Supplementary Material for "Selective Labeling via Error Bound Minimization"

Quanquan Gu[†], **Tong Zhang**[‡], **Chris Ding**[§], **Jiawei Han**[†]
[†]Department of Computer Science, University of Illinois at Urbana-Champaign
[‡]Department. of Statistics, Rutgers University
[§]Department. of Computer Science & Engineering, University of Texas at Arlington
qgu3@illinois.edu, tzhang@stat.rutgers.edu, chqding@uta.edu, hanj@cs.uiuc.edu

1 Derivation of Eq. (19)

Recall that the Lagrangian function is

$$L(\mathbf{S}) = \mathrm{tr}(\mathbf{X}^T(\mathbf{X}\mathbf{S}\mathbf{S}^T\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}) \tag{1}$$

where we ignore the term $\mathrm{tr}(\mathbf{\Lambda}(\mathbf{S}^T\mathbf{S} - \mathbf{I}))$, whose derivative is trivial.

To compute its derivative with respect to $\mathbf{S}$, we use the definition of derivative.

Let

$$\mathbf{D} = \frac{\partial L}{\partial \mathbf{S}} \tag{2}$$

Let $\mathbf{\Delta}$ be a small perturbation on $\mathbf{S}$. We have

$$\begin{aligned}
& \mathrm{tr}(\mathbf{D}\mathbf{\Delta}^T) \\
= \; & L(\mathbf{S}+\mathbf{\Delta}) - L(\mathbf{S}) \\
= \; & \mathrm{tr}(\mathbf{X}^T(\mathbf{X}(\mathbf{S}+\mathbf{\Delta})(\mathbf{S}+\mathbf{\Delta})^T\mathbf{L}(\mathbf{S}+\mathbf{\Delta})(\mathbf{S}+\mathbf{\Delta})^T\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}) \\
- \; & \mathrm{tr}(\mathbf{X}^T(\mathbf{X}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \mathbf{X}\mathbf{S}\mathbf{S}^T\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}) \\
\approx \; & \mathrm{tr}(\mathbf{X}^T(\mathbf{X}\mathbf{S}\mathbf{S}^T\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \lambda\mathbf{I} + \mathbf{X}(\mathbf{\Delta}\mathbf{S}^T + \mathbf{S}\mathbf{\Delta}^T)\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{L}(\mathbf{S}\mathbf{\Delta}^T + \mathbf{\Delta}\mathbf{S}^T)\mathbf{X}^T)^{-1}\mathbf{X}) \\
- \; & \mathrm{tr}(\mathbf{X}^T(\mathbf{X}\mathbf{S}\mathbf{S}^T\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X})
\end{aligned} \tag{3}$$

where we omit the second-order term of $\mathbf{\Delta}$ because it does not affect the calculation of first-order derivative.

Let

$$\mathbf{A} = \mathbf{X}\mathbf{S}\mathbf{S}^T\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \lambda\mathbf{I} \tag{4}$$

Eq. (3) can be further simplified as

$$\begin{aligned}
& \mathrm{tr}(\mathbf{D}\mathbf{\Delta}^T) \\
= \; & \mathrm{tr}(\mathbf{X}^T(\mathbf{A} + \mathbf{X}(\mathbf{\Delta}\mathbf{S}^T + \mathbf{S}\mathbf{\Delta}^T)\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{L}(\mathbf{S}\mathbf{\Delta}^T + \mathbf{\Delta}\mathbf{S}^T)\mathbf{X}^T)^{-1}\mathbf{X}) \\
- \; & \mathrm{tr}(\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X}) \\
= \; & \mathrm{tr}(\mathbf{X}^T(\mathbf{A}(\mathbf{I} + \mathbf{A}^{-1}\mathbf{X}(\mathbf{\Delta}\mathbf{S}^T + \mathbf{S}\mathbf{\Delta}^T)\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \mathbf{A}^{-1}\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{L}(\mathbf{S}\mathbf{\Delta}^T + \mathbf{\Delta}\mathbf{S}^T)\mathbf{X}^T))^{-1}\mathbf{X}) \\
- \; & \mathrm{tr}(\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X}) \\
= \; & \mathrm{tr}(\mathbf{X}^T(\mathbf{I} + \mathbf{A}^{-1}\mathbf{X}(\mathbf{\Delta}\mathbf{S}^T + \mathbf{S}\mathbf{\Delta}^T)\mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{X}^T + \mathbf{A}^{-1}\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{L}(\mathbf{S}\mathbf{\Delta}^T + \mathbf{\Delta}\mathbf{S}^T)\mathbf{X}^T)^{-1}\mathbf{A}^{-1}\mathbf{X}) \\
- \; & \mathrm{tr}(\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X})
\end{aligned} \tag{5}$$

Since $\mathbf{X}(\mathbf{\Delta S}^T + \mathbf{S\Delta}^T)\mathbf{LSS}^T\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{SS}^T\mathbf{L}(\mathbf{S\Delta}^T + \mathbf{\Delta S}^T)\mathbf{X}^T$ are small, using the first-order taylor expansion $(\mathbf{I}+\mathbf{C})^{-1} = \mathbf{I} - \mathbf{C}$, we have

$$\begin{aligned} & \text{tr}(\mathbf{D\Delta}^T) \\ = \ & \text{tr}(\mathbf{X}^T(\mathbf{I} - \mathbf{A}^{-1}\mathbf{X}(\mathbf{\Delta S}^T + \mathbf{S\Delta}^T)\mathbf{LSS}^T\mathbf{X}^T - \mathbf{A}^{-1}\mathbf{X}^T\mathbf{SS}^T\mathbf{L}(\mathbf{S\Delta}^T + \mathbf{\Delta S}^T)\mathbf{X}^T)\mathbf{A}^{-1}\mathbf{X}) \\ - \ & \text{tr}(\mathbf{X}^T\mathbf{A}^{-1}\mathbf{X}) \\ = \ & -\text{tr}(\mathbf{X}^T\mathbf{A}^{-1}(\mathbf{X}(\mathbf{\Delta S}^T + \mathbf{S\Delta}^T)\mathbf{LSS}^T\mathbf{X}^T + \mathbf{X}^T\mathbf{SS}^T\mathbf{L}(\mathbf{S\Delta}^T + \mathbf{\Delta S}^T)\mathbf{X}^T)\mathbf{A}^{-1}\mathbf{X}) \\ = \ & -\text{tr}(\mathbf{X}^T\mathbf{A}^{-1}(\mathbf{X}(\mathbf{\Delta S}^T + \mathbf{S\Delta}^T)\mathbf{LSS}^T\mathbf{X}^T + \mathbf{X}^T\mathbf{SS}^T\mathbf{L}(\mathbf{S\Delta}^T + \mathbf{\Delta S}^T)\mathbf{X}^T)\mathbf{A}^{-1}\mathbf{X}) \end{aligned} \tag{6}$$

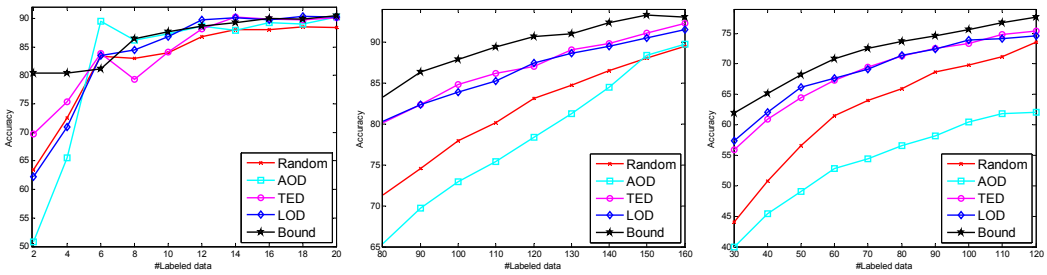
Therefore

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{S}} &= \mathbf{D} \\ &= -2(\mathbf{X}^T\mathbf{BXSS}^T\mathbf{LS} + \mathbf{LSS}^T\mathbf{X}^T\mathbf{BXS}) \end{aligned} \tag{7}$$

where $\mathbf{B} = \mathbf{A}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{A}^{-1}$

2 Additional Experiments

The experimental results using ridge regression (RR) are shown in Figure 1. In all subfigures, the x-axis represents the number of labeled points, while the y-axis is the averaged classification accuracy on the test data over 10 runs.



(a) wdbc (b) ORL (c) Isolet1

Figure 1: Comparison of different methods on (a) wdbc; (b) ORL; and (c) Isolet1 using ridge regression.

We can see that our proposed method is also much better than the other methods using RR.