Supplementary Material: Faster Greedy MAP Inference for Determinantal Point Processes

A. Proof of Theorem 1

For given $X \subseteq \mathcal{Y}$, we denote that the true marginal gain Λ_i and the approximated gain Δ_i (used in Algorithm 1) as

$$\Lambda_i := \log\det L_{X\cup\{i\}} - \log\det L_X$$
$$\Delta_i := \left\langle \left(\overline{L}_X^{(j)}\right)^{-1}, L_{X\cup\{i\}} - \overline{L}_X^{(j)} \right\rangle + \left(\log\det \overline{L}_X^{(j)} - \log\det L_X\right)$$

where an item $i \in \mathcal{Y} \setminus X$ is in the partition j. We also use $i_{\mathtt{OPT}} = \operatorname{argmax}_i \Lambda_i$ and $i_{\max} = \operatorname{argmax}_i \Delta_i$. Then, we have

$$\Lambda_{i_{\max}} \geq \Delta_{i_{\max}} - \varepsilon \geq \Delta_{i_{\mathtt{OPT}}} - \varepsilon \geq \Lambda_{i_{\mathtt{OPT}}} - 2\varepsilon$$

where the first and third inequalities are from the definition of ε, i.e., $|\Lambda_i - \Delta_i| \leq \varepsilon$, and the second inequality holds by the optimality of $i_{\max}$. In addition, when the smallest eigenvalue of L is greater than 1, $\log\det L_X$ is monotone and non-negative (Sharma et al., 2015). To complete the proof, we introduce following approximation guarantee of the greedy algorithm with a 'noise' during the selection (Streeter & Golovin, 2009).

Theorem. (Noisy greedy algorithm) *Suppose a submodular function f defined on ground set $\mathcal{Y}$ is monotone and non-negative. Let $X_0 = \emptyset$ and $X_k = X_{k-1} \cup \{i_{\max}\}$ such that*

$$f(X_{k-1} \cup \{i_{\max}\}) - f(X_{k-1}) \geq \max_{i \in \mathcal{Y}\setminus X_{k-1}} \left(f(X_{k-1}\cup\{i\}) - f(X_{k-1})\right) - \varepsilon_k$$

for some $\varepsilon_k \geq 0$. Then,

$$f(X_k) \geq (1 - 1/e) \max_{X\subseteq\mathcal{Y}, |X|\leq k} f(X) - \sum_{i=1}^{k} \varepsilon_i$$

Theorem 1 is straightforward by substituting 2ε into ε_k. This completes the proof of Theorem 1.

B. Proof of Theorem 2

As we explained in Section 2.3, Chebyshev expansion of $\log x$ in $[\delta, 1-\delta]$ with degree n is defined as $p_n(x)$. This can be written as

$$p_n(x) = \sum_{k=0}^{n} c_k T_k\left(\frac{2}{1-2\delta}x - \frac{1}{1-2\delta}\right) \tag{6}$$

where the coefficient c_k and the k-th Chebyshev polynomial $T_k(x)$ are defined as

$$c_k = \begin{cases} \dfrac{1}{n+1}\displaystyle\sum_{j=0}^{n} f\left(\frac{1-2\delta}{2}x_j + \frac{1}{2}\right) T_0(x_j) & \text{if } k = 0 \\ \dfrac{2}{n+1}\displaystyle\sum_{j=0}^{n} f\left(\frac{1-2\delta}{2}x_j + \frac{1}{2}\right) T_k(x_j) & \text{otherwise} \end{cases} \tag{7}$$

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \qquad \text{for } k \geq 1 \tag{8}$$

where $x_j = \cos\left(\frac{\pi(j+1/2)}{n+1}\right)$ for $j = 0, 1, \dots, n$ and $T_0(x) = 1$, $T_1(x) = x$ (Mason & Handscomb, 2002). For simplicity, we now use $H := p_n(A) - p_n(B)$ and denote $\widetilde{A} = \frac{2}{1-2\delta}A - \frac{1}{1-2\delta}\mathbf{I}$ where $\mathbf{I}$ is identity matrix with same dimension of A and same for $\widetilde{B}$.

We estimate the log-determinant difference while random vectors are shared, i.e.,

$$\log\det A - \log\det B \approx \frac{1}{m}\sum_{i=1}^{m} \mathbf{v}^{(i)\top} H \mathbf{v}^{(i)}.$$

To show that the variance of $\mathbf{v}^{(i)\top} H \mathbf{v}^{(i)}$ is small as $\|A - B\|_F$, we provide that

$$\begin{aligned} \mathbf{Var}\left[\frac{1}{m}\sum_{i=1}^{m} \mathbf{v}^{(i)\top} H \mathbf{v}^{(i)}\right] &= \frac{1}{m}\mathbf{Var}\left[\mathbf{v}^\top H \mathbf{v}\right] \\ &\leq \frac{2}{m}\|H\|_F^2 = \frac{2}{m}\|p_n(A) - p_n(B)\|_F^2 \\ &\leq \frac{2}{m}\left(\sum_{k=0}^{n} |c_k| \left\| T_k\left(\widetilde{A}\right) - T_k\left(\widetilde{B}\right)\right\|_F\right)^2 \end{aligned}$$

where the first inequality holds from (Avron & Toledo, 2011) and the second is from combining (6) with the triangle inequality. To complete the proof, we use following two lemmas.

Lemma 3. *Let $T_k(\cdot)$ be Chebyshev polynomial with k-degree and symmetric matrices B, E satisfied with $\|B\|_2 \leq 1$, $\|B+E\|_2 \leq 1$. Then, for $k \geq 0$,*

$$\|T_k(B+E) - T_k(B)\|_F \leq k^2 \|E\|_F.$$

Lemma 4. *Let c_k be the k-th coefficient of Chebyshev expansion for $f(x)$. Suppose f is analytic with $|f(z)| \leq M$ in the region bounded by the ellipse with foci ± 1 and the length of major and minor semiaxis summing to $\rho > 1$. Then,*

$$\sum_{k=0}^{n} k^2 |c_k| \leq \frac{2M\rho(\rho+1)}{(\rho-1)^3}.$$

In order to apply Lemma 4, we should consider $f(x) = \log\left(\frac{1-2\delta}{2}x + \frac{1}{2}\right)$. Then it can be easily obtained $M = 5\log(2/\delta)$ and $\rho = 1 + \frac{2}{\sqrt{2/\delta - 1} - 1}$ as provided in (Han et al., 2015).

Using Lemma 3 and 4, we can write

$$\begin{aligned}
&\mathbf{Var}\left[\frac{1}{m}\sum_{i=1}^{m} \mathbf{v}^{(i)\top} H \mathbf{v}^{(i)}\right] \\
&\leq \frac{2}{m}\left(\sum_{k=0}^{n} |c_k| \left\|T_k\left(\widetilde{A}\right) - T_k\left(\widetilde{B}\right)\right\|_F\right)^2 \\
&\leq \frac{2}{m}\left(\sum_{k=0}^{n} |c_k| k^2 \left\|\widetilde{A} - \widetilde{B}\right\|_F\right)^2 \\
&\leq \frac{2}{m}\left(\frac{2M\rho(\rho+1)}{(\rho-1)^3}\right)^2 \left(\frac{2}{1-2\delta}\|A-B\|_F\right)^2 \\
&= \frac{32M^2\rho^2(\rho+1)^2}{m(\rho-1)^6(1-2\delta)^2}\|A-B\|_F^2
\end{aligned}$$

where the second inequality holds from Lemma 3 and the thrid is from Lemma 4. This completes the proof of Theorem 2.

B.1. Proof of Lemma 3

Denote $R_k := T_k(B+E) - T_k(B)$. From the recurrence of Chebyshev polynomial (8), R_k has following

$$R_{k+1} = 2(B+E)R_k - R_{k-1} + 2E\,T_k(B) \tag{9}$$

for $k \geq 1$ where $R_1 = E$, $R_0 = \mathbf{0}$ where $\mathbf{0}$ is defined as zero matrix with the same dimension of B. Solving this, we obtain that

$$R_{k+1} = g_{k+1}(B+E)E + \sum_{i=0}^{k} h_i(B+E)E\,T_{k+1-i}(B) \tag{10}$$

for $k \geq 1$ where both $g_k(\cdot)$ and $h_k(\cdot)$ are polynomials with degree k and they have following recurrences

$$\begin{aligned}
g_{k+1}(x) &= 2xg_k(x) - g_{k-1}(x),\, g_1(x) = 1,\, g_0(x) = 0, \\
h_{k+1}(x) &= 2xh_k(x) - h_{k-1}(x),\, h_1(x) = 2,\, h_0(x) = 0.
\end{aligned}$$

In addition, we can easily verify that

$$2\max_{x\in[-1,1]}|g_k(x)| = \max_{x\in[-1,1]}|h_k(x)| = 2k.$$

Putting all together, we conclude that

$$\begin{aligned}
\|R_{k+1}\|_F &\leq \|g_{k+1}(B+E)E\|_F \\
&\quad + \left\|\sum_{i=0}^{k} h_i(B+E)E\,T_{k+1-i}(B)\right\|_F \\
&\leq \|g_{k+1}(B+E)\|_2\|E\|_F \\
&\quad + \sum_{i=0}^{k}\|h_i(B+E)\|_2\|E\|_F\|\,T_{k+1-i}(B)\|_2 \\
&\leq \left(\|g_{k+1}(B+E)\|_2 + \sum_{i=0}^{k}\|h_i(B+E)\|_2\right)\|E\|_F \\
&\leq \left(k+1+\sum_{i=0}^{k}2i\right)\|E\|_F \\
&= (k+1)^2\|E\|_F
\end{aligned}$$

where the second inequality holds from $\|YX\|_F = \|XY\|_F \leq \|X\|_2\|Y\|_F$ for matrix X, Y and the third inequality uses that $|T_k(x)| \leq 1$ for all $k \geq 0$. This completes the proof of Lemma 3.

B.2. Proof of Lemma 4

For general analytic function f, Chebyshev series of f is defined as

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k T_k(x), \quad a_k = \frac{2}{\pi}\int_{-1}^{1}\frac{f(x)T_k(x)}{\sqrt{1-x^2}}dx.$$

and from (Mason & Handscomb, 2002) it is known that

$$c_k - a_k = \sum_{j=1}^{\infty}(-1)^j\left(a_{2j(n+1)-k} + a_{2j(n+1)+k}\right)$$

and $|a_k| \leq \frac{2M}{\rho^k}$ for $0 \leq k \leq n$. We remind that c_k is defined in (7). Using this facts, we get

$$\begin{aligned}
k^2 |c_k| &\leq k^2 \left(|a_k| + \sum_{j=1}^{\infty} \left|a_{2j(n+1)-k}\right| + \left|a_{2j(n+1)+k}\right| \right) \\
&\leq k^2 |a_k| + \sum_{j=1}^{\infty} k^2 \left|a_{2j(n+1)-k}\right| + k^2 \left|a_{2j(n+1)+k}\right| \\
&\leq k^2 |a_k| + \sum_{j=1}^{\infty} (2j(n+1)-k)^2 \left|a_{2j(n+1)-k}\right| \\
&\qquad + (2j(n+1)+k)^2 \left|a_{2j(n+1)+k}\right|
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\sum_{k=0}^{n} k^2 |c_k| &\leq \sum_{k=0}^{n} k^2 |a_k| + \sum_{k=n+1}^{\infty} k^2 |a_k| \\
&\leq \sum_{k=0}^{\infty} k^2 |a_k| \leq \sum_{k=0}^{\infty} k^2 \frac{2M}{\rho^k} = \frac{2M\rho(\rho+1)}{(\rho-1)^3}
\end{aligned}$$

This completes the proof of Lemma 4.

References

Avron, Haim and Toledo, Sivan. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2): 8, 2011.

Bird, Steven. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72. Association for Computational Linguistics, 2006.

Boutsidis, Christos, Drineas, Petros, Kambadur, Prabhanjan, and Zouzias, Anastasios. A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix. *arXiv preprint arXiv:1503.00374*, 2015.

Buchbinder, Niv, Feldman, Moran, Seffi, Joseph, and Schwartz, Roy. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, 44(5):1384–1402, 2015.

Daley, Daryl J and Vere-Jones, David. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

De Avila, Sandra Eliza Fontes, Lopes, Ana Paula Brandão, da Luz, Antonio, and de Albuquerque Araújo, Arnaldo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

Feige, Uriel, Mirrokni, Vahab S, and Vondrak, Jan. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

Gillenwater, Jennifer, Kulesza, Alex, and Taskar, Ben. Near-optimal map inference for determinantal point processes. In *Advances in Neural Information Processing Systems*, pp. 2735–2743, 2012.

Gong, Boqing, Chao, Wei-Lun, Grauman, Kristen, and Sha, Fei. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, pp. 2069–2077, 2014.

Greenbaum, Anne. *Iterative methods for solving linear systems*. SIAM, 1997.

Han, Insu, Malioutov, Dmitry, and Shin, Jinwoo. Large-scale log-determinant computation through stochastic chebyshev expansions. In *ICML*, pp. 908–917, 2015.

Hausmann, Dirk, Korte, Bernhard, and Jenkyns, TA. Worst case analysis of greedy type algorithms for independence systems. In *Combinatorial Optimization*, pp. 120–131. Springer, 1980.

Hutchinson, Michael F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.

Johansson, Kurt. Course 1 random matrices and determinantal processes. *Les Houches*, 83:1–56, 2006.

Jordan, Michael Irwin. *Learning in graphical models*, volume 89. Springer Science & Business Media, 1998.

Kang, Byungkon. Fast determinantal point process sampling with application to clustering. In *Advances in Neural Information Processing Systems*, pp. 2319–2327, 2013.

Kathuria, Tarun and Deshpande, Amit. On sampling and greedy map inference of constrained determinantal point processes. *arXiv preprint arXiv:1607.01551*, 2016.

Krause, Andreas, Singh, Ajit, and Guestrin, Carlos. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284, 2008.

Kulesza, Alex and Taskar, Ben. Learning determinantal point processes. In *In Proceedings of UAI*. Citeseer, 2011.

Kulesza, Alex, Taskar, Ben, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Kumar, Ravi, Moseley, Benjamin, Vassilvitskii, Sergei, and Vattani, Andrea. Fast greedy algorithms in mapreduce and streaming. *ACM Transactions on Parallel Computing*, 2(3):14, 2015.

Li, Chengtao, Jegelka, Stefanie, and Sra, Suvrit. Efficient sampling for k-determinantal point processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 1328–1337, 2016a.

Li, Chengtao, Sra, Suvrit, and Jegelka, Stefanie. Gaussian quadrature for matrix inverse forms with applications. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1766–1775, 2016b.

Liu, Yajing, Zhang, Zhenliang, Chong, Edwin KP, and Pezeshki, Ali. Performance bounds for the k-batch greedy strategy in optimization problems with curvature. In *American Control Conference (ACC), 2016*, pp. 7177–7182. IEEE, 2016.

Macchi, Odile. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(01): 83–122, 1975.

Mason, John C and Handscomb, David C. *Chebyshev polynomials*. CRC Press, 2002.

Minoux, Michel. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, pp. 234–243. Springer, 1978.

Mirzasoleiman, Baharan, Badanidiyuru, Ashwinkumar, Karbasi, Amin, Vondrák, Jan, and Krause, Andreas. Lazier than lazy greedy. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Nemhauser, George L, Wolsey, Laurence A, and Fisher, Marshall L. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

Ouellette, Diane Valerie. Schur complements and statistics. *Linear Algebra and its Applications*, 36:187–295, 1981.

Pan, Xinghao, Jegelka, Stefanie, Gonzalez, Joseph E, Bradley, Joseph K, and Jordan, Michael I. Parallel double greedy submodular maximization. In *Advances in Neural Information Processing Systems*, pp. 118–126, 2014.

Peng, Wei and Wang, Hongxia. Large-scale log-determinant computation via weighted l_2 polynomial approximation with prior distribution of eigenvalues. In *International Conference on High Performance Computing and Applications*, pp. 120–125. Springer, 2015.

Saad, Yousef. *Iterative methods for sparse linear systems*. SIAM, 2003.

Sharma, Dravyansh, Kapoor, Ashish, and Deshpande, Amit. On greedy maximization of entropy. In *ICML*, pp. 1330–1338, 2015.

Streeter, Matthew and Golovin, Daniel. An online algorithm for maximizing submodular functions. In *Advances in Neural Information Processing Systems*, pp. 1577–1584, 2009.

Yao, Jin-ge, Fan, Feifan, Zhao, Wayne Xin, Wan, Xiaojun, Chang, Edward, and Xiao, Jianguo. Tweet timeline generation with determinantal point processes. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 3080–3086. AAAI Press, 2016.

Zhang, Martin J and Ou, Zhijian. Block-wise map inference for determinantal point processes with application to change-point detection. In *Statistical Signal Processing Workshop (SSP), 2016 IEEE*, pp. 1–5. IEEE, 2016.