Supplementary File: Fast Deterministic CUR Matrix Decomposition with Accuracy Assurance

Yasutoshi Ida [1 2] **Sekitoshi Kanai** [1] **Yasuhiro Fujiwara** [3] **Tomoharu Iwata** [3]
Koh Takeuchi [2 4] **Hisashi Kashima** [2 4]

A. Proof of Lemma 1

Proof *From subgradients at optimality (Sra et al., 2011), $\boldsymbol{W}_{(i)} = \boldsymbol{0}$ is an optimal solution if and only if the following condition holds for the objective in problem (1):*

$$-\boldsymbol{z}_i + \lambda \boldsymbol{v}_i = \boldsymbol{0}, \tag{A.1}$$

where $\boldsymbol{v}_i$ is an element of the subdifferential of $\|\boldsymbol{W}_{(i)}\|_2$. The subdifferential for the $l2$-norm is represented as $\partial\|\boldsymbol{W}_{(i)}\|_2 = \{\boldsymbol{v}_i \in \mathbb{R}^{1\times p} | \|\boldsymbol{v}_i\|_2 \leq 1\}$ if $\boldsymbol{W}_{(i)} = \boldsymbol{0}$ (Sra et al., 2011). Therefore, we obtain the condition $K_i \leq \lambda$ in Lemma 1 by using Equation (A.1) and the condition $\|\boldsymbol{v}_i\|_2 \leq 1$. □

We note that Lemma 1 is the known result from the optimality condition with the subdifferential (see (Hastie et al., 2015; Yuan & Lin, 2006; Simon et al., 2013; Friedman et al., 2010; Sra et al., 2011) in detail).

B. Proof of Lemma 2

Proof *From Equation (4) and $\|\boldsymbol{X}^{(i)}\|_2 = 1$, we obtain*

$$\boldsymbol{z}_i = \boldsymbol{G}_{(i)} - \boldsymbol{G}_{(i)}\boldsymbol{W} + \boldsymbol{W}_{(i)}. \tag{B.1}$$

If $\tilde{\boldsymbol{z}}_i := \boldsymbol{G}_{(i)} - \boldsymbol{G}_{(i)}\tilde{\boldsymbol{W}} + \tilde{\boldsymbol{W}_{(i)}}$ is $\boldsymbol{z}_i$ before entering the inner loop of the coordinate descent, Equation (B.1) is transformed into the following form:

$$\begin{aligned}\boldsymbol{z}_i &= \boldsymbol{G}_{(i)} - \boldsymbol{G}_{(i)}\tilde{\boldsymbol{W}} + \tilde{\boldsymbol{W}}_{(i)} - \boldsymbol{G}_{(i)}\Delta\boldsymbol{W} + \Delta\boldsymbol{W}_{(i)} \\ &= \tilde{\boldsymbol{z}}_i - \boldsymbol{G}_{(i)}\Delta\boldsymbol{W} + \Delta\boldsymbol{W}_{(i)}.\end{aligned} \tag{B.2}$$

From the aforementioned equation and the triangle equality, we obtain the following inequality:

$$\|\boldsymbol{z}_i\|_2 \leq \|\tilde{\boldsymbol{z}}_i\|_2 + \|\Delta\boldsymbol{W}_{(i)}\|_2 + \|\boldsymbol{G}_{(i)}\Delta\boldsymbol{W}\|_2. \tag{B.3}$$

For the term $\|\boldsymbol{G}_{(i)}\Delta\boldsymbol{W}\|_2$, we obtain the following inequality by using the Cauchy-–Schwarz inequality:

$$\|\boldsymbol{G}_{(i)}\Delta\boldsymbol{W}\|_2 \leq \|\boldsymbol{G}_{(i)}\|_2\|\Delta\boldsymbol{W}\|_F. \tag{B.4}$$

From Equation (B.3) and (B.4), we obtain the following upper bound in the lemma:

$$K_i \leq \tilde{K}_i + \|\Delta\boldsymbol{W}_{(i)}\|_2 + \|\boldsymbol{G}_{(i)}\|_2\|\Delta\boldsymbol{W}\|_F = \overline{K}_i. \tag{B.5}$$

We obtain $\|\boldsymbol{z}_i\|_2 \geq \|\tilde{\boldsymbol{z}}_i\|_2 - \|\Delta\boldsymbol{W}_{(i)}\|_2 - \|\boldsymbol{G}_{(i)}\Delta\boldsymbol{W}\|_2$ for the lower bound by using Equation (B.2) and the triangle inequality, similar to that in the case of the upper bound. From the inequality and Equation (B.4), we obtain the lower bound in the lemma:

$$K_i \geq \tilde{K}_i - \|\Delta\boldsymbol{W}_{(i)}\|_2 - \|\boldsymbol{G}_{(i)}\|_2\|\Delta\boldsymbol{W}\|_F = \underline{K}_i, \tag{B.6}$$

which completes the proof. □

C. Proof of Lemma 3

Proof *From Lemma 2, we have $\underline{K}_i \leq K_i \leq \overline{K}_i$. Therefore, the error bound of the upper bound is $|\overline{K}_i - K_i| \leq |\overline{K}_i - \underline{K}_i| = 2\|\Delta\boldsymbol{W}_{(i)}\|_2 + 2\|\boldsymbol{G}_{(i)}\|_2\|\Delta\boldsymbol{W}\|_F = \epsilon$. Similar to that for the upper bound, we obtain $|\underline{K}_i - K_i| \leq |\underline{K}_i - \overline{K}_i| = \epsilon$ for the lower bound.* □

D. Proof of Lemma 4

Proof *When $\overline{K}_i \leq \lambda$ holds, we have $K_i \leq \overline{K}_i \leq \lambda$ from Lemma 2. Therefore, we have $\boldsymbol{W}_{(i)} = \boldsymbol{0}$ from Lemma 1 as $K_i \leq \lambda$ holds.* □

E. Proof of Lemma 5

Proof *For the term $\|\boldsymbol{G}_{(i)}\|_2\|\Delta\boldsymbol{W}\|_F$ in Equation (5), since we have $\|\Delta\boldsymbol{W}\|_F^2 = \|[\|\Delta\boldsymbol{W}^{(1)}\|_2, ..., \|\Delta\boldsymbol{W}^{(p)}\|_2]\|_2^2 = \|\Delta\boldsymbol{W}^{(1)}\|_2^2 + ... + \|\Delta\boldsymbol{W}^{(p)}\|_2^2$, we can update $\|\Delta\boldsymbol{W}\|_F^2$ by using the following equation:*

$$\|\Delta\boldsymbol{W}\|_F^2 - \|\Delta\boldsymbol{W}_{(j)}\|_2^2 + \|\Delta\boldsymbol{W}'_{(j)}\|_2^2 = \delta^2 \tag{E.1}$$

Therefore, we obtain Equation (7) by using δ instead of $\|\Delta\boldsymbol{W}\|_F$ in Equation (5). □

[1]NTT Software Innovation Center, Tokyo, Japan [2]Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan [3]NTT Communication Science Laboratories, Kyoto, Japan [4]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. Correspondence to: Yasutoshi Ida <yasutoshi.ida@ieee.org>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

F. Proof of Lemma 6

Proof *We can precompute $\tilde{K}_i$ and $\|\boldsymbol{G}_{(i)}\|_2$ before entering the inner loop and outer loop, respectively. In addition, we have $\|\Delta\boldsymbol{W}_{(i)}\|_2$ and $\|\Delta\boldsymbol{W}\|_F$ as scalars. Thus, we obtain the terms $\tilde{K}_i$, $\|\boldsymbol{G}_{(i)}\|_2$, $\|\Delta\boldsymbol{W}_{(i)}\|_2$, and $\|\Delta\boldsymbol{W}\|_F$ at $\mathcal{O}(1)$ times. When $\Delta\boldsymbol{W}_{(j)}$ is updated to $\Delta\boldsymbol{W}'_{(j)}$, the computation of $\|\Delta\boldsymbol{W}'_{(j)}\|_2$ in Equation (8) requires $\mathcal{O}(p)$ time. Therefore, the total computation cost of Equation (7) is $\mathcal{O}(p)$ time.* □

G. Proof of Lemma 7

Proof *If $\boldsymbol{W}$ converges, we have $\Delta\boldsymbol{W}_{(i)} = \boldsymbol{0}$ and $\Delta\boldsymbol{W} = \boldsymbol{0}$. Since the error bound ϵ is $2\|\Delta\boldsymbol{W}_{(i)}\|_2 + 2\|\boldsymbol{G}_{(i)}\|_2\|\Delta\boldsymbol{W}\|_F$ in Lemma 3, we obtain $\epsilon = 0$. In addition, because we have $\tilde{K}_i = K_i$ when $\boldsymbol{W}$ converges, the upper bound $\overline{K}_i$ and the lower bound $\underline{K}_i$ converge to the condition score K_i from Equations (5) and (6).* □

H. Proof of Lemma 8

Proof *When $\underline{K}_i > \lambda$ holds, we have $K_i \geq \underline{K}_i > \lambda$ from Lemma 2. Therefore, since $K_i > \lambda$ holds, we have $\boldsymbol{W}_{(i)} \neq \boldsymbol{0}$ from Lemma 1.* □

I. Proof of Lemma 9

Proof *Lemma 9 holds since the set $\mathbb{M}$ includes indices of rows that must be nonzero vectors from Lemma 8.* □

J. Proof of Lemma 10

Proof *Since $\overline{K}_i \geq \underline{K}_i$ from Lemma 2, the lower bound is computed as $\underline{K}_i = \overline{K}_i - \epsilon$ from the proof of Lemma 3. Since we have $\epsilon = |\overline{K}_i - \underline{K}_i| = 2\|\Delta\boldsymbol{W}_{(i)}\|_2 + 2\delta\|\boldsymbol{G}_{(i)}\|_2$ after $\boldsymbol{W}_{(j)}$ is updated to $\boldsymbol{W}'_{(j)}$, we obtain Equation (10).* □

K. Proof of Lemma 11

Proof *The terms $\tilde{K}_i$, $\|\Delta\boldsymbol{W}_{(i)}\|_2$, $\delta = \|\Delta\boldsymbol{W}\|_F$, and $\|\boldsymbol{G}_{(i)}\|_2$ in Equation (6) have already been computed in Equations (7) and (8). Since we obtain these terms at $\mathcal{O}(1)$ times, the computation cost of Equation (8) is $\mathcal{O}(1)$ time.* □

L. Proof of Lemma 12

Proof *Lemma 12 holds from Lemma 11 since the computation of the construction for the set $\mathbb{M}$ checks $\underline{K}_i$, which requires $\mathcal{O}(1)$ time, for p rows.* □

M. Proof of Lemma 13

Proof *Suppose that $L(\boldsymbol{W}) := \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}\|_F^2$ in problem (11). By following the property of the overlapping norm (Jacob et al., 2009), $\boldsymbol{W}$ is an optimal solution of problem (11) if and only if the following conditions hold for any rows and columns: (i) $\boldsymbol{W}$ can be decomposed as $\boldsymbol{W} = \boldsymbol{V} + \boldsymbol{H}$, (ii) if $\boldsymbol{V}_{(i)} = \boldsymbol{0}$ then $\|\nabla_i L(\boldsymbol{W})\|_2 \leq \lambda_r$, (iii) if $\boldsymbol{V}_{(i)} \neq \boldsymbol{0}$ then $\nabla_i L(\boldsymbol{W}) = -\lambda_r \boldsymbol{V}_{(i)}/\|\boldsymbol{V}_{(i)}\|_2$, (iv) if $\boldsymbol{H}^{(j)} = \boldsymbol{0}$ then $\|\nabla^j L(\boldsymbol{W})\|_2 \leq \lambda_c$, and (v) if $\boldsymbol{H}^{(j)} \neq \boldsymbol{0}$ then $\nabla^j L(\boldsymbol{W}) = -\lambda_c \boldsymbol{H}^{(j)}/\|\boldsymbol{H}^{(j)}\|_2$, where $\nabla_i L(\cdot)$ and $\nabla^j L(\cdot)$ are the partial gradients of $L(\cdot)$ with respect to the parameters in i-th row and j-th column, respectively. We consider conditions (ii) and (iv) since Lemma 13 handles the condition for zero parameters. By using the covariate duplication method (Obozinski et al., 2011), if $\boldsymbol{V}_{(i)} = \boldsymbol{0}$, $\nabla_i L(\boldsymbol{W})$ in condition (ii) is computed as follows:*

$$\nabla_i L(\boldsymbol{W}) = \boldsymbol{X}^{(i)\mathrm{T}}\{\boldsymbol{X} - (\boldsymbol{X}\boldsymbol{W} - \boldsymbol{X}^{(i)}\boldsymbol{V}_{(i)})\boldsymbol{X}\}\boldsymbol{X}^{\mathrm{T}}. \quad \text{(M.1)}$$

Therefore, we obtain the condition $R_i \leq \lambda_r$ for $\boldsymbol{V}_{(i)} = \boldsymbol{0}$ in Lemma 13 from $\|\nabla_i L(\boldsymbol{W})\|_2 = R_i$ and condition (ii). Similarly, if $\boldsymbol{H}^{(j)} = \boldsymbol{0}$, $\nabla^j L(\boldsymbol{W})$ in condition (iv) is computed as follows:

$$\nabla^j L(\boldsymbol{W}) = \boldsymbol{X}^{\mathrm{T}}\{\boldsymbol{X} - \boldsymbol{X}(\boldsymbol{W}\boldsymbol{X} - \boldsymbol{H}^{(j)}\boldsymbol{X}_{(j)})\}\boldsymbol{X}_{(j)}^{\mathrm{T}}. \quad \text{(M.2)}$$

From $\|\nabla^j L(\boldsymbol{W})\|_2 = C_j$ and condition (iv), we obtain the condition $C_j \leq \lambda_c$ in Lemma 13. □

We can solve problem (11) by combining Lemma 13 and coordinate descent with the covariate duplication method (Obozinski et al., 2011). We note that the columns $\mathbf{C}$ and the rows $\mathbf{R}$ are selected on the basis of the indices corresponding to the nonzero rows in $\mathbf{V}$ and nonzero columns in $\mathbf{H}$, respectively. Namely, if $\mathcal{I} \subseteq \{1, ..., p\}$ and $\mathcal{J} \subseteq \{1, ..., n\}$ are the indices corresponding to the nonzero rows in $\mathbf{V}$ and nonzero columns in $\mathbf{H}$ respectively, we obtain $\mathbf{X}^{\mathcal{I}}$ and $\mathbf{X}_{\mathcal{J}}$ as the columns $\mathbf{C}$ and the rows $\mathbf{R}$, respectively.

Similar to the proof of Lemma 2, we obtain the upper and lower bounds of R_i and C_j in Definition 3 from Lemma 13. We note that the assumption of $\|\mathbf{X}^{(i)}\|_2 = 1$ in Lemma 2 is not required for Definition 3.

N. Proof of Theorem 1

Proof *The precomputations of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and $\|\boldsymbol{G}_{(i)}\|_2$ for all the rows require $\mathcal{O}(p^2 n)$ and $\mathcal{O}(p^2)$ times, respectively. Since the lower bound $\underline{K}_i$ is computed at $\mathcal{O}(1)$ time from Lemma 11, the construction of the set $\mathbb{M}$ requires $\mathcal{O}(p)$ time as shown in Lemma 12. According to the sequential rule, the total cost of the construction is $\mathcal{O}(pQ)$ time. Since computation of all the rows of $\tilde{K}_i$ requires $\mathcal{O}(p^2 n)$ time, the total cost is $\mathcal{O}(p^2 n t_u)$ time for all the outer loops of the coordinate descent with the upper bounds. From Lemma 6, the total computation cost of the upper bounds $\overline{K}_i$ is $\mathcal{O}(p^2 t_u)$ because $\mathcal{O}(p^2)$ time is required to compute the upper bounds of all the rows. To update the parameter vectors,*

we need $\mathcal{O}(pnt_m)$ time for the coordinate descent on the set $\mathbb{M}$. For the coordinate descent with the upper bounds, $\mathcal{O}(p^2nt_uS)$ time is required for the updates. This is because the total number of inner loops is pt_u, and the updates are performed only when they are unskipped by the upper bound. Thus, Algorithm 2 needs $\mathcal{O}(p\{n(t_m+pt_uS)+Q\})$ time. □

O. Proof of Theorem 2

Proof *Since Algorithm 2 preferentially updates the rows that must be nonzero vectors (lines 5–13), the updating order is different from that of the original algorithm. On the other hand, after lines 5–13, Algorithm 2 performs coordinate descent, which updates the parameter vectors in a cyclic order the same as the original algorithm (lines 14–25). In other words, Algorithm 2 performs cyclic coordinate descent the same as the original algorithm with different initial parameters. In addition, from Lemma 4, Algorithm 2 safely skips the computations of the cyclic coordinate descent (lines 20–21). Since we assume that the cyclic coordinate descent converges in the theorem and that problem (1) is a convex optimization problem (Bien et al., 2010), Algorithm 2 converges to the same objective values as those of the original algorithm if their regularization constants are the same.* □

References

Bien, J., Xu, Y., and Mahoney, M. W. CUR from a Sparse Optimization Viewpoint. In *NeurIPS*, pp. 217–225, 2010.

Friedman, J., Hastie, T., and Tibshirani, R. A Note on The Group Lasso and a Sparse Group Lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

Jacob, L., Obozinski, G., and Vert, J. Group Lasso with Overlap and Graph Lasso. In *ICML*, pp. 433–440, 2009.

Obozinski, G., Jacob, L., and Vert, J.-P. Group Lasso with Overlaps: the Latent Group Lasso approach. Research report, 2011.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Sra, S., Nowozin, S., and Wright, S. J. *Optimization for Machine Learning*. The MIT Press, 2011.

Yuan, M. and Lin, Y. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the royal statistical society, series B*, 68:49–67, 2006.