ORIE 6334 Spectral Graph Theory	November 3, 2016
Lecture 21	
Lecturer: David P. Williamson	*Scribe: Seung Won (Wilson) Yoo*

1 Matrix Multiplicative Weights

Just like matrix Chernoff bounds were a generalization of scalar Chernoff bounds, the multiplicative weights algorithm can be generalized to matrices. Recall that in the setup for the multiplicative weight update algorithm, we had a sequence of time steps $t = 1, \dots, T$; in each time step t, we made a decision $i \in \{1...N\}$ and got a value $v_t(i) \in [0, 1]$. After we made a decision in time step t, we got to see all the values

In matrix multiplicative weights, we make a decision $u \in \mathbb{R}^n$, $||u|| = 1$ and get a value $u^T M_t u$ where $0 \preceq M_t \preceq I$, $M_t \in \mathbb{R}^{n\times n}$, so that $u^T M_t u \in [0, 1]$. As with multiplicative weights, we make a randomized decision for the vector u based on some weights. We now maintain a weight matrix $W_t \in \mathbb{R}^{n\times n}$, $W_t \succeq 0$. Let $P_t = \frac{W_t}{\mathrm{tr}(W_k)}$ so that $\mathrm{tr}(P_t) = 1$ and $P_t \succeq 0$. If λ_{it} are eigenvalues of P_t, and x_{it} are the corresponding orthonormal eigenectors, then $P_t = \sum_{i=1}^n \lambda_{it} x_{it} x_{it}^T$, $\lambda_{it} \geq 0$, $\sum_{i=1}^n \lambda_{it} = 1$; that is, P_t is a discrete distribution over the vectors x_{it}, and we will choose the vector x_{it} with probability λ_{it}.

Algorithm 1: Matrix Multiplicative Weights

$W_t \leftarrow I$
for $t \leftarrow 1$ **to** T **do**
  $P_t \leftarrow \frac{W_t}{tr(W_t)}$
  Make decision $u_t = x_{it}$ with prob. λ_{it} for x_{it}, λ_{it} eigenvectors/eigenvalues of P_t
  Get value $u_t^T M_t u_t$
  $W_{t+1} \leftarrow \exp(\epsilon \sum_{k=1}^T M_k)$.
end

This is a generalization of the multiplicative weights algorithm as one can think of all of the matrices as diagonal, and the values that are associated with each of the n decisions as each entry on the diagonal of M_t. In this case, the weights are maintained on the diagonal of W_t as well.

We introduce a new piece of notation:

$$A \bullet B \equiv \sum_{i,j} a_{ij} b_{ij},\ A = (a_{ij}), B = (b_{ij})$$

[0]This lecture is drawn from Arora and Kale 2016 `http://dl.acm.org/citation.cfm?doid=2837020`; Kale's thesis `http://www.satyenkale.com/papers/thesis.pdf`; and de Carli Silva, Harvey, and Sato 2015 `https://www.cs.ubc.ca/~nickhar/Publications/SparsifierMMWUM/SparsifierMMWUM.pdf`.

Then the expected value of the algorithm is:

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{n}\lambda_{it}(x_{it}^T M_t x_{it}) &= \sum_{t=1}^{T}\sum_{i=1}^{n}\lambda_{it}(x_{it}x_{it}^T \bullet M_t)\\ &= \sum_{t=1}^{T}(\sum_{i=1}^{n}\lambda_{it}x_{it}x_{it}^T)\bullet M_t)\\ &= \sum_{t=1}^{T} P_t \bullet M_t.\end{aligned}$$

We want to show that the algorithm does as well as any fixed decision u, $||u|| = 1$. Note that for a fixed decision u,

$$\sum_{t=1}^{T} u^T M_t u = u^T\left(\sum_{t=1}^{T} M_t\right)u \le \max_{u:||u||=1} u^T\left(\sum_{t=1}^{T} M_t\right)u = \lambda_{\max}\left(\sum_{t=1}^{T} M_t\right).$$

Thus the best fixed decision is the eigenvector corresponding to the maximum eigenvalue of $\sum_{t=1}^{T} M_t$.

To carry out our analysis, we need the following facts.

Theorem 1 (Golden-Thompson Inequality)

$$\mathrm{tr}(\exp(A+B)) \le \mathrm{tr}(\exp(A)\exp(B)).$$

Claim 2 $\mathrm{tr}(AB) = A \bullet B$ *for either* A, B *symmetric*

Claim 3 $X \bullet A \le X \bullet B$ *if* $A \preceq B$, $X \succeq 0$.

Claim 4 *If* $0 \preceq A \preceq I$, *then*

$$\exp(\epsilon A) \preceq I + (e^{\epsilon} - 1)A.$$

We can now prove a theorem analogous to the one we proved for the multiplicative weights update algorithm.

Theorem 5 *Let* $0 \le \epsilon \le \frac{1}{2}$. *Then* $\sum_{t=1}^{T} P_t \cdot M_t \ge \frac{1}{1+\epsilon}\lambda_{max}(\sum_{t=1}^{T} M_t) - \frac{1}{\epsilon} ln\, n$.

Proof: The proof mirrors that of the scalar multiplicative weights algorithm's proof. We start by getting an upper and lower bound on $\mathrm{tr}(W_{T+1})$.

$$\begin{aligned}\mathrm{tr}(W_{t+1}) &= \mathrm{tr}\left(\exp\left(\epsilon\sum_{k=1}^{t} M_k\right)\right)\\ &\le \mathrm{tr}\left(\exp\left(\epsilon\sum_{k=1}^{t-1} M_k\right)\exp(\epsilon M_t)\right)\\ &= W_t \bullet \exp(\epsilon M_t)\\ &= \mathrm{tr}(W_t)P_t \bullet \exp(\epsilon M_t)\\ &\le \mathrm{tr}(W_t)P_t \bullet (I + (e^{\epsilon}-1)M_t)\\ &= \mathrm{tr}(W_t)(1 + (e^{\epsilon}-1)P_t \bullet M_t)\\ &\le \mathrm{tr}(W_t)(\exp(e^{\epsilon}-1)P_t \bullet M_t)).\end{aligned}$$

The first inequality follows from Golden-Thompson, the second follows from Claims 3 and 4 combined, and the third follows from $1+x \leq \exp(x)$. We can determine $\mathrm{tr}(W_{T+1})$ by a telescoping product, getting that

$$\mathrm{tr}(W_{T+1}) \leq \mathrm{tr}(W_1) \exp\left((e^\epsilon - 1)\sum_{t=1}^{T} P_t \bullet M_t\right) = n \exp\left((e^\epsilon - 1)\sum_{t=1}^{T} P_t \cdot M_t\right).$$

For the lower bound,

$$\begin{aligned}
\mathrm{tr}(W_{T+1}) &\geq \lambda_{\max}(W_{T+1}) \\
&= \lambda_{\max}\left(\exp\left(\epsilon\sum_{t=1}^{T} M_t\right)\right) \\
&= \exp\left(\lambda_{\max}\left(\epsilon\sum_{t=1}^{T} M_t\right)\right).
\end{aligned}$$

The last step follows from the fact that taking maximum eigenvalue of a matrix derived by exponentiating all of the eigenvalues is the same as taking the exponential of the maximum eigenvalue.

Given the upper bound and lower bound on $\mathrm{tr}(W_{T+1})$, we then get

$$n \exp\left((e^\epsilon - 1)\sum_{t=1}^{T} P_t \bullet M_t\right) \geq \exp\left(\lambda_{max}\left(\epsilon\sum_{t=1}^{T} M_t\right)\right).$$

Taking the log of both sides and rearranging, we get

$$\begin{aligned}
\sum_{t=1}^{T} P_t \cdot M_t &\geq \frac{\epsilon}{e^\epsilon - 1}\lambda_{\max}\left(\sum_{t=1}^{T} M_t\right) - \frac{1}{e^\epsilon - 1}\ln n \\
&\geq \frac{1}{1+\epsilon}\lambda_{\max}\left(\sum_{t=1}^{T} M_t\right) - \frac{1}{\epsilon}\ln n.
\end{aligned}$$

In the last inequality we use $e^\epsilon - 1 \leq \epsilon(1+\epsilon)$, for $0 \leq \epsilon \leq \frac{1}{2}$, and $e^\epsilon - 1 \geq \epsilon$. $\square$

2 A Feasibility Problem and Application to Spectral Sparsification

Just as we did with the multiplicative weights algorithm, we now want to apply matrix multiplicative weights to a feasibility problem. We do so here as follows. Suppose we have B_i, $i = 1, \ldots, m$, with $B_i \succeq 0$ for all i, and $\sum_{i=1}^m B_i = I$. We want to find a sparse weighting $y \in \mathbb{R}^m \geq 0$ such that $(1-\epsilon)I \preceq \sum_{i=1}^m y(i)B_i \preceq (1+\epsilon)I$. Assume we have an oracle such that given $P, \tilde{P} \succeq 0$ with $\mathrm{tr}(P) = \mathrm{tr}(\tilde{P}) = 1$, the oracle returns a y such that $y(i) \neq 0$ at only one entry i, $y(i) = \alpha$ and $\alpha P \bullet B_i \leq (1+\epsilon)$ and $\alpha \tilde{P} \bullet B_i \geq (1-\epsilon)$.

We define the *width* of the oracle as

$$\rho \equiv \max_y \alpha\, tr(B_i)$$

over all y returned by oracle.

The application to spectral sparsification is as follows. We have m matrices, and one matrix for every edge in our graph. Let us index those matrices by the edges in our graph:

$$B_{(i,j)} = L_G^{\dagger/2}(e_i - e_j)(e_i - e_j)^T L_G^{\dagger/2}$$

We want the sum of them to be the identity matrix. We showed it last time but we show it again.

$$\begin{aligned}\sum_{(i,j)\in E} B_{(i,j)} &= L_G^{\dagger/2}(\sum_{(i,j)\in E}(e_i - e_j)(e_i - e_j)^T)L_G^{\dagger/2} \\ &= L_G^{\dagger/2} L_G L_G^{\dagger/2} \\ &= I^*\end{aligned}$$

(Recall that this is the identity when multiplied by any vector orthogonal to e.) So what's our sparse solution going to be? If this algorithm works, we get a sparse y such that

$$(1-\epsilon)I \preceq L_G^{\dagger/2}\left(\sum_{(i,j)\in E} y_{(i,j)}(e_i - e_j)(e_i - e_j)^T\right)L_G^{\dagger/2} \preceq (1+\epsilon)I.$$

We showed last time that this equation is satisfied for some vector y if and only if if subgraph H of G is a spectral sparsifier using the weights given by $y_{(i,j)}$.

$$(1-\epsilon)L_G \preceq L_H \preceq (1+\epsilon)L_G.$$

In the following algorithm, the two weight matrices W_t and $\tilde{W}_t$ ensure that the resulting sparse sum does not get larger than $(1+\epsilon)I$ and does not get smaller than $(1-\epsilon)I$.

Algorithm 2: Algorithm for Feasibility

$W_1 \leftarrow I, \tilde{W}_1 \leftarrow I$
for $t \leftarrow 1$ **to** T **do**
  $P_t \leftarrow \frac{W_t}{\mathrm{tr}(W_t)}, \tilde{P}_t \leftarrow \frac{\tilde{W}_t}{\mathrm{tr}(\tilde{W}_t)}$
  Run oracle to find y_t such that only one i st $y_t(i) = \alpha_t \geq 0$, $\alpha_t P_t \bullet B_{it} \leq (1+\epsilon)$, $\alpha \tilde{P}_t \bullet B_{it} \geq (1-\epsilon)$
  $W_t \leftarrow \exp(\frac{\epsilon}{\rho}\sum_{k=1}^{t}\sum_{i=1}^{m} y_k(i)B_i)$
  $\tilde{W}_t \leftarrow \exp(-\frac{\epsilon}{\rho}\sum_{k=1}^{t}\sum_{i=1}^{m} y_k(i)B_i)$
end
return $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$

An upper bound on the number on the number of nonzeros in $\bar{y}$ is T because at every timestep we increase exactly one index of $\bar{y}$. We also notice that $\frac{1}{\rho}\sum_{i=1}^{m} y_t(i)B_i$ plays the role of M_t from matrix multiplicative weights in the algorithm above since

$$\alpha_t \,\mathrm{tr}(B_i) \leq \rho \implies 0 \preceq \frac{1}{\rho}\sum_{i=1}^{m} y_t(i)B_i \preceq I.$$

It then follows that
$$\sum_{t=1}^{T} P_t \bullet \left(\frac{1}{\rho}\sum_{i=1}^{m} y(i)B_i\right) \leq \frac{T(1+\epsilon)}{\rho}.$$

Theorem 5 guarantees that

$$\sum_{t=1}^{T} P_t \bullet \left(\frac{1}{\rho}\sum_{i=1}^{m} y_t(i)B_i\right) \geq \frac{1}{1+\epsilon}\lambda_{\max}\left(\frac{1}{\rho}\sum_{t=1}^{T}\sum_{i=1}^{m} y_t(i)B_i\right) - \frac{1}{\epsilon}\ln n.$$

If we choose $T = \frac{(1+\epsilon)\rho}{\epsilon^2}\ln n$, we have that

$$\begin{aligned}
\frac{1}{1+\epsilon}\lambda_{\max}\left(\frac{1}{\rho}\sum_{t=1}^{T}\sum_{i=1}^{m} y_t(i)B_i\right) - \frac{1}{\epsilon}\ln n &\leq \sum_{t=1}^{T} P_t \bullet \left(\frac{1}{\rho}\sum_{i=1}^{m} y(i)B_i\right) \leq \frac{T(1+\epsilon)}{\rho} \\
\frac{T}{\rho(1+\epsilon)}\lambda_{\max}\left(\sum_{i=1}^{m} \bar{y}(i)B_i\right) - \frac{1}{\epsilon}\ln n &\leq \frac{T(1+\epsilon)}{\rho} \\
\lambda_{\max}\left(\sum_{i=1}^{m} \bar{y}(i)B_i\right) &\leq (1+\epsilon)^2 + \frac{(1+\epsilon)\rho}{T\epsilon}\ln n \\
&\leq (1+\epsilon)^2 + \epsilon \\
&\leq (1+4\epsilon).
\end{aligned}$$

Similarly, we can show that

$$\lambda_{\min}\left(\sum_{i=1}^{m} \bar{y}(i)B_i\right) \geq 1 - 4\epsilon,$$

so that we have
$$(1-4\epsilon)I \preceq \sum_{i=1}^{m} \bar{y}(i)B_i \preceq (1+4\epsilon)I.$$

As stated above, $\bar{y}$ has at most $T = O(\frac{\rho}{\epsilon^2}\ln n)$ nonzeroes. In the lecture we did not have time to show the lemma below, which states that we can find an oracle with $\rho = O(\frac{(1+\epsilon)n}{\epsilon})$, which implies $O((n\ln n)/\epsilon^3)$ nonzeroes. It is possible to modify the algorithm to obtain $O((n\ln n)/\epsilon^2)$ nonzeroes.

Lemma 6 *There is an oracle with width $\rho = O(\frac{(1+\epsilon)n}{\epsilon})$.*

Proof: Recall that the oracle needs to find i and α such that $\alpha P \bullet B_i \leq 1+\epsilon$, $\alpha\tilde{P} \bullet B_i \geq 1-\epsilon$, and $\alpha\,\mathrm{tr}(B_i) \leq \rho = (1+\epsilon)n/\epsilon$.

Define $\tilde{p}_i = B_i \bullet \tilde{P}$. Then $\tilde{p}_i \geq 0$ since $P \succeq 0$ and $B_i \succeq 0$. Also

$$\sum_{i=1}^{n} \tilde{p}_i = \tilde{P} \bullet \left(\sum_{i=1}^{n} B_i\right) = \tilde{P} \bullet I = \mathrm{tr}(\tilde{P}) = 1.$$

So $\tilde{p}_i$ is a probability distribution.

Then

$$E_i\left[\frac{\operatorname{tr}(B_i)}{\tilde{p}_i}\right] = \sum_{i=1}^{m} \operatorname{tr}(B_i) = \operatorname{tr}(I) = n,$$

so that

$$\Pr\left[\frac{\operatorname{tr}(B_i)}{\tilde{p}_i} \leq \frac{(1+\epsilon)n}{\epsilon}\right] = 1 - \Pr\left[\frac{\operatorname{tr}(B_i)}{\tilde{p}_i} > \frac{(1+\epsilon)n}{\epsilon}\right] > 1 - \frac{\epsilon}{1+\epsilon} = \frac{1}{1+\epsilon},$$

by Markov's inequality. Similarly,

$$E_i\left[\frac{P \bullet B_i}{\tilde{p}_i}\right] = \sum_{i=1}^{m} P \bullet B_i = P \bullet I = \operatorname{tr}(P) = 1,$$

so that

$$\Pr\left[\frac{P \bullet B_i}{\tilde{p}_i} \leq 1+\epsilon\right] = 1 - \Pr\left[\frac{P \bullet B_i}{\tilde{p}_i} > 1+\epsilon\right] > 1 - \frac{1}{1+\epsilon},$$

again by Markov's inequality.

So there must exist an index i such that both

$$\frac{P \bullet B_i}{\tilde{p}_i} \leq 1+\epsilon \text{ and } \frac{\operatorname{tr}(B_i)}{\tilde{p}_i} \leq \frac{(1+\epsilon)n}{\epsilon} \equiv \rho.$$

Thus if we set $\alpha = 1/\tilde{p}_i$, we get that $\alpha P \bullet B_i \leq 1+\epsilon$, $\alpha \operatorname{tr}(B_i) \leq \rho$, and

$$\alpha \tilde{P} \bullet B_i = \frac{1}{\tilde{p}_i} \tilde{P} \bullet B_i = 1 \geq 1 - \epsilon,$$

where the final equation follows by the definition of $\tilde{p}_i$. $\square$