Supplementary Material for the paper: Efficient Optimization For Low-Rank Integrated Bilinear Classifiers

Takumi Kobayashi and Nobuyuki Otsu

National Institute of Advanced Industrial Science and Technology,
1-1-1, Umezono, Tsukuba, Japan
{takumi.kobayashi,otsu.n}@aist.go.jp

In this supplementary material, we prove the convexity of the proposed formulation,

$$\min_{\boldsymbol{\Sigma}_w \succcurlyeq 0} \left[J(\boldsymbol{\Sigma}_w) \triangleq \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_w) + \sum_i \alpha_i^{*}(\boldsymbol{\Sigma}_w) - \frac{1}{2}\sum_{i,j} \alpha_i^{*}(\boldsymbol{\Sigma}_w)\alpha_j^{*}(\boldsymbol{\Sigma}_w) y_i y_j K_{ij(\boldsymbol{\Sigma}_w)} \right], \quad (1)$$

where $K_{ij(\boldsymbol{\Sigma}_w)} = \mathrm{tr}(\boldsymbol{\Sigma}_w \boldsymbol{X}_i^\top \boldsymbol{X}_j)$,

$$\boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w) = \arg\max_{\boldsymbol{\alpha}\in\Theta_{\boldsymbol{y}}} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j K_{ij(\boldsymbol{\Sigma}_w)}, \quad (2)$$

$$\Theta_{\boldsymbol{y}} = \left\{ \boldsymbol{\alpha} \,|\, \forall i,\ 0 \le \alpha_i \le C,\ \textstyle\sum_i y_i\alpha_i = 0 \right\},$$

which appears as (10) in the paper. Since the constraint $\boldsymbol{\Sigma}_w \succcurlyeq 0$, *i.e.*, positive semidefinite cone, is a convex set, the convexity of the proposed optimization problem (1) will be established if the objective cost function J is proven to be convex.

We use the following notations:

$$\boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w) \in \mathbb{R}^n := [\alpha_1^{*}(\boldsymbol{\Sigma}_w), \cdots, \alpha_n^{*}(\boldsymbol{\Sigma}_w)]^\top,$$
$$\bar{\boldsymbol{K}}_{(\boldsymbol{\Sigma}_w)} \in \mathbb{R}^{n\times n} : \bar{K}_{ij(\boldsymbol{\Sigma}_w)} = y_i y_j K_{ij(\boldsymbol{\Sigma}_w)},$$

and then the objective cost function is written by

$$J(\boldsymbol{\Sigma}_w) = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_w) + \mathbf{1}^\top \boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w) - \frac{1}{2}\boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w)^\top \bar{\boldsymbol{K}}_{(\boldsymbol{\Sigma}_w)} \boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w). \quad (3)$$

1 Derivative of J

Before proceeding to the proof of the convexity, we first show the form of the derivative of J with respect to $\boldsymbol{\Sigma}_w$ based on Lemma 2 in [1].

Let σ_{ij} denote the i,j-th component of $\boldsymbol{\Sigma}_w$, and the derivative of J with respect to σ_{ij} is simply obtained by

$$\frac{\partial J}{\partial \sigma_{ij}} = \frac{1}{2}\delta_{ij} + \mathbf{1}^\top \frac{\partial \boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}} - \frac{1}{2}\boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w)^\top \frac{\partial \bar{\boldsymbol{K}}(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}} \boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w) - \boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w)^\top \bar{\boldsymbol{K}}_{(\boldsymbol{\Sigma}_w)} \frac{\partial \boldsymbol{\alpha}^{*}(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}, \quad (4)$$

where δ_{ij} is the Kronecker delta and note that $\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)$ is the function of $\boldsymbol{\Sigma}_w$.

The Lagrangian function for (2) is given by

$$L(\boldsymbol{\alpha}, \lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top \bar{\boldsymbol{K}}(\boldsymbol{\Sigma}_w)\boldsymbol{\alpha} - \lambda \boldsymbol{y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \boldsymbol{\alpha} + \boldsymbol{\gamma}^\top (C\mathbf{1} - \boldsymbol{\alpha}), \tag{5}$$

where $\lambda \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^n_+$ and $\boldsymbol{\gamma} \in \mathbb{R}^n_+$ are the Lagrange multipliers for the constraints $\Theta_{\boldsymbol{y}}$. Thus, the unique optimizer $\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)$ in (2) satisfies the following conditions:

$$\begin{aligned} \mathbf{1} - \bar{\boldsymbol{K}}(\boldsymbol{\Sigma}_w)\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w) - \lambda^*(\boldsymbol{\Sigma}_w)\boldsymbol{y} + \boldsymbol{\beta}^*(\boldsymbol{\Sigma}_w) - \boldsymbol{\gamma}^*(\boldsymbol{\Sigma}_w) = \mathbf{0} \\ \Rightarrow \mathbf{1} - \bar{\boldsymbol{K}}(\boldsymbol{\Sigma}_w)\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w) = \lambda^*(\boldsymbol{\Sigma}_w)\boldsymbol{y} - \boldsymbol{\beta}^*(\boldsymbol{\Sigma}_w) + \boldsymbol{\gamma}^*(\boldsymbol{\Sigma}_w), \end{aligned} \tag{6}$$

$$\boldsymbol{y}^\top \boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w) = 0 \Rightarrow \boldsymbol{y}^\top \frac{\partial \boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}} = \mathbf{0}, \tag{7}$$

$$\forall l, \beta_l^*(\boldsymbol{\Sigma}_w)\alpha_l^*(\boldsymbol{\Sigma}_w) = 0 \Rightarrow \beta_l^*(\boldsymbol{\Sigma}_w)\frac{\partial \alpha_l^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}} + \frac{\partial \beta_l^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}\alpha_l^*(\boldsymbol{\Sigma}_w) = 0, \tag{8}$$

$$\forall l, \gamma_l^*(\boldsymbol{\Sigma}_w)\{C - \alpha_l^*(\boldsymbol{\Sigma}_w)\} = 0 \Rightarrow -\gamma_l^*(\boldsymbol{\Sigma}_w)\frac{\partial \alpha_l^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}} + \frac{\partial \gamma_l^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}\{C - \alpha_l^*(\boldsymbol{\Sigma}_w)\} = 0, \tag{9}$$

where $\lambda^*(\boldsymbol{\Sigma}_w), \boldsymbol{\beta}^*(\boldsymbol{\Sigma}_w)$ and $\boldsymbol{\gamma}^*(\boldsymbol{\Sigma}_w)$ are the optimum Lagrange multipliers, depending on the variable $\boldsymbol{\Sigma}_w$. By using the fact that either $\beta_l^*(\boldsymbol{\Sigma}_w) = 0$ or $\alpha_l^*(\boldsymbol{\Sigma}_w) = 0$ in (8) and either $\gamma_l^*(\boldsymbol{\Sigma}_w) = 0$ or $\alpha_l^*(\boldsymbol{\Sigma}_w) = C$ in (9), we can further obtain the following conditions from (8) and (9):

$$\beta_l^*(\boldsymbol{\Sigma}_w)\frac{\partial \alpha_l^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}} = -\frac{\partial \beta_l^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}\alpha_l^*(\boldsymbol{\Sigma}_w) = 0, \tag{10}$$

$$\gamma_l^*(\boldsymbol{\Sigma}_w)\frac{\partial \alpha_l^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}} = \frac{\partial \gamma_l^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}(C - \alpha_l^*(\boldsymbol{\Sigma}_w)) = 0. \tag{11}$$

Thus, the derivative (4) results in

$$\frac{\partial J}{\partial \sigma_{ij}} = \frac{1}{2}\delta_{ij} - \frac{1}{2}\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)^\top \frac{\partial \bar{\boldsymbol{K}}(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w) + \frac{\partial \boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}^\top \left\{\mathbf{1} - \bar{\boldsymbol{K}}(\boldsymbol{\Sigma}_w)\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)\right\} \tag{12}$$

$$= \frac{1}{2}\delta_{ij} - \frac{1}{2}\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)^\top \frac{\partial \bar{\boldsymbol{K}}(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w) + \frac{\partial \boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}^\top \{\lambda^*(\boldsymbol{\Sigma}_w)\boldsymbol{y} - \boldsymbol{\beta}^*(\boldsymbol{\Sigma}_w) + \boldsymbol{\gamma}^*(\boldsymbol{\Sigma}_w)\} \tag{13}$$

$$= \frac{1}{2}\delta_{ij} - \frac{1}{2}\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w)^\top \frac{\partial \bar{\boldsymbol{K}}(\boldsymbol{\Sigma}_w)}{\partial \sigma_{ij}}\boldsymbol{\alpha}^*(\boldsymbol{\Sigma}_w), \tag{14}$$

where we use (6) for transforming (12) into (13), and use (7), (10) and (11) to get (14) from (13).

Finally, the derivative of J with respect to $\boldsymbol{\Sigma}_w$ is given by

$$\frac{\partial J}{\partial \boldsymbol{\Sigma}_w} = \frac{1}{2}\boldsymbol{I} - \frac{1}{2}\sum_{i,j}\alpha_i^*(\boldsymbol{\Sigma}_w)\alpha_j^*(\boldsymbol{\Sigma}_w)\frac{\partial \bar{K}_{ij}(\boldsymbol{\Sigma}_w)}{\partial \boldsymbol{\Sigma}_w} \tag{15}$$

$$= \frac{1}{2}\Big\{\boldsymbol{I} - \sum_{i,j}\alpha_i^*(\boldsymbol{\Sigma}_w)\alpha_j^*(\boldsymbol{\Sigma}_w)y_i y_j \boldsymbol{X}_i^\top \boldsymbol{X}_j\Big\}. \tag{16}$$

2 Convexity of J

We prove the convexity of J by verifying the following first order condition:

$$J(\boldsymbol{B}) \geq J(\boldsymbol{A}) + \mathrm{tr}\left\{(\boldsymbol{B}-\boldsymbol{A})\frac{\partial J}{\partial \boldsymbol{\Sigma}_w}\bigg|_{\boldsymbol{A}}\right\}, \tag{17}$$

where $\boldsymbol{A} \succcurlyeq 0,\ \boldsymbol{B} \succcurlyeq 0$.

Proof: From (3), the left-hand side in (17) is written by

$$J(\boldsymbol{B}) = \frac{1}{2}\mathrm{tr}(\boldsymbol{B}) + \mathbf{1}^\top \boldsymbol{\alpha}^*_{(\boldsymbol{B})} - \frac{1}{2}\boldsymbol{\alpha}^*_{(\boldsymbol{B})}{}^\top \bar{\boldsymbol{K}}_{(\boldsymbol{B})}\boldsymbol{\alpha}^*_{(\boldsymbol{B})}, \tag{18}$$

while by using (16), the right-hand side in (17) results in

$$\begin{aligned}
&J(\boldsymbol{A}) + \mathrm{tr}\left\{(\boldsymbol{B}-\boldsymbol{A})\frac{\partial J}{\partial \boldsymbol{\Sigma}_w}\bigg|_{\boldsymbol{A}}\right\} \\
&= \frac{1}{2}\mathrm{tr}(\boldsymbol{A}) + \mathbf{1}^\top \boldsymbol{\alpha}^*_{(\boldsymbol{A})} - \frac{1}{2}\boldsymbol{\alpha}^*_{(\boldsymbol{A})}{}^\top \bar{\boldsymbol{K}}_{(\boldsymbol{A})}\boldsymbol{\alpha}^*_{(\boldsymbol{A})} \\
&\quad + \frac{1}{2}\mathrm{tr}(\boldsymbol{B}-\boldsymbol{A}) - \frac{1}{2}\sum_{ij}\alpha^*_{i(\boldsymbol{A})}\alpha^*_{j(\boldsymbol{A})}y_i y_j \mathrm{tr}\big\{(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{X}_i^\top \boldsymbol{X}_j\big\} \\
&= \frac{1}{2}\mathrm{tr}(\boldsymbol{B}) + \mathbf{1}^\top \boldsymbol{\alpha}^*_{(\boldsymbol{A})} - \frac{1}{2}\boldsymbol{\alpha}^*_{(\boldsymbol{A})}{}^\top \bar{\boldsymbol{K}}_{(\boldsymbol{B})}\boldsymbol{\alpha}^*_{(\boldsymbol{A})}.
\end{aligned} \tag{19}$$

Since $\boldsymbol{\alpha}^*_{(\boldsymbol{B})} = \arg\max_{\boldsymbol{\alpha}\in\Theta_{\boldsymbol{y}}}\{\mathbf{1}^\top\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top\bar{\boldsymbol{K}}_{(\boldsymbol{B})}\boldsymbol{\alpha}\}$ and $\boldsymbol{\alpha}^*_{(\boldsymbol{A})} \in \Theta_{\boldsymbol{y}}$, it is shown that

$$\mathbf{1}^\top \boldsymbol{\alpha}^*_{(\boldsymbol{B})} - \frac{1}{2}\boldsymbol{\alpha}^*_{(\boldsymbol{B})}{}^\top \bar{\boldsymbol{K}}_{(\boldsymbol{B})}\boldsymbol{\alpha}^*_{(\boldsymbol{B})} \geq \mathbf{1}^\top \boldsymbol{\alpha}^*_{(\boldsymbol{A})} - \frac{1}{2}\boldsymbol{\alpha}^*_{(\boldsymbol{A})}{}^\top \bar{\boldsymbol{K}}_{(\boldsymbol{B})}\boldsymbol{\alpha}^*_{(\boldsymbol{A})}. \tag{20}$$

From (18–20), the inequality (17) holds. □

References

1. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. Machine Learning **46** (2002) 131–159