# Lecture 7: More on Sparsification

Instructor: Lorenzo Orecchia Scribe: Erasmo Tani

## Sparsification

We reduced the problem to the setting in which we have a collection of vectors in $n$ dimension $\mathbf{u}_i \in \mathbb{R}^n$ and :

$$\sum_i \mathbf{u}_i \mathbf{u}_i^T = I \tag{1}$$

here we want to select and weight some of these elements. So that:

$$\sum_{i=1}^{m} w_i \mathbf{u}_i \mathbf{u}_i^T \approx I$$

and the vector $\mathbf{w} \in \mathbb{R}^m$ is sparse. The strategy we will use is oblivious sampling at each iteration $t = 1, ..., T$ we sample a vector $i_t$ w.p.: $p_{i_t} = p_i$ (independent of $t$):

$$w_{i_t} = \frac{1}{p_{i_t}}$$

Note that Equation (1) implies $\sum_i ||\mathbf{u}_i||^2 = n$

**Theorem 1.** *If we choose $p_i \propto ||u_i||^2$ (importance sampling) then, after $T = O(\frac{n \log n}{\epsilon^2})$ samples:*

$$\left\| \sum_{t=1}^{T} w_{i_t} \mathbf{u}_{i_t} \mathbf{u}_{i_t}^T - I \right\| \leq \epsilon$$

*with constant probability (the norm above is the spectral norm w.r.t. $\ell_2$).*

To prove this, we will make use of the Matrix Chernoff Bound. The $O(\frac{n \log n}{\epsilon^2})$ is optimal for oblivious sampling (probability doesn't change with iteration counter), but not in general. Adaptive sampling allows for $O(\frac{n}{\epsilon^2})$. Consider the example of a dumbbell graph. The problem with oblivious sampling is that the probability of the middle edge to be sampled is roughly $\frac{1}{n}$ which gives the $O(n \log n)$ (this has interesting connections with the balls into bins process). We will now prove Theorem 1:

*Proof.*

**Claim 1.** $\mathbb{E}[w_{i_t} \mathbf{u}_{i_t} \mathbf{u}_{i_t}^T] = \sum_{i=1}^{m} p_i \frac{1}{p_i} \mathbf{u}_{i_t} \mathbf{u}_{i_t}^T = \sum \mathbf{u}_{i_t} \mathbf{u}_{i_t}^T = I$

Now, we will introduce the Laplacian transform potential function. Let $A_t = \sum_{i=1}^{t} w_{i_s} u_{i_s} u_{i_s}^T$. Then:

$$\lambda_{\max}(A_t) \approx \frac{1}{\eta} \log Tr(e^{\eta A_t}) = \frac{1}{\eta} \log(\sum e^{\eta \lambda_i (A_t)})$$

we define the following potential function:

$$\phi_t = \mathbb{E}[Tre^{\eta A_t}].$$

How does the potential change?

$$\mathbb{E}[\phi_{(t+1)}] = \mathbb{E}[Tre^{\eta A_{t+1}}] = \mathbb{E}[Tre^{\eta(A_t + A_{t+1})}]$$

Where the expecttion is conditional on al randomness up to time $t$. From here we can apply the Golden-Thomson inequality:

**Theorem 2** (Golden-Thomson Inequality)**.** *For symmetric matrices $X$ and $Y$:*

$$Tr(e^{x+y}) \leq Tr(e^x e^y).$$

We then get:

$$\mathbb{E}[\phi_{(t+1)}] = \mathbb{E}[Tre^{\eta A_{t+1}}] = \mathbb{E}[Tre^{\eta(A_t + A_{t+1})}] \leq Tr(e^{\eta A_t} \cdot \mathbb{E}_{t+1}[e^{\eta \Delta_{t+1}}]) = e^{\eta A_t} \bullet \mathbb{E}_{t+1}[e^{\eta \Delta_{t+1}}]$$

Where we define:

$$A \bullet B := Tr(A^T B)$$

We then have:

$$e^{\eta \Delta} \preceq I + (e^{\eta ||\Delta||} - 1) \frac{\Delta}{||\Delta||}$$

Note that we have:

$$0 \leq \Delta \leq ||\Delta|| \vec{1}$$

$$e^{\eta x} \leq 1 + (e^{\eta ||\Delta||} - 1) \frac{X}{||\Delta||}$$

(Note: We also have that if $A \succeq 0$, $B \preceq C$ then:

$$A \bullet B \leq A \bullet C$$

). We then have:

$$\mathbb{E}[\phi_{(t+1)}] \leq Tr(e^{\eta A_t}) + e^{\eta A_t} \bullet \frac{\Delta}{||\Delta||} \cdot (e^{\eta ||\Delta||} - 1)$$

and hence:

$$\frac{\mathbb{E}[\phi_{t+1}]}{\phi_{t+1}} = 1 + \frac{e^{\eta A_t}}{Tr(e^{\eta A_t})} \bullet \mathbb{E}\left[\frac{\Delta}{||\Delta||}(e^{\eta ||\Delta||} - 1)\right]$$

but also:

$$||\Delta_{t+1}|| = \left\| \frac{n}{||u||^2} u u^T \right\| = \frac{n}{||u||^2} ||u||^2 n$$

so that:

$$\frac{\mathbb{E}[\phi_{t+1}]}{\phi_{t+1}} \leq 1 + \left( \frac{e^{\eta n} - 1}{n} \right)$$

$$\frac{\mathbb{E}[\phi_T]}{\phi_0} \leq \left(1 + \frac{e^{\eta n} - 1}{n}\right)^T \phi_0 \leq \left(1 + \frac{e^{\eta n} - 1}{n}\right)^T n.$$

Now:

$$\begin{aligned}
\Pr\left[\lambda_{\max}(A_T) > T(1+\varepsilon)\right] &\leq \Pr\left[\phi_T > e^{\eta T(1+\varepsilon)}\right] \\
&\leq \frac{\left(1 + \frac{e^{\eta n} - 1}{n}\right)^T n}{e^{\eta T(1+\varepsilon)}} \\
&= e^{\frac{\varepsilon}{2}\frac{T}{n}(-\varepsilon) + \log n}
\end{aligned}$$

So that $T = c\frac{n \log n}{\varepsilon^2}$.

$\square$

We want:

$$p_e = \frac{\left\| L^{-\frac{1}{2}} \chi_e \right\|^2}{n} = \frac{\chi_e^T L^{-1} \chi_e}{n}$$

We will see that we can compute all the $p_e$s in time $O(m \log n)$.