A deduction for dictionary learning

Yuchen Jin

May 15, 2019

In this article, we would discuss the trick about training and testing phases for dictionary learning (sparse coding). The original work could be referred in [1]. As extra reading materials, we suggest reading [2] for understanding how to apply Lagrangian method and [3] to refer some conclusions about how to calculate gradients for matrices.

1 Solve the Lasso problem

Consider the testing phase of sparse coding which could be formulated as

$$\min_{\{\boldsymbol{\alpha}_i\}_{i=1}^N} \sum_{i=1}^{N} \Big(\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1 \Big). \tag{1-1}$$

Then we could find the stationary point according to the first-order gradient,

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_k} \left(\sum_{i=1}^{N} \Big(\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1 \Big) \right) \\ &\quad = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_k} \Big(\|\mathbf{x}_k - \mathbf{D}\boldsymbol{\alpha}_k\|_2^2 + \lambda \|\boldsymbol{\alpha}_k\|_1 \Big) \\ &\quad = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_k} \big((\mathbf{x}_k - \mathbf{D}\boldsymbol{\alpha}_k)^T (\mathbf{x}_k - \mathbf{D}\boldsymbol{\alpha}_k) \big) + \lambda \mathrm{sign}(\boldsymbol{\alpha}_k) \\ &\quad = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_k} \big(-2\boldsymbol{\alpha}_k^T \mathbf{D}^T \mathbf{x}_k + \boldsymbol{\alpha}_k^T \mathbf{D}^T \mathbf{D}\boldsymbol{\alpha}_k \big) + \lambda \mathrm{sign}(\boldsymbol{\alpha}_k) \\ &\quad = -2\mathbf{D}^T \mathbf{x}_k + 2\mathbf{D}^T \mathbf{D}\boldsymbol{\alpha}_k + \lambda \mathrm{sign}(\boldsymbol{\alpha}_k) = 0. \end{aligned} \tag{2}$$

(2) indicates the analytical solution for Lasso problem implicitly. To find the best $\boldsymbol{\alpha}_k$, we still need to apply some tricks. Denote that $\boldsymbol{\theta}_k = \mathrm{sign}(\boldsymbol{\alpha}_k)$, we could rewrite (2) as

$$\boldsymbol{\alpha}_k = (\mathbf{D}^T\mathbf{D})^{-1}(\mathbf{D}^T\mathbf{x}_k - \frac{\lambda}{2}\boldsymbol{\theta}_k). \tag{3}$$

Since $\boldsymbol{\theta}_k = \mathrm{sign}(\boldsymbol{\alpha}_k)$, we could apply $\mathrm{sign}(\cdot)$ to both sides of (3), then we get

$$\begin{aligned}
\mathrm{sign}\left(\frac{1}{\lambda}\frac{\mathrm{dLasso}}{\mathrm{d}\boldsymbol{\alpha}_k}\right) &= \mathrm{sign}\left((\mathbf{D}^T\mathbf{D})^{-1}(\mathbf{D}^T\mathbf{x}_k - \frac{\lambda}{2}\boldsymbol{\theta}_k)\right) - \boldsymbol{\theta}_k \\
&= \mathrm{sign}\left(\mathbf{D}^T\mathbf{x}_k - \frac{\lambda}{2}\boldsymbol{\theta}_k\right) - \boldsymbol{\theta}_k \\
&= \mathrm{sign}\left(\mathbf{y}_k - \lambda\boldsymbol{\theta}_k\right) - \boldsymbol{\theta}_k = 0.
\end{aligned} \tag{4}$$

(4) indicates a fast solution for (2). It proves that considering that the i^{th} element of $\mathbf{y}_k$, i.e. y_{ki}, we would find that when $y_{ki} > \lambda$, $\alpha_{ki} > 0$, and when $y_{ki} < -\lambda$, $\alpha_{ki} < 0$. Furthermore, when $y_{ki} \in [-\lambda,\ \lambda]$, $\alpha_{ki} = 0$. After confirming $\boldsymbol{\theta}_k$, it will be easy to get $\boldsymbol{\alpha}_k$ from (3) directly.

2 Learn the dictionary

If we rewrite the coding as a matrix as below,

$$\mathbf{A} = \left[\begin{array}{cccc} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \cdots & \boldsymbol{\alpha}_N \end{array}\right], \tag{5}$$

then we could rewrite the dictionary learning problem by Frobenius norm,

$$\min_{\mathbf{D}} \ \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2, \tag{6-1}$$

$$\text{s.t. } \|D(:,k)\|_2 \leqslant 1,\ \forall k \in \{1, 2, \dots, K\}. \tag{6-2}$$

Training dictionary requires us to train $\mathbf{D}$ and $\mathbf{A}$ alternatively. The method for training $\mathbf{A}$ has been discussed before, hence we would discuss how to train $\mathbf{D}$ in the following part. Subsequently, we only note that the trainable variable is $\mathbf{D}$ (6-1), which means (1-1) and (6-1) require to be solved alternatively.

Solving the dictionary training need us to use Lagrangian multiplier method. Denote the multiplier as μ_j, we could incorporate the constraints into the problem,

$$\mathcal{L}(\mathbf{D},\ \boldsymbol{\mu}) = \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \sum_{j=1}^{K} \mu_j \sum_{i=1}^{D} (D_{ij}^2 - 1). \tag{7}$$

The first term of (7) could be expanded as

$$\begin{aligned}
\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 &= \mathrm{Tr}((\mathbf{X} - \mathbf{D}\mathbf{A})(\mathbf{X} - \mathbf{D}\mathbf{A})^T) \\
&= \mathrm{Tr}(\mathbf{X}\mathbf{X}^T) + \mathrm{Tr}(\mathbf{D}\mathbf{A}\mathbf{A}^T\mathbf{D}^T) - 2\mathrm{Tr}(\mathbf{D}\mathbf{A}\mathbf{X}^T).
\end{aligned} \tag{8}$$

Denote a diagonal matrix $\boldsymbol{\Lambda}$ where each element is μ_j, Then the second term could be rewritten as

$$\begin{aligned}
\sum_{j=1}^{K} \mu_j \sum_{i=1}^{D} (D_{ij}^2 - 1) &= \sum_{j=1}^{K} \mu_j \sum_{i=1}^{D} (D_{ij}^2) - \sum_{j=1}^{K} \mu_j \\
&= \mathrm{Tr}(\mathbf{D}\boldsymbol{\Lambda}\mathbf{D}^T - \boldsymbol{\Lambda}).
\end{aligned} \tag{9}$$

Hence we could rewrite (7) as

$$\mathcal{L}(\mathbf{D},\ \mathbf{\Lambda}) = \mathrm{Tr}(\mathbf{X}\mathbf{X}^T + \mathbf{D}\mathbf{A}\mathbf{A}^T\mathbf{D}^T - 2\mathbf{D}\mathbf{A}\mathbf{X}^T + \mathbf{D}\mathbf{\Lambda}\mathbf{D}^T - \mathbf{\Lambda}). \tag{10}$$

Apply the first-order partial gradient to $\mathbf{D}$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\mathbf{D}}\left(\mathcal{L}(\mathbf{D},\ \mathbf{\Lambda})\right) &= \frac{\mathrm{d}}{\mathrm{d}\mathbf{D}}\left(\mathrm{Tr}(\mathbf{D}\mathbf{A}\mathbf{A}^T\mathbf{D}^T - 2\mathbf{D}\mathbf{A}\mathbf{X}^T + \mathbf{D}\mathbf{\Lambda}\mathbf{D}^T\right) \\ &= 2\mathbf{D}\mathbf{A}\mathbf{A}^T - 2\mathbf{X}\mathbf{A}^T + 2\mathbf{D}\mathbf{\Lambda} = 0. \\ \mathbf{D} &= \mathbf{X}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})^{-1}.\end{aligned} \tag{11}$$

Substitute (11) into (10), we would have

$$\begin{aligned}\mathcal{L}(\mathbf{\Lambda}) &= \min_{\mathbf{D}}\ \mathcal{L}(\mathbf{D},\ \mathbf{\Lambda}) \\ &= \mathrm{Tr}(\mathbf{X}\mathbf{X}^T + \mathbf{D}\mathbf{A}\mathbf{A}^T\mathbf{D}^T - 2\mathbf{D}\mathbf{A}\mathbf{X}^T + \mathbf{D}\mathbf{\Lambda}\mathbf{D}^T - \mathbf{\Lambda}) \\ &= \mathrm{Tr}(\mathbf{X}\mathbf{X}^T + \mathbf{D}(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})\mathbf{D}^T - 2\mathbf{D}\mathbf{A}\mathbf{X}^T - \mathbf{\Lambda}) \\ &= \mathrm{Tr}(\mathbf{X}\mathbf{X}^T + \mathbf{X}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})^{-1}(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})^{-1}\mathbf{A}\mathbf{X}^T \\ &\quad - 2\mathbf{X}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})^{-1}\mathbf{A}\mathbf{X}^T - \mathbf{\Lambda}) \\ &= \mathrm{Tr}(\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})^{-1}\mathbf{A}\mathbf{X}^T - \mathbf{\Lambda}).\end{aligned} \tag{12}$$

Note that $\mathbf{D}$ has been represented by $\mathbf{\Lambda}$, we would know that minimizing $\mathcal{L}(\cdot)$ could be applied on $\mathbf{\Lambda}$ solely. Hence we have

$$\begin{aligned}\frac{\partial \min_{\mathbf{D}}\ \mathcal{L}}{\partial \mu_i} &= \mathrm{Tr}\left(\frac{\partial \mathbf{X}\mathbf{X}^T}{\partial \mu_i} - \frac{\partial \mathbf{X}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})^{-1}\mathbf{A}\mathbf{X}^T}{\partial \mu_i} - \frac{\partial \mathbf{\Lambda}}{\partial \mu_i}\right) \\ &= -\mathrm{Tr}\left(\frac{\partial \mathbf{X}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})^{-1}\mathbf{A}\mathbf{X}^T}{\partial \mu_i}\right) - 1.\end{aligned} \tag{13}$$

In [3], there has been a conclusion that

$$\mathrm{Tr}\left(\frac{\partial \mathbf{P}^T(\mathbf{X} + \mathbf{A})^{-1}\mathbf{P}}{\partial x_i}\right) = -\left\|\mathbf{P}^T(\mathbf{X} + \mathbf{A})^{-1}\mathbf{e}_i\right\|_2^2. \tag{14}$$

Apply (14) to (13), we have

$$\frac{\partial \min_{\mathbf{D}}\ \mathcal{L}}{\partial \mu_i} = \left\|\mathbf{X}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \mathbf{\Lambda})^{-1}\mathbf{e}_i\right\|_2^2 - 1 = 0. \tag{15}$$

(15) is in the quadratic form, hence it is convex and we could find the analytical solution for $\mathbf{\Lambda}$. Substitute $\mathbf{\Lambda}$ into (11), we would solve $\mathbf{D}$.

An interesting thing is that if anyone substitute (11) into (15), then it will be

$$\left\|\mathbf{D}\mathbf{e}_i\right\|_2^2 = \left\|D(:, i)\right\|_2 = 1, \tag{16}$$

which shows that the solution revealed in (11) and (15) fulfills the constraints in (6-2) strictly.

References

[1] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in neural information processing systems*, 2007, pp. 801–808.

[2] S. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

[3] K. B. Petersen, M. S. Pedersen *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.