Proceedings of Machine Learning Research 80:1–3, 2018 ACML 2018

Supplementary Material: Unsupervised Heterogeneous Domain Adaptation with Sparse Feature Transformation

1. Minimization Over B

Given the current fixed $A^{(k)}$ and $\Lambda^{(k)}$, B can be updated by minimizing the augmented Lagrangian:

$$\begin{aligned} B^{(k+1)} &:= \arg\min_{B} \; L_\rho(A^{(k)}, B, \Lambda^{(k)}) \\ &:= \arg\min_{B} \; \ell(B) + \frac{\gamma}{q} \|B\|_{p,q}^{q} \end{aligned} \tag{10}$$

where the smooth part of function is

$$\ell(B) = \frac{1}{2} \left\| A^{(k)\top} C_s B - C_t \right\|_F^2 + \frac{\alpha}{2} \left\| X_s^0 B - X_t^0 \right\|_F^2 - \mathrm{tr}(\Lambda^{(k)\top} B) + \frac{\rho}{2} \left\| A^{(k)} - B \right\|_F^2$$

This minimization problem is a convex quadratic programing with a non-smooth sparsity regularizer. We solve it using a fast proximal gradient descent method with a quadratic convergence rate (Beck and Teboulle, 2009), which tackles Eq.(10) by solving a sequence of intermediate problems iteratively with proximity operators. The algorithm is given in Algorithm 1 below. The convergence of the algorithm is proved in (Beck and Teboulle, 2009).

Algorithm 1 Fast Proximal Gradient Descent Algorithm

Initialization: $Q^{(1)} = B^{(0)}$=starting point, $\beta_1 = 1$, $t = 0$.
For iter = 1:maxiters
  1. Set $t = t + 1$
  2. Update: $B^{(t)} = \mathcal{P}_\eta(Q^{(t)})$, $\beta_{t+1} = \frac{1+\sqrt{1+4\beta_t^2}}{2}$,
  $Q^{(t+1)} = B^{(t)} + \left(\frac{\beta_t - 1}{\beta_{t+1}}\right)(B^{(t)} - B^{(t-1)})$
End For

For the t-th iteration, the intermediate problem at point $Q^{(t)}$ is in the following form:

$$\mathcal{P}_\eta(Q^{(t)}) = \arg\min_{B} \left\{ \frac{1}{2} \|B - \widehat{Q}^{(t)}\| + \frac{\gamma}{q\eta} \|B\|_{p,q}^{q} \right\} \tag{11}$$

where $\widehat{Q}^{(t)}$ is derived from the gradient of $\ell(Q^{(t)})$ such that

$$\widehat{Q}^{(t)} = Q^{(t)} - \frac{1}{\eta} \nabla \ell(Q^{(t)})$$

and η is the Lipschitz constant of the general gradient function $\nabla\ell(B)$. The gradient can be computed as

$$\nabla\ell(B) = \left(C_s^\top A^{(k)} A^{(k)\top} C_s + \alpha X_s^{0\top} X_s^0 + \rho I\right) B - \left(C_s^\top A^{(k)} C_t + \alpha X_s^{0\top} X_t^0 + \Lambda^{(k)} + \rho A^{(k)}\right)$$

A Lipschitz constant η of $\nabla\ell(B)$ needs to satisfy the property

$$\|\nabla\ell(B) - \nabla\ell(B')\|_F \leq \eta \|B - B'\|_F, \text{ for any feasible } B, B'.$$

Lemma 1 *Let*

$$\eta = \sigma_{\max}\left(C_s^\top A^{(k)} A^{(k)\top} C_s + \alpha X_s^{0\top} X_s^0 + \rho I\right),$$

where $\sigma_{\max}(\cdot)$ denotes the largest singular value of the corresponding matrix. Then η is a Lipschitz constant of $\nabla\ell(B)$.

Proof Let $H = C_s^\top A^{(k)} A^{(k)\top} C_s + \alpha X_s^{0\top} X_s^0 + \rho I$. We have the following derivations

$$\begin{aligned}
&\|\nabla\ell(B) - \nabla\ell(B')\|_F \\
=&\left\|(C_s^\top A^{(k)} A^{(k)\top} C_s + \alpha X_s^{0\top} X_s^0 + \rho I)(B - B')\right\|_F \\
=&\left\|H(B - B')\right\|_F \\
=&\Big(\sum_j \|H(B_{:j} - B'_{:j})\|_2^2\Big)^{1/2} \\
\leq&\Big(\|H\|_2^2 \sum_j \|B_{:j} - B'_{:j}\|_2^2\Big)^{1/2} \qquad \text{(since spectral norm is induced by the Euclidean norm)} \\
=&\left\|H\|_2 \|B - B'\right\|_F \\
=&\sigma_{\max}(H)\|B - B'\|_F
\end{aligned}$$

where $\|\cdot\|_2$ denotes the spectral norm of the corresponding matrix or the Euclidean norm of a vector; $B_{:j}$ denotes the j-th column of matrix B. ■

The nice property about the intermediate problem in Eq.(11) is that it allows us to exploit closed-form solutions for the proximity operator $\mathcal{P}_\eta(Q^{(t)})$ with either the ℓ_1-norm regularizer ($p = 1$ and $q = 1$) or the $\ell_{1,2}$-norm regularizer ($p = 1$ and $q = 2$). According to (Kowalski et al., 2009), we have the following closed-form solution for the proximity operations:

If $p = 1$ and $q = 1$ (ℓ_1-norm), we have

$$\mathcal{P}_\eta(Q^{(t)}) = sign(\widehat{Q}^{(t)}) \circ \left(|\widehat{Q}^{(t)}| - \frac{\gamma}{\eta}\right)_+$$

where $(\cdot)_+ = \max(0, \cdot)$ and $\circ$ denotes the entrywise Hadamard product operator.

If $p = 1$ and $q = 2$ ($\ell_{1,2}$-norm), we have

$$\mathcal{P}_\eta(Q^{(t)}) = \tilde{Q}$$

such that

$$\tilde{Q}_{i,j} = sign(\widehat{Q}^{(t)}_{i,j}) \left(|\widehat{Q}^{(t)}_{i,j}| - \frac{\gamma \sum_{r=1}^{m_j} \overrightarrow{Q}_{r,j}}{(\eta + \gamma m_j) \|\widehat{Q}^{(t)}_{:j}\|_2} \right)_+$$

where $\overrightarrow{Q}_{:j}$ denotes a reordered j-th column $|\widehat{Q}^{(t)}_{:j}|$ with a descending order of the entries, and the corresponding m_j is the number such that

$$\overrightarrow{Q}_{m_j+1,j} \leq \frac{\gamma}{\eta} \sum_{r=1}^{m_j+1} \left(\overrightarrow{Q}_{r,j} - \overrightarrow{Q}_{m_j+1,j} \right)$$
$$\overrightarrow{Q}_{m_j,j} > \frac{\gamma}{\eta} \sum_{r=1}^{m_j} \left(\overrightarrow{Q}_{r,j} - \overrightarrow{Q}_{m_j,j} \right)$$

References

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

M. Kowalski, M. Szafranski, and L. Ralaivola. Multiple indefinite kernel learning with mixed norm regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 545–552. ACM, 2009.