# Supplementary of "WPML$^3$CP: Wasserstein Partial Multi-Label Learning with Dual Label Correlation Perspectives"

## A Optimization of Regularized Wasserstein Distance

In this section, we present the optimization details of regularized Wasserstein distance:

$$W_\lambda\left(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbf{M}_\mathcal{K}\right) = \inf_{\mathbf{T}\in U(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{T}, \mathbf{M}_\mathcal{K}\rangle - \frac{1}{\lambda} H(\mathbf{T}). \tag{1}$$

We convert the above regularized Wasserstein distance into its dual problem, and solve them as well as optimize the parameters $\{\boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{M}_\mathcal{K}\}$ by employing the Sinkhorn's algorithm [Cuturi, 2013; Cuturi and Doucet, 2014].

Following [Cuturi and Doucet, 2014], its dual problem can be formulated by:

$${}^dW_\lambda\left(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbf{M}_\mathcal{K}\right) = \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})} \boldsymbol{\alpha}^\top \boldsymbol{\mu} + \boldsymbol{\beta}^\top \boldsymbol{\nu} - \sum_{i,j} \frac{e^{-\lambda\left(\mathbf{M}_\mathcal{K}(i,j) - \alpha_i - \beta_j\right)}}{\lambda}.$$

Then, the regularized Wasserstein distance and its dual problem can be both efficiently solved by the Sinkhorn's algorithm with $O(K^2)$ complexity [Cuturi, 2013; Cuturi and Doucet, 2014]. Thanks to their efficient computations, one can utilize this regularized Wasserstein distance as the loss function under various learning paradigms. Specifically, given models with parameters of interest, *i.e.,* denoted by $\{\boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{M}_\mathcal{K}\}$, we can optimize them by leveraging their subgradients of $W_\lambda\left(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbf{M}_\mathcal{K}\right)$, being equivalent to the optimum $\{\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \mathbf{T}^*\}$ of the dual problem ${}^dW_\lambda\left(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbf{M}_\mathcal{K}\right)$ [Bertsimas and Tsitsiklis, 1997; Cuturi and Doucet, 2014; Frogner *et al.*, 2015]:

$$\frac{\partial W_\lambda(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbf{M}_\mathcal{K})}{\partial \boldsymbol{\mu}} \triangleq \boldsymbol{\alpha}^* = -\frac{\log(\mathbf{a})}{\lambda} + \frac{\log(\mathbf{a})^\top \mathbf{1}}{\lambda K}\mathbf{1}, \tag{2}$$

$$\frac{\partial W_\lambda(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbf{M}_\mathcal{K})}{\partial \boldsymbol{\nu}} \triangleq \boldsymbol{\beta}^* = -\frac{\log(\mathbf{b})}{\lambda} + \frac{\log(\mathbf{b})^\top \mathbf{1}}{\lambda K}\mathbf{1}, \tag{3}$$

$$\frac{\partial W_\lambda(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbf{M}_\mathcal{K})}{\partial \mathbf{M}_\mathcal{K}} \triangleq \mathbf{T}^* = \mathrm{diag}(\mathbf{a})\mathbf{K}\mathrm{diag}(\mathbf{b}), \tag{4}$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^K$ can be computed by solving a matrix balancing problem with the Sinkhorn's algorithm [Cuturi, 2013; Cuturi and Doucet, 2014]:

$$(\mathbf{a}, \mathbf{b}) \leftarrow (\boldsymbol{\mu} \oslash \mathbf{K}\mathbf{b}, \boldsymbol{\nu} \oslash \mathbf{K}^\top \mathbf{a}), \tag{5}$$

$\mathbf{K} = e^{-\lambda \mathbf{M}_\mathcal{K}}$ denotes the element-wise exponential of $-\lambda \mathbf{M}_\mathcal{K}$, and $\oslash$ represents the element-wise division. For clarity, the full computation process of subgradients is summarized in *Algorithm 1*.

**Algorithm 1** Regularized Wasserstein distance's subgradient

**Input:** parameters $\{\boldsymbol{\mu}, \boldsymbol{\nu}\}$, matrix $\mathbf{K} = e^{-\lambda \mathbf{M}_\mathcal{K}}$ and regularization parameter $\lambda > 0$;

**Initialize** $\mathbf{a} = \mathbf{1}$, $\mathbf{b} = \mathbf{1}$;
**while** $\{\mathbf{a}, \mathbf{b}\}$ have not converged **do**
    $(\mathbf{a}, \mathbf{b}) \leftarrow (\boldsymbol{\mu} \oslash \mathbf{K}\mathbf{b}, \boldsymbol{\nu} \oslash \mathbf{K}^\top \mathbf{a})$, *i.e.,* Eq.(5)
**end while**

**Output:** $\frac{\partial W_\lambda(\boldsymbol{\mu},\boldsymbol{\nu};\mathbf{M}_\mathcal{K})}{\partial \boldsymbol{\mu}}$, $\frac{\partial W_\lambda(\boldsymbol{\mu},\boldsymbol{\nu};\mathbf{M}_\mathcal{K})}{\partial \boldsymbol{\nu}}$, $\frac{\partial W_\lambda(\boldsymbol{\mu},\boldsymbol{\nu};\mathbf{M}_\mathcal{K})}{\partial \mathbf{M}_\mathcal{K}}$, *i.e.,* Eqs.(2), (3) and (4)

## B Optimization of WPML$^3$CP

In this section, we describe the optimization details of WPML$^3$CP. We first revisit the objective of WPML$^3$CP:

$$\begin{aligned}
\min_{\mathbf{W},\mathbf{Q},\mathbf{C},\mathbf{E}} \quad & \sum_{i=1}^n W_\lambda\big(\mathfrak{s}(\mathbf{q}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big) \\
& + \frac{\beta_1}{2}\sum_{i=1}^l \sum_{j=1}^l \mathbf{C}_{ij} \|\mathbf{w}_i - \mathbf{w}_j\|_2^2 + \frac{\beta_2}{2}\|\mathbf{W}\|_F^2 \\
& + \beta_3 \|\mathbf{Q}\|_* + \beta_4 \|\mathbf{E}\|_1 \\
\mathrm{s.t.} \quad & \mathbf{Y} = \mathbf{Q} + \mathbf{E}.
\end{aligned} \tag{6}$$

By employing the LADMAP method [Lin *et al.*, 2011] over its augmented Lagrangian, we reformulate the optimization problem in Eq.(6) as follows:

$$\begin{aligned}
\min_{\mathbf{W},\mathbf{Q},\mathbf{C},\mathbf{E},\mathbf{H}} \quad & \sum_{i=1}^n W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big) + \frac{\beta_1}{2}\mathrm{tr}(\mathbf{W}^\top \mathbf{L}\mathbf{W}) \\
& + \frac{\beta_2}{2}\|\mathbf{W}\|_F^2 + \beta_3\|\mathbf{Q}\|_* + \beta_4\|\mathbf{E}\|_1 \\
& + \frac{\mu_1}{2}\|\mathbf{Y} - \mathbf{Q} - \mathbf{E} + \frac{\mathbf{Y}_1}{\mu_1}\|_F^2 \\
& + \frac{\mu_2}{2}\|\mathbf{Q} - \mathbf{H} + \frac{\mathbf{Y}_2}{\mu_2}\|_F^2,
\end{aligned} \tag{7}$$

where $\mathbf{L} = \mathrm{diag}(\mathbf{C}\mathbf{1}) - \mathbf{C}$ is the Laplacian matrix of $\mathbf{C}$. Accordingly, we employ the gradient decent approach to optimize $\{\mathbf{W}, \mathbf{C}, \mathbf{H}\}$, whose gradients can be easily calculated with some simple derivations and the Sinkhorn algorithm in *Algorithm* 1, and update $\{\mathbf{Q}, \mathbf{E}\}$ as well as $\{\mathbf{Y}_1, \mathbf{Y}_2, \mu_1, \mu_2\}$

with the linear ADM method following [Liu *et al.*, 2010a]. Details are described in the following part.

**Updating W**: Fixing $\{\mathbf{Q}, \mathbf{C}, \mathbf{E}, \mathbf{H}\}$ as constants, the subproblem of Eq.(7) with respect to $\mathbf{W}$ can be compactly formulated as follows:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big) + \frac{\beta_1}{2}\mathrm{tr}(\mathbf{W}^\top \mathbf{L}\mathbf{W}) + \frac{\beta_2}{2}\|\mathbf{W}\|_F^2. \tag{8}$$

After some simple derivations, the gradient of $\mathbf{W}$ can be computed based on the chain rule:

$$g(\mathbf{W}) = \sum_{i=1}^{n}\Big(g\big(\mathfrak{s}(\mathbf{W}\mathbf{x}_i)\big) \times \frac{\partial \mathfrak{s}(\mathbf{W}\mathbf{x}_i)}{\partial(\mathbf{W}\mathbf{x}_i)}\Big)\mathbf{x}_i^\top + \frac{\beta_1}{2}(\mathbf{L}\mathbf{W} + \mathbf{L}^\top\mathbf{W}) + \beta_2\mathbf{W}, \tag{9}$$

where

$$g\big(\mathfrak{s}(\mathbf{W}\mathbf{x}_i)\big) = \frac{\partial W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big)}{\partial \mathfrak{s}(\mathbf{W}\mathbf{x}_i)}.$$

Then, $\mathbf{W}$ can be updated with the gradient decent method as:

$$\mathbf{W} \leftarrow \mathbf{W} - \rho_t g(\mathbf{W}). \tag{10}$$

**Updating H**: When keeping $\{\mathbf{W}, \mathbf{Q}, \mathbf{C}, \mathbf{E}\}$ fixed, the subproblem of Eq.(7) with respect to $\mathbf{H}$ is given by:

$$\min_{\mathbf{H}} \sum_{i=1}^{n} W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big) + \frac{\mu_2}{2}\|\mathbf{Q} - \mathbf{H} + \frac{\mathbf{Y}_2}{\mu_2}\|_F^2. \tag{11}$$

With some simple derivations, we can compute the gradient of $\mathbf{H}$ by leveraging the chain rule:

$$g(\mathbf{H}) = \sum_{i=1}^{n} g\big(\mathfrak{s}(\mathbf{h}_i)\big) \times \frac{\partial \mathfrak{s}(\mathbf{h}_i)}{\partial \mathbf{h}_i} - \mu_2(\mathbf{Q} - \mathbf{H} + \frac{\mathbf{Y}_2}{\mu_2}), \tag{12}$$

where

$$g\big(\mathfrak{s}(\mathbf{h}_i)\big) = \frac{\partial W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big)}{\partial \mathfrak{s}(\mathbf{h}_i)}.$$

Consequently, we can update $\mathbf{H}$ with:

$$\mathbf{H} \leftarrow \mathbf{H} - \rho_t g(\mathbf{H}). \tag{13}$$

**Updating C**: Fixing $\{\mathbf{W}, \mathbf{Q}, \mathbf{E}, \mathbf{H}\}$ as constants, the subproblem of Eq.(7) with respect to $\mathbf{C}$ can be compactly formulated as follows:

$$\min_{\mathbf{C}} \sum_{i=1}^{n} W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big) + \frac{\beta_1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\mathbf{C}_{ij}\|\mathbf{w}_i - \mathbf{w}_j\|_2^2 \tag{14}$$

After some simple derivations, the gradient of $\mathbf{C}$ can be computed based on the chain rule:

$$g(\mathbf{C}) = g\big(\mathfrak{m}(\mathbf{C})\big) \times \frac{\partial \mathfrak{m}(\mathbf{C})}{\partial \mathbf{C}} + \frac{\beta_1}{2}\mathbf{A}, \tag{15}$$

where

$$g(\mathfrak{m}(\mathbf{C})) = \sum_{i=1}^{n}\mathbf{T}_i^*,$$

$\mathbf{T}_i^*$ is the optimal transport plan of $W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big)$, and $\mathbf{A} \in \mathbb{R}^{l\times l}$ is defined by $\mathbf{A}_{ij} = \|\mathbf{w}_i - \mathbf{w}_j\|_2^2$. Then, $\mathbf{C}$ can be updated with the gradient decent method as follows:

$$\mathbf{C} \leftarrow \mathbf{C} - \rho_t g(\mathbf{C}). \tag{16}$$

**Updating $\{\mathbf{Q}, \mathbf{E}\}$**: Holding $\{\mathbf{W}, \mathbf{C}, \mathbf{H}\}$ fixed, the subproblem of Eq.(7) with respect to $\{\mathbf{Q}, \mathbf{E}\}$ can be rewritten as follows:

$$\min_{\mathbf{Q},\mathbf{E}} \beta_3\|\mathbf{Q}\|_* + \beta_4\|\mathbf{E}\|_1 + \frac{\mu_1}{2}\|\mathbf{Y} - \mathbf{Q} - \mathbf{E} + \frac{\mathbf{Y}_1}{\mu_1}\|_F^2 + \frac{\mu_2}{2}\|\mathbf{Q} - \mathbf{H} + \frac{\mathbf{Y}_2}{\mu_2}\|_F^2. \tag{17}$$

The above optimization problem can be solved by employing a robust PCA (RPCA) technique, and its Lineared Alternating Direction Method (LADM) solution is given by:

$$\mathbf{Q}^{k+1} = \mathcal{D}_{1/\beta_{\mathbf{Q}}}\left[\mathbf{Q}^k - \frac{\mathbf{F}_{\mathbf{Q}}^k}{\beta_{\mathbf{Q}}}\right], \tag{18}$$

$$\mathbf{E}^{k+1} = \mathcal{S}_{\beta_4/\mu_1}\left[\mathbf{Y} - \mathbf{Q}^{k+1} + \frac{\mathbf{Y}_1^k}{\mu_1^k}\right], \tag{19}$$

where $\mathcal{D}_{1/\beta_{\mathbf{Q}}}(\cdot)$ is the singular value thresholding [Liu *et al.*, 2010b], $\mathcal{S}_{\beta_4/\mu_1}(\cdot)$ is the shrinkage operator [Zhuang *et al.*, 2012], $\beta_{\mathbf{Q}} = (\mu_1 + \mu_2)\tau_{\mathbf{Q}}/2$, $\tau_z > \rho(\mathbf{I}^\top\mathbf{I})$ is the proximal parameter, $\tau_{\mathbf{P}} > \rho(\mathbf{I}^\top\mathbf{I})$ denotes the spectral radius of $\mathbf{I}^\top\mathbf{I}$, and $\mathbf{F}_{\mathbf{Q}}^k$ is derivated by $\mathbf{Q}^k$ for the third and fourth terms in Eq.(17):

$$\mathbf{F}_{\mathbf{Q}}^k = \mu_1(\mathbf{Q} - \mathbf{Y} + \mathbf{E}) + \mu_2(\mathbf{Q} - \mathbf{H}) + \mathbf{Y}_2 - \mathbf{Y}_1.$$

**Updating $\{\mathbf{Y}_1, \mathbf{Y}_2, \mu_1, \mu_2\}$**: The Lagrange multiplier matrixes $\{\mathbf{Y}_1, \mathbf{Y}_2\}$ and the corresponding regularization parameters $\{\mu_1, \mu_2\}$ can be updated by utilizing the LADM as follows:

$$\begin{aligned}
\mathbf{Y}_1^{k+1} &\leftarrow \mathbf{Y}_1^k + \mu_1^k(\mathbf{Y} - \mathbf{Q} - \mathbf{E}),\\
\mathbf{Y}_2^{k+1} &\leftarrow \mathbf{Y}_2^k + \mu_2^k(\mathbf{Q} - \mathbf{H}),\\
\mu_1^{k+1} &\leftarrow \min(\mu_{max}, \psi\mu_1^k),\\
\mu_2^{k+1} &\leftarrow \min(\mu_{max}, \psi\mu_2^k),
\end{aligned} \tag{20}$$

where $\psi$ is a positive scalar.

Note that both gradients of Eqs.(9), (12) and (15) can be efficiently calculated. **First**, we can compute the subgradients of regularized Wasserstein distance, *i.e.*, $g\big(\mathfrak{s}(\mathbf{W}\mathbf{x}_i)\big)$, $g\big(\mathfrak{s}(\mathbf{h}_i)\big)$ and $g\big(\mathfrak{m}(\mathbf{C})\big)$, by directly using *Algorithm 1* mentioned in Section A, specifically substituting $\{\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i), \mathfrak{m}(\mathbf{C})\}$ into $\{\boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{M}_{\mathcal{K}}\}$. **Second**, we can directly calculate the two gradients of the softmax function, *i.e.*, $\partial\mathfrak{s}\big(\mathbf{W}\mathbf{x}_i\big)/\partial\big(\mathbf{W}\mathbf{x}_i\big)$ and $\partial\mathfrak{s}\big(\mathbf{h}_i\big)/\partial\mathbf{h}_i$, as well as the gradient of the sigmoid function, *i.e.*, $\partial\mathfrak{m}\big(\mathbf{C}\big)/\partial\mathbf{C}$.

**Algorithm 2** Optimization for WPML$^3$CP

**Input:** Training dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{i=n}$, regularization parameters $\{\beta_1, \beta_2, \beta_3, \beta_4, \lambda\}$; LADM parameters $\{\psi, \mu_{max}\}$;
**Output:** Model parameter $\mathbf{W}^*$.
**Initialize** $\{\mathbf{W}, \mathbf{Q}, \mathbf{E}, \mathbf{H}\}$ and $\{\mathbf{Y}_1, \mathbf{Y}_2, \mu_1, \mu_2\}$;
Calculate the initial pairwise similarity matrix $\mathbf{C}$;
**for** $t = 1$ **to** $N_{iter}$ **do**
  **for** $i = 1$ **to** $n$ **do**
    Calculate $g\big(\mathfrak{s}(\mathbf{h}_i)\big)$, $g\big(\mathfrak{s}(\mathbf{W}\mathbf{x}_i)\big)$ , $\mathbf{T}_i^*$ by *Algorithm 1*;
  **end for**
  Calculate $g(\mathbf{H})$, $g(\mathbf{W})$ and $g(\mathbf{C})$ by Eqs.(12), (9) and (15);
  Update $\{\mathbf{W}, \mathbf{Q}, \mathbf{C}, \mathbf{E}, \mathbf{H}\}$ and $\{\mathbf{Y}_1, \mathbf{Y}_2, \mu_1, \mu_2\}$ by Eqs.(10), (18), (13), (16) and (20);
**end for**

**Full Algorithm**: In summary, we iteratively update parameters $\{\mathbf{W}, \mathbf{Q}, \mathbf{C}, \mathbf{E}, \mathbf{H}\}$ and Lagrange multiplier variables $\{\mathbf{Y}_1, \mathbf{Y}_2, \mu_1, \mu_2\}$. Finally, we can obtain the optimal model parameter $\mathbf{W}^*$ for predicting future instances. For clarity, the full optimization procedure of WPML$^3$CP is summarized in *Algorithm 2*.

## References


[Bertsimas and Tsitsiklis, 1997] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific Belmont, MA, 1997.

[Cuturi and Doucet, 2014] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *ICML*, pages 685–693, 2014.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300, 2013.

[Frogner *et al.*, 2015] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi an Mauricio Araya-Polo, and Tomaso Pogglo. Learning with a wasserstein loss. In *NeurIPS*, pages 2053–2061, 2015.

[Lin *et al.*, 2011] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NeurIPS*, pages 612–620, 2011.

[Liu *et al.*, 2010a] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[Liu *et al.*, 2010b] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[Zhuang *et al.*, 2012] Liansheng Zhuang, Haoyuan Gao, Zhouchen Lin, Yi Ma, Xin Zhang, and Nenghai Yu. Nonnegative low rank and sparse graph for semi-supervised learning. In *IEEE CVPR*, pages 2328–2335, 2012.