The Sparse-Plus-Low-Rank Quasi-Newton Method for Entropic Regularized Optimal Transport

Yixuan Qiu [1] **Chenrui Wang** [1]

Abstract

We propose an extension to the Sinkhorn Algorithm by introducing a new Hessian sparsification scheme that significantly reduces the number of non-zero entries. Remarkably, it can be shown that the condition number of the sparsified Hessian is strictly smaller than the original Hessian. We also introduce two low-rank terms to adapt the algorithm to dense Hessian. Finally, a number of numerical experiments are conducted to demonstrate the effectiveness of the proposed algorithm in solving large-scale OT problems.

1. Introduction

Optimal transport (OT) theory, as described by Gaspard Monge (1746–1818), addresses the problem of moving a distribution (e.g., a pile of sand) to match a target configuration (e.g., a prescribed shape) while minimizing a cost, such as the total distance or effort required. The OT problem between two distributions μ and ν can be characterized by the following linear programming problem:

$$\min_{P\mathbf{1}_m=a, P^T\mathbf{1}_n=b} \langle P, M\rangle$$

where a, b are two vectors satisfying $a^T\mathbf{1}_n = b^T\mathbf{1}_m = 1$, P is a matrix with all positive entries and M is a given cost matrix.

2. Background and Related Work

Notation Determine a set of consistent notations.

The Sinkhorn Distance The definition of Sinkhorn Distance and the main objective, etc..

Sparsification Schemes How other algorithms sparsify the Hessian.

Computational Sinkhorn OT Introduce algorithms to solve the Sinkhorn OT problem.

OT in machine learning SNS and SSNS both have this part.

3. The Proposed Algorithm

In this article we propose the Sparse-Plus-Low-Rank(SPLR) algorithm, which mainly consists of two steps: first, we construct an approximation of the original Hessian; second, we update the solution iteratively with a Newton-type subroutine.

3.1. Sparsifying the Hessian Matrix

A crucial point that we care about is how to design an algorithm to sparsify a matrix such that the sparsified Hessian matrix remains positive definite while retaining as few non-zero elements as possible. We conducted a simple test of several sparsification methods, sparsifying T and using the sparsified $\tilde{T}$ to construct the sparsified Hessian H_δ. Then, we computed the eigenvalues of both H and H_δ, and the results are very interesting: it always holds that $\lambda_{\max}(H) > \lambda_{\max}(H_\delta), \lambda_{\min}(H) < \lambda_{\min}(H_\delta)$, which inspires us to hypothesize under what conditions such results may occur.

But instead of directly comparing H and H_δ, we first consider the effect of setting one element of T to 0.

Theorem 3.1. *T_{δ_1} is T with some elements set to 0, T_{δ_2} is T_{δ_1} with one element set to 0. Suppose that $T_{\delta_1} \in \mathbb{R}^{n\times m}_{\geq 0} = \left(t^{(1)}_{\delta_1} \quad \dots \quad t^{(m)}_{\delta_1}\right) = \left(\tilde{T_{\delta_1}} \quad t^m_{\delta_1}\right), \|\tilde{T_{\delta_1}}\|_\infty > 0, T_{\delta_2} = T_{\delta_1} - T_{ij}\mathbf{e}_i\mathbf{e}_j^T, i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m-1\}$. The corresponding Hessian matrices are:*

$$H_{\delta_1} = \begin{pmatrix} \boldsymbol{diag}(T\mathbf{1}_m) & \tilde{T_{\delta_1}} \\ \tilde{T_{\delta_1}}^T & \boldsymbol{diag}(T^T\mathbf{1}_n) \end{pmatrix}$$

$$H_{\delta_2} = \begin{pmatrix} \boldsymbol{diag}(T\mathbf{1}_m) & \tilde{T_{\delta_2}} \\ \tilde{T_{\delta_2}}^T & \boldsymbol{diag}(T^T\mathbf{1}_n) \end{pmatrix}$$

If there exists a row and a column in $\tilde{T_{\delta_1}}$ that is strictly

greater than 0*, then we have:*

$$\lambda_{\max}(H_{\delta_1}) > \lambda_{\max}(H_{\delta_2})$$
$$\lambda_{\min}(H_{\delta_1}) < \lambda_{\min}(H_{\delta_2})$$

Proof. First we prove $\lambda_{\max}(H_{\delta_1}) > \lambda_{\max}(H_{\delta_2})$.

Define $p = i, q = n + j, \beta = (H_{\delta_1})_{pq}$, then the difference of Hessian is:

$$H_{\delta_2} - H_{\delta_1} = -\beta \left(\mathbf{e}_p \mathbf{e}_q^T + \mathbf{e}_q \mathbf{e}_p^T\right) \triangleq -\beta \mathbf{J}$$

Define $M(\kappa) = H_{\delta_1} - \kappa J, l(\kappa) = \lambda_{\max}(M(\kappa)), \kappa \in \mathbb{R}$, then we have:

$$\lambda_{\max}(H_{\delta_2}) - \lambda_{\max}(H_{\delta_1}) = l(\beta) - l(0)$$

According to A.4, we can show that $\lambda_{\max}$ is differentiable on $\{M(\kappa) | \kappa \in [0, \beta]\}$. Suppose the eigenvector associated with $\lambda_{\max}(M(\kappa))$ is $\mathbf{u}_\kappa$, then the derivative at $M(\kappa)$ is:

$$\frac{\partial \lambda_{\max}}{\partial M}\Big|_{M=M(\kappa)} = \mathbf{u}_\kappa \mathbf{u}_\kappa^T$$

thus $l'(\kappa) = \langle \mathbf{u}_\kappa \mathbf{u}_\kappa^T, -J \rangle$. According to the Lagrange's mean value theorem, $\exists \xi \in (0, \beta)$ such that:

$$\begin{aligned} l(\beta) - l(0) &= l'(\xi)(\beta - 0) \\ &= \langle \mathbf{u}_\xi \mathbf{u}_\xi^T, -J \rangle \beta \\ &= -\beta \left[\boldsymbol{tr}(\mathbf{u}_\xi \mathbf{u}_\xi^T \mathbf{e}_p \mathbf{e}_q^T) + \boldsymbol{tr}(\mathbf{u}_\xi \mathbf{u}_\xi^T \mathbf{e}_q \mathbf{e}_p^T)\right] \\ &= -\beta \left[\boldsymbol{tr}(\mathbf{u}_\xi^T \mathbf{e}_p \mathbf{e}_q^T \mathbf{u}_\xi) + \boldsymbol{tr}(\mathbf{u}_\xi^T \mathbf{e}_q \mathbf{e}_p^T \mathbf{u}_\xi)\right] \\ &= -2\beta (\mathbf{u}_\xi)_p (\mathbf{u}_\xi)_q \end{aligned} \tag{1}$$

According to A.4, $\mathbf{u}_\xi$ can be normalized to have strictly positive entries so that $(\mathbf{u}_\xi)_p (\mathbf{u}_\xi)_q > 0$, which means $l(\beta) - l(0) = \lambda_{\max}(H_{\delta_2}) - \lambda_{\max}(H_{\delta_1}) < 0$.

Then we prove $\lambda_{\min}(H_{\delta_1}) < \lambda_{\min}(H_{\delta_2})$.

According to [SSNS], we have $M(\kappa) \succ 0$ thus invertible, then according to A.1, $M(\kappa)^{-1}$ is of the form:

$$M(\kappa)^{-1} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$$

where $A > 0, C > 0, B < 0$. Then according to A.2, we can show that $\lambda_{\max}(M(\kappa)^{-1})$ is a simple positive eigenvalue, the corresponding eigenvector $\mathbf{v}_\kappa = (\mathbf{v}_1^T, \mathbf{v}_2^T)^T$ where $\mathbf{v}_1 \in \mathbb{R}^n, \mathbf{v}_2 \in \mathbb{R}^{m-1}$ can be normalized such that $\mathbf{v}_1 > 0, \mathbf{v}_2 < 0$. Since $\mathbf{v}_\kappa$ is also the eigenvector of $M(\kappa)$ corresponding to $\lambda_{\min}(M(\kappa))$, we can tell that $(\mathbf{v}_\kappa)_p (\mathbf{v}_\kappa)_q < 0$, where $p \le n, q > n$.

Define $h(\kappa) = \lambda_{\min}(M(\kappa))$, then similar to 1 we have $h(\beta) - h(0) = -2\beta (\mathbf{v}_\xi)_p (\mathbf{v}_\xi)_q$, where $\xi \in (0, \beta)$. Recall that $(\mathbf{v}_\xi)_p (\mathbf{v}_\xi)_q < 0$, so $h(\beta) - h(0) > 0$, which means $\lambda_{\min}(H_{\delta_1}) < \lambda_{\min}(H_{\delta_2})$. □

Theorem 3.2. *Suppose that $\tilde{T}_\delta$ is $\tilde{T}$ with some elements set to* 0*. If there exists a row and a column in T_δ that is strictly greater than* 0*, then we have:*

$$\lambda_{\max}(H) > \lambda_{\max}(H_\delta)$$
$$\lambda_{\min}(H) < \lambda_{\min}(H_\delta)$$

Proof. According to mathematical induction, we can continuously apply 3.1 to prove this conclusion. □

Therefore, our sparsifying scheme is to keep the first row and first column of T, then set the lower 90% of other elements to 0. It is worth noting that the result of 3.2 not only guarantees H_δ is positive definite, but also shows that the condition number of H_δ is strictly smaller than H which makes CG faster theoretically.

Algorithm 1 Sparsification Algorithm: Setting lower 90% of elements to 0 while keeping the first row and first column

Input: A matrix $\tilde{T}$ of size $n \times m - 1$
Output: A sparsified matrix $\tilde{T}_\delta$ where the lower 90% of elements are set to 0 while entries at the first row and first column are positive
Return: $\tilde{T}_\delta$

3.2. Low-Rank Terms

When $\tilde{T}$ is generally sparse, then H_δ should be a good sparse approximation of H. However, there are cases where $\tilde{T}$ is not as sparse as expected. Inspired by the BFGS algorithm, we approximate the Hessian matrix by adding two low-rank terms to solve this problem.

Assume that $H(x^k) \approx H_\delta(x^k) + a\mathbf{u}\mathbf{u}^T + b\mathbf{v}\mathbf{v}^T \triangleq \hat{H(x^k)}, a, b \in \mathbb{R}, \mathbf{u}, \mathbf{v} \in \mathbb{R}^{n+m-1}$, and we will determine $a, b, \mathbf{u}, \mathbf{v}$ by the first-order approximation of g at x^{k+1}:

$$g(x) = g(x^{k+1}) + H(x^{k+1})(x - x^{k+1}) + O(\|x - x^{k+1}\|^2)$$

Let $x = x^k, s^k = x^{k+1} - x^k, y^k = g(x^{k+1}) - g(x^k)$ and we will get:

$$H(x^{k+1}) s^k + O(\|s^k\|^2) = y^k$$

Ignoring $O(\|s^k\|^2)$, we expect $\hat{H(x^{k+1})}$ to satisfy:

$$\begin{aligned} \hat{H(x^{k+1})} s^k &= y^k \\ (H_\delta(x^{k+1}) + a\mathbf{u}\mathbf{u}^T + b\mathbf{v}\mathbf{v}^T) s^k &= y^k \\ (a \cdot \mathbf{u}^T s^k)\mathbf{u} + (b \cdot \mathbf{v}^T s^k)\mathbf{v} &= y^k - H_\delta(x^{k+1}) s^k \end{aligned}$$

For the sake of simplicity, we set:

$$\begin{cases} \mathbf{u} = y^k \\ \mathbf{v} = H_\delta(x^{k+1})s^k \\ a \cdot \mathbf{u}^T s^k = 1 \\ b \cdot \mathbf{v}^T s^k = 1 \end{cases}$$

Then we get:

$$\begin{cases} a = \dfrac{1}{\mathbf{u}^T s^k} \\ b = -\dfrac{1}{\mathbf{v}^T s^k} \\ \mathbf{u} = y^k \\ \mathbf{v} = H_\delta(x^{k+1})s^k \end{cases}$$

Another significant reason for us to employ low-rank terms for approximation is that we can compute the inverse of $H_\delta(\hat{x^{k+1}})$ very conveniently by using the [Woodbury matrix identity]:

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$$

where

$$\begin{cases} A = H_\delta(x^{k+1}) \\ U = \begin{pmatrix} \mathbf{u} & \mathbf{v} \end{pmatrix} \\ C = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \\ V = \begin{pmatrix} \mathbf{u}^T \\ \mathbf{v}^T \end{pmatrix} \end{cases}$$

3.3. The SPLR Algorithm

3.2 shows that H_δ is guaranteed to be positive definite with specific sparsification schemes.

4. Numerical Experiments

Here are some numerical experiments.

5. Conclusion

Finally here is the conclusion.

References

Meyer Jr, C. and Stadelmaier, M. Singular m-matrices and inverse positivity. *Linear Algebra and its Applications*, 22:139–156, 1978.

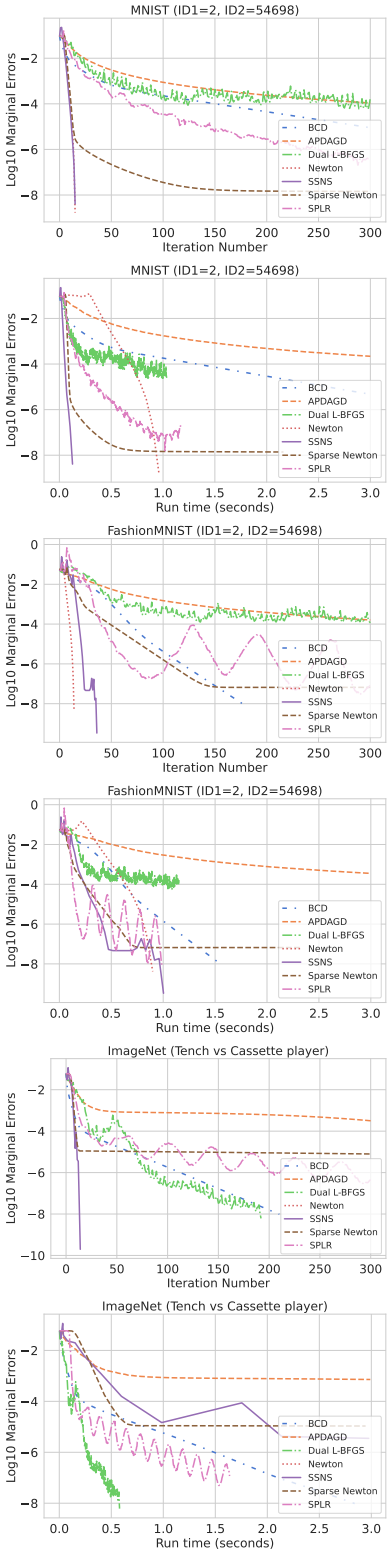

Figure 1. Top: Marginal error vs. iteration number for different algorithms on two datasets. Bottom: Marginal error vs. run time. The cost matrix is based on the l_1 -distance, and $\eta = 0.01$.

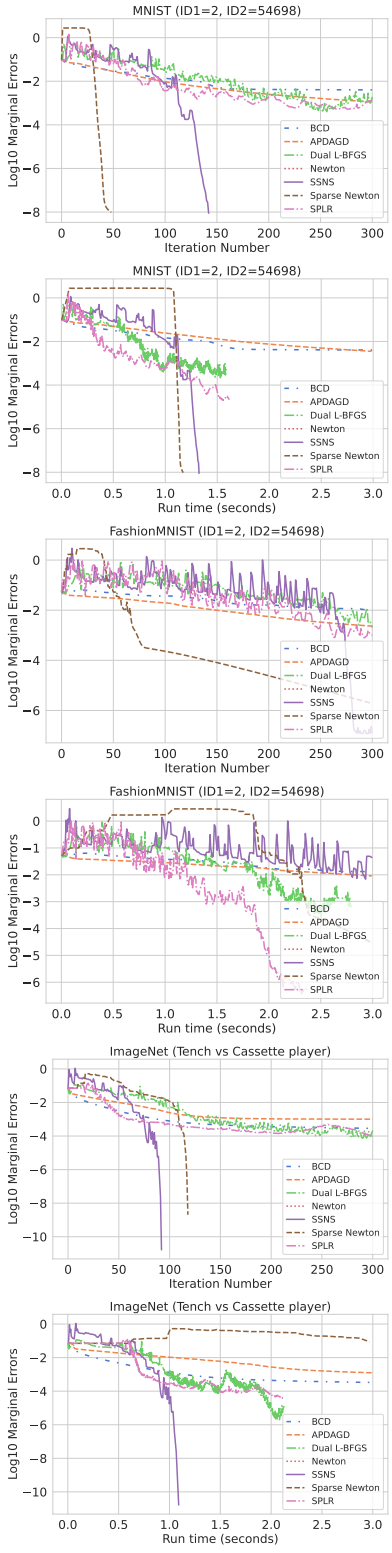

Figure 2. Top: Marginal error vs. iteration number for different algorithms on two datasets. Bottom: Marginal error vs. run time. The cost matrix is based on the l_1 -distance, and $\eta = 0.001$.

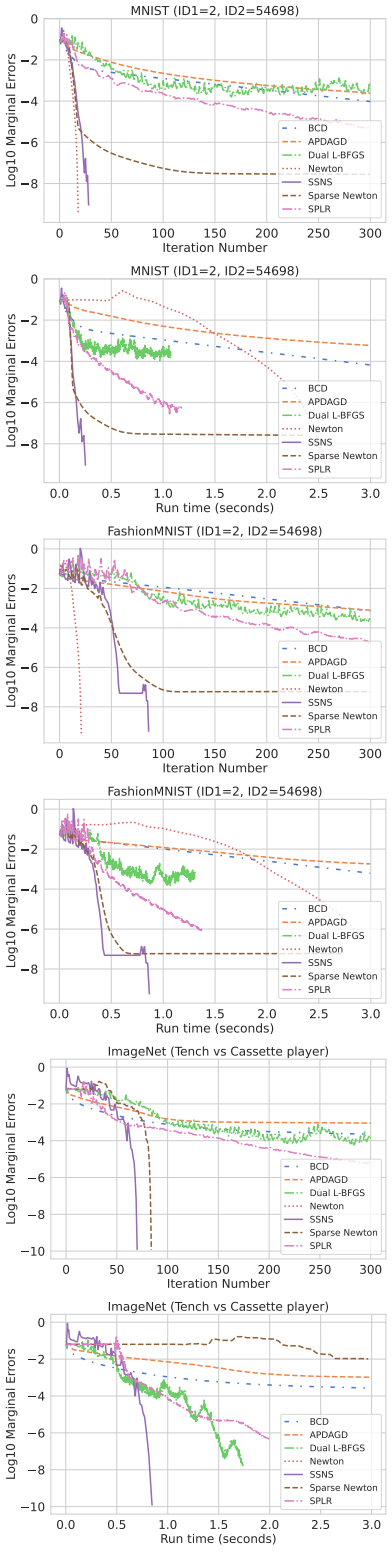

Figure 3. Top: Marginal error vs. iteration number for different algorithms on two datasets. Bottom: Marginal error vs. run time. The cost matrix is based on the l_2 -distance, and $\eta = 0.001$.

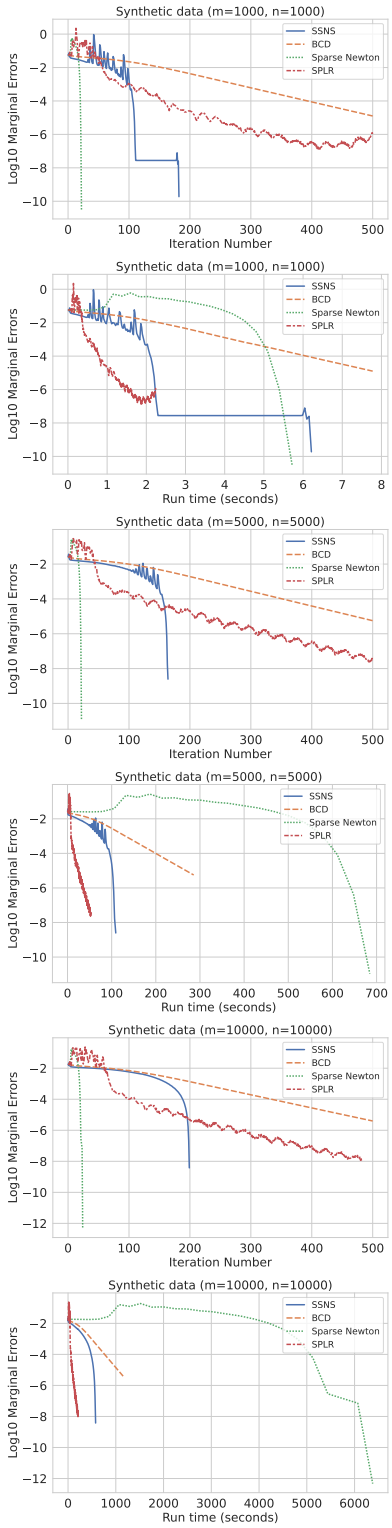

Figure 4. Comparing BCD and SSNS on large OT problems. Top: Marginal error vs. iteration number for different problem sizes. Bottom: Marginal error vs. run time.

Algorithm 2 The SPLR Algorithm

Input: $\varepsilon_{tol} > 0, \varepsilon > 0$
Output: x_k
Initialize $x_0 \leftarrow \mathbf{0}$, $x_1 \leftarrow optimize\_\beta(\mathbf{0})$
Compute $f_0 = f(x^0), g_0 = g(x^0)$
for $k = 1, 2, \ldots$ **do**
  Compute $f_k = f(x^k), g_k = g(x^k), H_{\delta_k} = H_\delta(x^k)$
  if $\|g_k\| < \varepsilon_{tol}$ **then**
    return x_k
  end if
  Compute $s \leftarrow x^k - x^{k-1}, y \leftarrow g_k - g_{k-1}$
  Compute $\mathbf{u} \leftarrow y, \mathbf{v} = H_{\delta_k} s$
  Compute $a \leftarrow \frac{1}{\mathbf{u}^T s}, b \leftarrow -\frac{1}{\mathbf{v}^T s}$
  if $\frac{\langle y,s \rangle}{\langle y,y \rangle} > \varepsilon$ **then**
    Compute $p_k = -(H_{\delta_k} + a\mathbf{u}\mathbf{u}^T + b\mathbf{v}\mathbf{v}^T)^{-1} g_k$ using the Woodbury Identity
  else
    Compute $p_k = -(H_{\delta_k})^{-1} g_k$
  end if
  Compute $\alpha_k = armijo\_line\_search(p_k, x_k)$
  Update $x_{k+1} \leftarrow x_k + \alpha_k p_k$
end for

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.

A.1. Technical Lemmas

Lemma A.1. *Let M be a matrix of the form*

$$M = \begin{bmatrix} D_1 & R \\ R^T & D_2 \end{bmatrix},$$

where $D_1 \in \mathbb{R}^{r\times r}$ and $D_2 \in \mathbb{R}^{s\times s}$ are diagonal matrices, and $R \in \mathbb{R}^{r\times s}$ has strictly positive entries. Suppose that M is positive definite, and M^{-1} has the partition

$$M^{-1} = \begin{bmatrix} A_{r\times r} & B_{r\times s} \\ B^T & C_{s\times s} \end{bmatrix},$$

and then we have $A > 0$, $C > 0$, and $B < 0$, where the inequality signs apply elementwisely.

Proof. Since M is positive definite, all of its diagonal elements must be strictly positive, implying that both D_1 and D_2 are invertible. By the inversion formula of block matrices, we have

$$\begin{aligned} A &= (D_1 - RD_2^{-1}R^T)^{-1}, \\ C &= (D_2 - R^T D_1^{-1} R)^{-1}, \\ B &= -D_1^{-1}R(D_2 - R^T D_1^{-1}R)^{-1} = -D_1^{-1}RC. \end{aligned}$$

Clearly, $J \coloneqq D_1 - RD_2^{-1}R^T$ is the Schur complement of the block D_2 of the matrix M. By the properties of the Schur complement, we know that J is positive definite, and hence it is nonsingular.

Moreover, since R has strictly positive entries and D_2 has positive diagonal elements, we have that $RD_2^{-1}R^T$ has strictly positive entries. Therefore, J can be represented in the form $J = sI - L$, where $L > 0$, $s \geq \rho(L)$, and $\rho(\cdot)$ stands for the spectral radius. Clearly, J is irreducible, so by Theorem A(ii) of Meyer Jr & Stadelmaier (1978), $A = J^{-1}$ has strictly positive entries. The same argument can be used to prove that $C > 0$.

Finally, recall that $B = -D_1^{-1}RC$. Since $R > 0$, $C > 0$, and D_1 has positive diagonal elements, we conclude that $B < 0$.

□

Lemma A.2. *Let M be a matrix of the form*

$$M = \begin{bmatrix} A_{r\times r} & -B_{r\times s} \\ -B^T & C_{s\times s} \end{bmatrix},$$

where $A > 0$, $B > 0$, $C > 0$, and the inequality signs apply elementwisely. Then M has a positive eigenvalue r such that any other eigenvalue of M in absolute value is strictly smaller than r, and the eigenvector $v = (v_1^T, v_2^T)^T$ associated with r, where $v_1 \in \mathbb{R}^r$ and $v_2 \in \mathbb{R}^s$, can be normalized such that $v_1 > 0$ and $v_2 < 0$.

Proof. Define

$$Q = \begin{bmatrix} I_r & O \\ O & -I_s \end{bmatrix},$$

and then it is easy to show that $Q^{-1} = Q$, and

$$QMQ^{-1} = \begin{bmatrix} I_r & O \\ O & -I_s \end{bmatrix} \begin{bmatrix} A & -B \\ -B^T & C \end{bmatrix} \begin{bmatrix} I_r & O \\ O & -I_s \end{bmatrix} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \coloneqq \tilde{M}.$$

Therefore, M and $\tilde{M}$ are similar to each other, and hence they must share the same eigenvalues. Clearly, $\tilde{M}$ is a positive matrix, so by the Perron–Frobenius theorem, it must have a positive and simple eigenvalue r such that any other eigenvalue of $\tilde{M}$ in absolute value is strictly smaller than r. Moreover, $\tilde{M}$ has an eigenvector $\tilde{v} = (\tilde{v}_1^T, \tilde{v}_2^T)^T$ such that $\tilde{v}_1 \in \mathbb{R}^r$, $\tilde{v}_2 \in \mathbb{R}^s$, $\tilde{v} > 0$, and $\tilde{M}\tilde{v} = r\tilde{v}$.

Now let $v = Q^{-1}\tilde{v}$, and then

$$Mv = MQ^{-1}\tilde{v} = Q^{-1}QMQ^{-1}\tilde{v} = Q^{-1}\tilde{M}\tilde{v} = Q^{-1}r\tilde{v} = rv.$$

Therefore, v is an eigenvector of M. Note that

$$v = Q^{-1}\tilde{v} = \begin{bmatrix} I_r & O \\ O & -I_s \end{bmatrix} \begin{bmatrix} \tilde{v}_1 \\ \tilde{v}_2 \end{bmatrix} = \begin{bmatrix} \tilde{v}_1 \\ -\tilde{v}_2 \end{bmatrix},$$

so $v_1 = \tilde{v}_1 > 0$ and $v_2 = -\tilde{v}_2 < 0$, which implies the stated result. □

Lemma A.3. *T_δ is T with some elements set to 0, then:*

$$H_\delta = \begin{pmatrix} \boldsymbol{diag}(T\mathbf{1}_m) & \tilde{T}_\delta \\ \tilde{T}_\delta^{\,T} & \boldsymbol{diag}(\tilde{T}^T)\mathbf{1}_n \end{pmatrix}$$

is positive semi-definite.

Proof. According to the [Gershgorin circle theorem], if $R_i = \sum_{i \neq j} (H_\delta)_{ij}$, then all eigenvalues of H_δ satisfies:

$$|\lambda - H_{ii}| \leq R_i \leq (T\mathbf{1}_m)_i = H_{ii} \rightarrow 0 \leq \lambda \leq 2H_{ii}$$

which shows that H_δ is positive semi-definite. □

Lemma A.4. *Suppose that $G \in \mathbb{R}^{n \times (m-1)}$ is a matrix with the first row and first column strictly greater than 0 and all other elements equal to 0, i.e.:*

$$G_{ij} = \begin{cases} g_{ij} > 0, i = 1 \textit{ or } j = 1 \\ 0, i \neq 1 \textit{ and } j \neq 1 \end{cases}$$

Then the corresponding Hessian matrix:

$$H_G = \begin{pmatrix} \boldsymbol{diag}(T\mathbf{1}_m) & G \\ G^T & \boldsymbol{diag}(\tilde{T}^T)\mathbf{1}_n \end{pmatrix}$$

has the following properties:

i. The eigenvector $\mathbf{u}(H_G)$ corresponding to $\lambda_{\max}(H_G)$ can be normalized to have strictly positive entries.

ii. There exists a neighborhood $N(H_G)$ of H_G such that $\lambda_{\max}$ is differentiable.

Proof. First we show that $H_G^4 > 0$. Consider elements of ${H_G}^2$:

i. when $i \leq n, j \leq n$, $(H_G)_{ij}^2 = {H_G}^{(i)} \cdot {H_G}^{(j)} \geq g_{i,1} \times g_{j,1} > 0$

ii. when $i > n, j > n$, $(H_G)_{ij}^2 \geq g_{1,i-n} \times g_{1,j-n} > 0$

iii. when $i \leq n, j = n+1$, $(H_G)_{ij}^2 \geq H_{ii} \times g_{i,1} > 0$

iv. when $i = n+1, j \leq n$, $(H_G)_{ij}^2 \geq H_{jj} \times g_{j,1} > 0$

v. when $i = 1, j > n$, $(H_G^2)_{ij} \geq H_{11} \times g_{1,j-n} > 0$

vi. when $i > n, j = 1$, $(H_G^2)_{ij} \geq H_{11} \times g_{1,i-n} > 0$

Then consider elements of H_G^4:

i. Similarly we have $(H_G^4)_{ij} > 0$ when $i, j \leq n$ or $i, j > n$.

ii. When $i \leq n, j > n$, $(H_G^4)_{ij} = (H_G^2)^{(i)} \cdot (H_G^2)^{(j)} \geq (H_G^2)_{n+1,i} \times (H_G^2)_{n+1,j} > 0$

iii. When $i > n, j \leq n$, $(H_G^4)_{ij} = (H_G^2)^{(i)} \cdot (H_G^2)^{(j)} \geq (H_G^2)_{n+1,j} \times (H_G^2)_{n+1,i} > 0$

So when $k = 4$, we have $H_G^4 > 0$. According to the [Frobenious-Perron theorem], $\mathbf{u}(H_G)$ can be normalized to have strictly positive entries and $\lambda_{\max}(H_G)$ is strictly greater than all other eigenvalues of H_G, which means that $\lambda_{\max}$ is a simple positive eigenvalue, hence $\lambda_{\max}$ is differentiable at H_G according to [Magnus (1985)]. □