equations $\boldsymbol{\mathcal{U}}\boldsymbol{z} = \boldsymbol{c}$ in time $O(\text{nnz}(\boldsymbol{\mathcal{U}}))$.*

We omit the proof, which is a straight-forward exercise. The algorithms for solving linear equations in upper and lower triangular matrices are known as forward and back substitution respectively.

Remark 2.2. Strictly speaking, the lemma requires us to have access an adjacency list representation of $\boldsymbol{\mathcal{L}}$ so that we can quickly tell where the non-zero entries are.

Using forward and back substitution, if we have a decomposition of an invertible matrix $\boldsymbol{M}$ into $\boldsymbol{M} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{L}}^\top$, we can now solve linear equations in $\boldsymbol{M}$ in time $O(\text{nnz}(\boldsymbol{\mathcal{L}}))$.

Remark 2.3. We have learned about decompositions using a lower triangular matrix, and later we will see an algorithm for computing these. In fact, we can have more flexibility than that. From

an algorithmic perspective, it is sufficient that there exists a permutation matrix $\boldsymbol{P}$ s.t. $\boldsymbol{P}\boldsymbol{\mathcal{L}}\boldsymbol{P}^\top$ is lower triangular. If we know the ordering under which the matrix becomes lower triangular, we can perform substitution according to that order to solve linear equations in the matrix without having to explicitly apply a permutation to the matrix.

Dealing with pseudoinverses. But how can we solve a linear equation in $\boldsymbol{L} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{L}}^\top$, where $\boldsymbol{L}$ is not invertible? For graph Laplacians we have a simple characterization the kernel, and because of this, dealing with the lack of invertibility turns out to be fairly easy.

We can use the following lemma which you will prove in an exercise next week.

Lemma 2.4. *Consider a real symmetric matrix $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}^\top$, where $\boldsymbol{X}$ is real and invertible and $\boldsymbol{Y}$ is real symmetric. Let $\boldsymbol{\Pi}_{\boldsymbol{M}}$ denote the orthogonal projection to the image of $\boldsymbol{M}$. Then $\boldsymbol{M}^+ = \boldsymbol{\Pi}_{\boldsymbol{M}}(\boldsymbol{X}^\top)^{-1}\boldsymbol{Y}^+\boldsymbol{X}^{-1}\boldsymbol{\Pi}_{\boldsymbol{M}}$.*

The factorizations $\boldsymbol{L} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{L}}^\top$ that we produce will have the property that all diagonal entries of $\boldsymbol{\mathcal{L}}$ are strictly non-zero, except that $\boldsymbol{\mathcal{L}}(n,n) = 0$. From let us $\widehat{\boldsymbol{\mathcal{L}}}$ as the matrix whose entries agree with $\boldsymbol{\mathcal{L}}$, except that $\widehat{\boldsymbol{\mathcal{L}}}(n,n) = 1$. Let $\boldsymbol{\mathcal{D}}$ be the diagonal matrix with $\boldsymbol{\mathcal{D}}(i,i) = 1$ for $i < n$ and $\boldsymbol{\mathcal{D}}(n,n) = 0$. Then $\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{L}}^\top = \widehat{\boldsymbol{\mathcal{L}}}\boldsymbol{\mathcal{D}}\widehat{\boldsymbol{\mathcal{L}}}$, and $\widehat{\boldsymbol{\mathcal{L}}}$ is invertible, and $\boldsymbol{\mathcal{D}}^+ = \boldsymbol{\mathcal{D}}$. Finally, $\boldsymbol{\Pi}_{\boldsymbol{L}} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$, because this matrix is acts like identity on vectors orthogonal to $\boldsymbol{1}$ and ensures $\boldsymbol{\Pi}_{\boldsymbol{L}}\boldsymbol{1} = \boldsymbol{0}$, and this matrix can be applied to a vector in $O(n)$ time. Thus $\boldsymbol{L}^+ = \boldsymbol{\Pi}_{\boldsymbol{L}}(\widehat{\boldsymbol{\mathcal{L}}}^\top)^{-1}\boldsymbol{\mathcal{D}}\widehat{\boldsymbol{\mathcal{L}}}^{-1}\boldsymbol{\Pi}_{\boldsymbol{L}}$, and this matrix can be applied in time $O(\mathrm{nnz}(\boldsymbol{\mathcal{L}}))$.

An additive view of Gaussian Elimination. The following theorem describes Gaussian Elimination / Cholesky decompostion of a graph Laplacian.

Theorem 2.5 (Cholesky Decomposition on graph Laplacians)**.** *Let $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ be a graph Laplacian of a connected graph $G = (V, E, \boldsymbol{w})$, where $|V| = n$. Using Gaussian Elimination, we can compute in $O(n^3)$ time a factorization $\boldsymbol{L} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{L}}^\top$ where $\boldsymbol{\mathcal{L}}$ is lower triangular, and has positive diagonal entries except $\boldsymbol{\mathcal{L}}(n,n) = 0$.*

Proof. Let $\boldsymbol{L}^{(0)} = \boldsymbol{L}$. We will use $\boldsymbol{A}(:,i)$ to denote the the ith column of a matrix $\boldsymbol{A}$. Now, for $i = 1$ to $i = n-1$ we define

$$\boldsymbol{l}_i = \frac{1}{\sqrt{\boldsymbol{L}^{(i-1)}(i,i)}}\boldsymbol{L}^{(i-1)}(:,i) \text{ and } \boldsymbol{L}^{(i)} = \boldsymbol{L}^{(i-1)} - \boldsymbol{l}_i\boldsymbol{l}_i^\top$$

Finally, we let $\boldsymbol{l}_n = \boldsymbol{0}_{n\times 1}$. We will show later that

$$\boldsymbol{L}^{(i-1)} = \boldsymbol{0}_{n\times n}. \tag{4}$$

It follows that $\boldsymbol{L} = \sum_i \boldsymbol{l}_i\boldsymbol{l}_i^\top$, provided this procedure is well-defined, i.e. $\boldsymbol{L}^{(i-1)}(i,i) \neq 0$ for all $i < n$. We will sketch a proof of this later, while also establishing several other properties of the procedure.

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $U \subseteq [n]$, we will use $\boldsymbol{A}(U,U)$ to denote the principal submatrix of $\boldsymbol{A}$ obtained by restricting to the rows and columns with index in U, i.e. all entries $\boldsymbol{A}(i,j)$ where $i,j \in U$.

Claim 2.6. *Fix some $i < n$ Let $U = \{i+1, \ldots, n\}$. Then $\boldsymbol{L}^{(i)}(i,j) = 0$ if $i \notin U$ or $j \notin U$. And $\boldsymbol{L}^{(i)}(U,U)$ is a graph Laplacian of a connected graph on the vertex set U.*

From this claim, it follows that $\boldsymbol{L}^{(i-1)}(i,i) \neq 0$ for $i < n-1$, since a connected graph Laplacian on a graph with $|U| > 1$ vertices cannot have a zero on the diagonal, and it follows that $\boldsymbol{L}^{(n-1)}(i,i) = 0$, because the only graph we allow on one vertex is the empty graph. This shows Equation (4) holds. □

Sketch of proof of Claim 2.6. We will focus on the first elimination, as the remaining are similar. Adopting the same notation as in Equation (1), we write

$$\boldsymbol{L}^{(0)} = \boldsymbol{L} = \begin{pmatrix} W & -\boldsymbol{a}^\top \\ -\boldsymbol{a} & \mathrm{diag}(\boldsymbol{a}) + \boldsymbol{L}_{-1} \end{pmatrix}$$

and, noting that

$$\boldsymbol{l}_1 \boldsymbol{l}_1^\top = \begin{pmatrix} W & -\boldsymbol{a}^\top \\ -\boldsymbol{a} & \frac{1}{W}\boldsymbol{a}\boldsymbol{a}^\top \end{pmatrix}$$

we see that

$$\boldsymbol{L}^{(1)} = \boldsymbol{L}^{(0)} - \boldsymbol{l}_1 \boldsymbol{l}_1^\top = \begin{pmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{0} & \mathrm{diag}(\boldsymbol{a}) - \frac{1}{W}\boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{L}_{-1} \end{pmatrix}.$$

Thus the first row and column of $\boldsymbol{L}^{(1)}$ are zero claimed. It also follows by Claim 1.1 that $\boldsymbol{L}^{(1)}(\{2, \ldots, n\}, \{2, \ldots, n\})$ is the Laplacian of a connected graph. This proves Claim 2.6 for the case $i = 1$. An induction following the same pattern can be used to prove the claim for all $i < n$. □

3 Matrix Sampling and Approximation

We want to begin understanding how sums of random matrices behave, in particular, whether they exhibit a tendency to concentrate in the same way that sum of scalar random variables do under various conditions.

First, let's recall a scalar Chernoff bound, which shows that a sum of bounded, non-negative random variables tend to concentrate around their mean.

Theorem 3.1 (A Chernoff Concentration Bound)**.** *Suppose $X_1, \ldots, X_k \in \mathbb{R}$ are independent, non-negative, random variables with $X_i \leq R$ always. Let $X = \sum_i X_i$, and $\mu = \mathbb{E}[X]$, then for $0 < \epsilon \leq 1$*

$$\Pr[X \geq (1+\epsilon)\mu] \leq \exp\left(\frac{-\epsilon^2 \mu}{4R}\right) \text{ and } \Pr[X \leq (1-\epsilon)\mu] \leq \exp\left(\frac{-\epsilon^2 \mu}{4R}\right).$$

The Chernoff bound should be familiar to most of you, but you may not have seen the following very similar bound. The Bernstein bound, which we will state in terms of zero-mean variables, is much like the Chernoff bound. It also requires bounded variables. But, when the variables have small variance, the Bernstein bound is sometimes stronger.

Theorem 3.2 (A Bernstein Concentration Bound)**.** *Suppose $X_1, \ldots, X_k \in \mathbb{R}$ are independent, zero-mean, random variables with $|X_i| \leq R$ always. Let $X = \sum_i X_i$, and $\sigma^2 = Var[X] = \sum_i \mathbb{E}[X_i^2]$, then for $\epsilon > 0$*

$$\Pr[|X| \geq t] \leq 2\exp\left(\frac{-t^2}{2Rt + 4\sigma^2}\right).$$

We will now prove the Bernstein concentration bound for scalar random variables, as a warm-up to next lecture, where we will prove a version of it for matrix-valued random variables. To help us prove Bernstein's bound, first let's recall Markov's inequality. This is a very weak concentration inequality, but also very versatile, because it requires few assumptions.

Lemma 3.3 (Markov's Inequality)**.** *Suppose $\boldsymbol{X} \in \mathbb{R}$ is a non-negative random variable, with a finite expectation. Then for any $t > 0$,*

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \Pr[X \geq t]\,\mathbb{E}[X \mid X \geq t] + \Pr[X < t]\,\mathbb{E}[X \mid X < t] \\
&\geq \Pr[X \geq t]\,\mathbb{E}[X \mid X \geq t] \\
&\geq Pr[X \geq t] \cdot t.
\end{aligned}$$

We can rearrange this to get the desired statment. □

Now, we are ready to prove Bernstein's bound.

Proof of Theorem 3.2. We will focus on bounding the probability that $\Pr[X \geq t]$. The proof that $\Pr[-X \geq t]$ is small proceeds in the same way.

First we observe that

$$\begin{aligned}
\Pr[X \geq t] &= \Pr[\exp(\theta X) \geq \exp(\theta t)] && \text{for any } \theta > 0, \text{ because } x \to \exp(\theta x) \text{ is strictly increasing.} \\
&\leq \exp(-\theta t)\,\mathbb{E}[\exp(\theta X)] && \text{by Lemma 3.3 (Markov's Inequality).}
\end{aligned}$$

Now, let's require that $\theta \leq 1/R$ This will allow us to using the following bound: For all $|z| \leq 1$,

$$\exp(z) \leq 1 + z + z^2. \tag{5}$$

We omit a proof of this, but the plots in Figure 1 suggest that this upper bound holds. The reader should consider how to prove this. With this in mind, we see that

$$\begin{aligned}
\mathbb{E}[\exp(\theta X)] &= \mathbb{E}\left[\exp\left(\theta \sum_i X_i\right)\right] \\
&= \mathbb{E}[\Pi_i \exp(\theta X_i)] \\
&= \Pi_i\,\mathbb{E}[\exp(\theta X_i)] && \text{because } \mathbb{E}[YZ] = \mathbb{E}[Y]\,\mathbb{E}[Z] \text{ for independent } Y \text{ and } Z. \\
&= \Pi_i\,\mathbb{E}\left[1 + \theta X_i + (\theta X_i)^2\right] \\
&= \Pi_i(1 + \theta^2\,\mathbb{E}\left[X_i^2\right]) \\
&\leq \Pi_i \exp(\theta^2\,\mathbb{E}\left[X_i^2\right]) && \text{because } 1 + z \leq \exp(z) \text{ for all } z \in R. \\
&= \exp\left(\sum_i \theta^2\,\mathbb{E}\left[X_i^2\right]\right) = \exp(\theta^2\sigma^2).
\end{aligned}$$

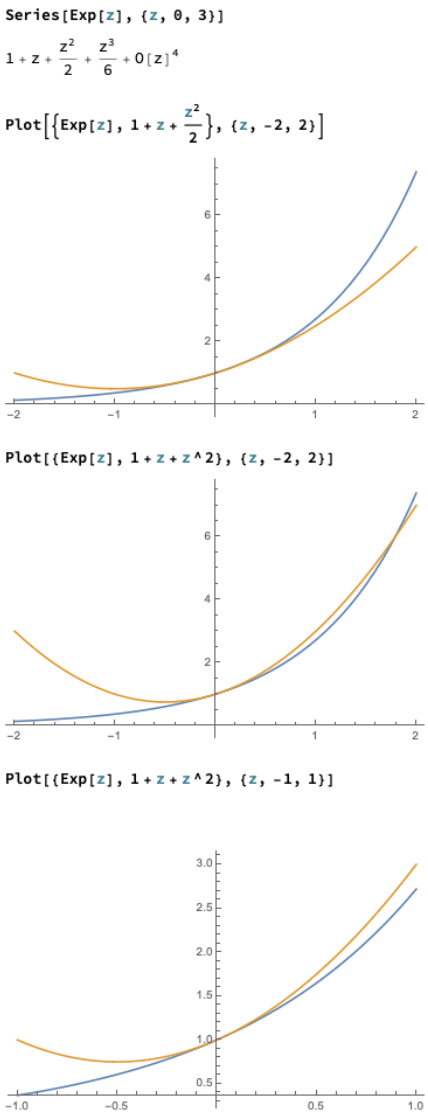


Figure 1: Plotting $\exp(z)$ compared to $1 + z + z^2$.

Thus $\Pr[X \geq t] \leq \exp(-\theta t)\,\mathbb{E}\left[\exp(\theta X)\right] \leq \exp(-\theta t + \theta^2\sigma^2)$. Now, to get the best possible bound, we'd like to minimize $-\theta t + \theta^2\sigma^2$ subject to the constraint $0 < \theta \leq 1/R$. Setting

$$\frac{\partial}{\partial \theta}\left(-\theta t + \theta^2\sigma^2\right) = -t + 2\theta\sigma^2.$$

Setting this derivative to zero gives $\theta = \frac{t}{2\sigma^2}$, and plugging that in gives

$$-\theta t + \theta^2\sigma^2 = -\frac{t^2}{4\sigma^2}$$

This choice only satisfies our constraints on θ if $\frac{t}{2\sigma^2} \leq 1/R$. Otherwise, we let $\theta = 1/R$ and note that in this case

$$-\theta t + \theta^2\sigma^2 = -\frac{t}{R} + \frac{\sigma^2}{R^2} \leq -\frac{t}{R} + \frac{t}{2R} = -\frac{t}{2R}$$

where we got the inequality from $t > 2\sigma^2/R$. Altogether, we can conclude that there always is a choice of θ s.t.

$$-\theta t + \theta^2\sigma^2 \leq -\min\left(\frac{t}{2R}, \frac{t^2}{4\sigma^2}\right) \leq -\frac{t^2}{2Rt + 4\sigma^2}.$$

In fact, with the benefit of hindsight, and a little algebra, we arrive at the same conclusion in another way: One can check that the following choice of θ is always valid and achives the same bound: $\theta = \frac{1}{2\sigma^2}\left(t - \frac{\sqrt{R}\cdot t^{3/2}}{\sqrt{2\sigma^2 + Rt}}\right)$. □

We use $\|\cdot\|$ to denote the spectral norm on matrices. Let's take a look at a version of Bernstein's bound that applies to sums of random matrices.

Theorem 3.4 (A Bernstein Matrix Concentration Bound (Tropp 2011))**.** *Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k \in \mathbb{R}^{n\times n}$ are independent, symmetric matrix-valued random variables. Assume each $\boldsymbol{X}_i$ is zero-mean, i.e. $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{0}_{n\times n}$, and that $\|\boldsymbol{X}_i\| \leq R$ always. Let $\boldsymbol{X} = \sum_i \boldsymbol{X}_i$, and $\sigma^2 = Var[\boldsymbol{X}] = \sum_i \mathbb{E}\left[\boldsymbol{X}_i^2\right]$, then for $\epsilon > 0$*

$$\Pr[\|\boldsymbol{X}\| \geq t] \leq 2n \exp\left(\frac{-t^2}{2Rt + 4\sigma^2}\right).$$

This basically says that probability of $\boldsymbol{X}$ being large in spectral norm behaves like the scalar case, except the bound is larger by a factor n, where the matrices are $n \times n$. We can get a feeling for why this might be a reasonable bound by considering the case of random diagonal matrices. Now $\|\boldsymbol{X}\| = \max_j |\boldsymbol{X}(j,j)| = \max_j |\sum_i \boldsymbol{X}_i(j,j)|$. In this case, we need to bound the largest of the n diagonal entries: We can do this by a union bound over n instances of the scalar problem – and this also turns out to be essentially tight in some cases, meaning we can't expect a better bound in general.