Low Rank Matrix Completion Problem Made Unified with Convex Optimization Theory

J.R

July 11, 2021

In this note, I want to clarify the derivation of the optimization algorithm for solving the low-rank matrix completion problem. Though it's a note on low-rank matrix completion, I'm not going to discuss about the motivation since it's well discussed in a lot of blogs. The main focus of this note is the logic of deriving the optimization algorithm from general convex optimization theory.

1 Singular Value Shrinkage

I have discussed about singular value shrinkage in a previous note, but I think I can make a better explanation this time.

Consider the following optimization problem

$$\min_X \quad \tau||X||_* + \frac{1}{2}||X - Y||_F^2 \tag{1.1}$$

where $||X||_*$ is the nuclear norm(the sum of all singular values) of X. (1.1) has a close-form solution

$$X = \mathcal{D}_\tau(Y) \tag{1.2}$$

where $\mathcal{D}_\tau(\cdot)$ is the singular value shrinkage operator.

If Y is a diagonal matrix, then the singular value shrinkage operator is defined as

$$\mathcal{D}_\tau(Y)_{ii} = \max(Y_{ii} - \tau, 0) \tag{1.3}$$

If Y is not diagonal, then the singular value shrinkage operator is defined as

$$\mathcal{D}_\tau(Y)_{ii} = U\mathcal{D}_\tau(\Sigma)V^T \tag{1.4}$$

where $Y = U\Sigma V^T$ is the singular value decomposition of Y.

To see why $X = \mathcal{D}_\tau(Y)$ solves (1.1), let's see the singular value decomposition of Y. The singular value decomposition of Y can be written as $Y = U_0\Sigma_0V_0^T + U_1\Sigma_1V_1^T$ where the diagonal elements $\text{diag}(\Sigma_0) > \tau$ and $\text{diag}(\Sigma_1) \leq \tau$. The SVD of Y is shown as the following picture.

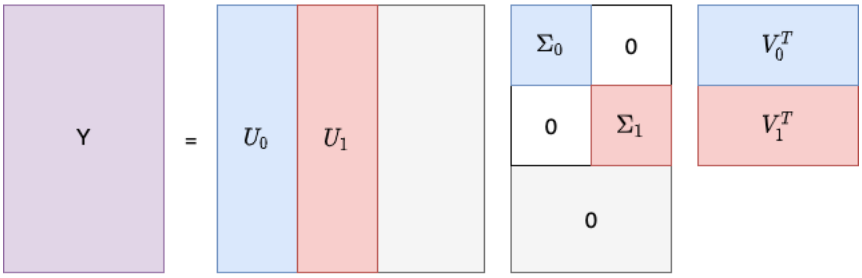

Figure 1: SVD of Y

The τ-singular value shrinkage of Y is

$$\mathcal{D}_\tau(Y) = U_0(\Sigma_0 - \tau I)V_0^T \tag{1.5}$$

and $Y - \mathcal{D}_\tau(Y) = \tau U_0 V_0^T + U_1\Sigma_1 V_1^T = \tau(U_0V_0^T + U_1\Sigma_1/\tau V_1^T)$.

When $X = \mathcal{D}_\tau(Y) = U_0(\Sigma_0 - \tau I)V_0^T + U_1\text{diag}(\vec{0})V_1^T$, notice that diagonal elements of $\Sigma_0 - \tau I$ is positive. the subdifferential of nuclear norm at X is defined as

$$\partial||X||_* = \{U_0V_0^T + U_1\text{diag}(\sigma)V_1^T | -1 \leq \sigma_i \leq 1\} \tag{1.6}$$

We can see that

$$Y - \mathcal{D}_\tau(Y) = Y - X = \tau U_0V_0^T + U_1\Sigma_1V_1^T \in \tau\partial||X||_* \tag{1.7}$$

Thus $\vec{0} \in X - Y + \tau\partial||X||_*$ and $X = \mathcal{D}_\tau(Y)$ solves (1.1).

When you think about the subdifferential of $||X||_*$, since the nuclear norm is the one norm of the singular values of X, the subdifferential of $||X||_*$ is just like the subdifferential of one norm of the singular values of X.

Suppose X has $n1$ nonzero singular values and $n2$ zero singular values, its SVD can be written as

$$X = U_0\Sigma_0V_0^T + U_1\Sigma_1V_1^T \tag{1.8}$$

where diagonal elements of Σ_0 are nonzero and diagonal elements of Σ_1 are zero.

The subgradient of one norm is defined as

$$(\partial||x||_1)i = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ [-1, 1] & x = 0 \end{cases}$$

Similarly, singular vectors corresponding to nonzero singular values Σ_0 contribute $U_0IV_0^T$ to the subgradient and singular vectors corresponding to zero singular values Σ_1 contribute $U_1\text{diag}(\sigma)V_0^T$ to the subgradient where $-1 \leq \sigma_i \leq 1$.

To make it more clear, the subdifferential of nuclear norm at $X = U\Sigma V^T$ is defined as $\partial||X||_* = U\text{diag}(\partial||\sigma||_1)V^T$.

For the detail of subdifferential of matrix norm, please refer to my note `http://lovinglavigne.com/matnorm/MatNormSubdiff.pdf`

2 Low Rank Matrix Completion Problem

The low-rank matrix completion problem of formulated as

$$\begin{aligned} \min_X. \quad & \text{rank}(X) \\ \text{s.t.} \quad & X_{i,j} = M_{i,j} \quad \forall (i,j) \in \Omega \end{aligned} \tag{2.1}$$

Since minimizing the rank is NP hard, instead of minimizing the rank, we can minimize the nuclear norm, the convex envelop of rank$(\cdot)$. The intuition is, minimizing the one norm of singular values yields sparsity of singular values and if there are many zero singular values, the matrix is low-rank.

So we can approximately solve (2.1) by solving the convex relaxation

$$\begin{aligned} \min_X. \quad & ||X||_* \\ \text{s.t.} \quad & P_\Omega(X) = P_\Omega(M) \end{aligned} \tag{2.2}$$

where $P_\Omega(\cdot)$ is the projection on indices in Ω.

Consider the following optimization problem

$$\begin{aligned} \min_X. \quad & \tau||X||_* + \frac{1}{2}||X||_F^2 \\ \text{s.t.} \quad & P_\Omega(X) = P_\Omega(M) \end{aligned} \tag{2.3}$$

If the τ is big enough, then solving (2.3) approximately solves (2.2). Thus, if we want to solve the low-rank matrix completion problem, we can just solve (2.3) with a big enough τ.

3 Lagrangian Multiplier and Gradient Method for Dual

Consider the optimization problem we got

$$\begin{aligned} \min_X. \quad & f(X) = \tau||X||_* + \frac{1}{2}||X||_F^2 \\ \text{s.t.} \quad & P_\Omega(X) = P_\Omega(M) \end{aligned} \tag{3.1}$$

the lagrangian is

$$\mathcal{L}(X,Y) = \tau||X||_* + \frac{1}{2}||X||_F^2 + < Y, P_\Omega(X-M) > \tag{3.2}$$

and the dual function is

$$\begin{aligned} g(Y) &= \inf_X \mathcal{L}(X,Y) \\ &= \inf_X \{\tau||X||_* + \frac{1}{2}||X||_F^2 + < Y, P_\Omega(X-M) >\} \\ &\leq \inf_{P_\Omega(X)=P_\Omega(M)} \{\tau||X||_* + \frac{1}{2}||X||_F^2 + < Y, P_\Omega(X-M) >\} \\ &= \inf_{P_\Omega(X)=P_\Omega(M)} \{\tau||X||_* + \frac{1}{2}||X||_F^2\} \\ &= p^* \end{aligned} \tag{3.3}$$

which is always a lower bound for p^*.

The optimization $\min_X.\tau||X||_* + \frac{1}{2}||X||_F^2 + < Y, P_\Omega(X-M) >$ is equivalent to $\min_X.\tau||X||_* + \frac{1}{2}||X - P_\Omega(Y)||_F^2$ and the close-form solution is given by $X = \mathcal{D}_\tau(Y)$. Thus the dual function can be written as

$$g(Y) = \tau||\hat{X}||_* + \frac{1}{2}||\hat{X}||_F^2 + < Y, P_\Omega(\hat{X}-M) > \tag{3.4}$$

Since problem (3.1) is convex and (strictly) feasible, the strong duality holds. that is

$$\begin{aligned} f(X^*) &= g(Y^*) \\ &= \inf_X L(X, Y^*) \\ &= \inf_X \{f(X) + < Y^*, P_\Omega(X-M) >\} \\ &\leq f(X^*) + < Y^*, P_\Omega(X^* - M) > \\ &= f(X^*) \end{aligned} \tag{3.5}$$

and we get $f(X^*) \leq f(X^*)$, so all inequalities holds with equalities.

We can solve the primal problem by solving the dual problem

$$\max_Y. \quad g(Y) = \tau||\hat{X}||_* + \frac{1}{2}||\hat{X}||_F^2 + < Y, P_\Omega(\hat{X}-M) > \tag{3.6}$$

since the dual function is linear, thus very simple, in Y, we can solve easily solve the dual problem with gradient ascent.

At each step, we calculate the gradient of $g(Y)$ at Y

$$\partial_Y g(Y) = P_\Omega(\hat{X} - M) \tag{3.7}$$

where

$$\hat{X} = \text{argmin}_X \mathcal{L}(X,Y) \tag{3.8}$$

and apply gradient ascent. Thus the updating rule can be describe as

$$\begin{cases} \hat{X} = \mathcal{D}_\tau(Y) \\ \partial_Y g(Y) = P_\Omega(\hat{X} - M) \\ Y := Y + \alpha \partial_Y g(Y) \end{cases}$$

where α is the step size for gradient ascent.

Once we have the updating rule, what we need is a stopping criterion. First, we can produce a upper bound of p^*(denoting the optimal value of (3.1) as p^*) by plugging a feasible X in to $f(X)$. At each step, once we calculated $\hat{X} = \mathcal{D}_\tau(Y)$, we can produce a feasible X by

$$X_{ij} = \begin{cases} \hat{X}_{ij} & (i,j) \notin \Omega \\ M_{ij} & (i,j) \in \Omega \end{cases} \tag{3.9}$$

and we can produce a upper bound for p^* by calculating $f(X)$.

Other than a upper bound of p^*, we need a lower bound of p^* which we can get from evaluating the dual function at the current dual variable Y

$$\begin{aligned} g(Y) &= \inf_X \mathcal{L}(X, Y) \\ &= \mathcal{L}(\mathcal{D}_\tau(Y), Y) \end{aligned}$$

Since we have a upper bound and a lower bound of p^*, once these two bounds meet with each other, we get the optimal value in a range of error. The complete algorithm for solving (3.1) can be described as

```
Algorithm 1: Solving (3.1)

Input:$\tau, M, \epsilon$ ;
Initialization: Dual variable Y=0;
while True do
    Calculate $\hat{X} = \mathcal{D}_\tau(Y)$ ;
    Calculate the gradient $\partial_Y g(Y) = P_\Omega(\hat{X} - M)$ ;
    Apply gradient ascent $Y := Y + \alpha \partial_Y g(Y)$ ;
    Produce a feasible point $X$ by (3.9) ;
    Calculate a upper bound $u = \tau||X||_* + \frac{1}{2}||X||_F^2$ ;
    Calculate a lower bound $l = \mathcal{L}(\mathcal{D}_\tau(Y), Y)$ ;
    if $u - l \leq epsilon$ then
        return $X$;
    else
        pass ;
    end
end
```

The following figure shows the optimization progress of a problem of size 30x30 with 50% of its entries fixed.

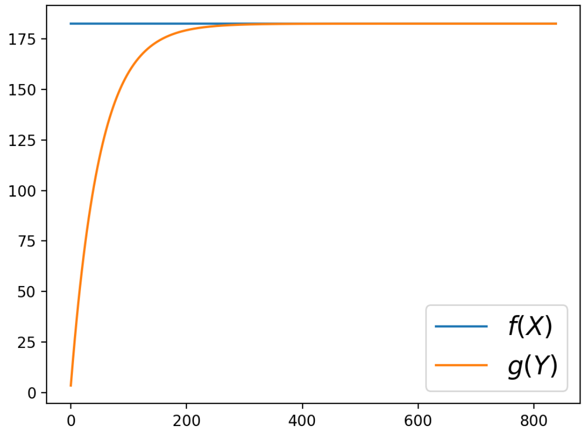

Figure 2: An example of size 30x30 with 50% of its entries

4 Compared with A Simple Problem

The optimization algorithm for solving the low-rank matrix completion problem is nothing more than a gradient method on the dual problem and is very easy to understand. One thing I found when I was learning things is that when things get complicated, you get lost in the intricate derivations of formulas and lose the big picture. Thus I want to compare the low-rank matrix completion problem to a very simple quadratic programming over a unit box to show you the "big picture" of how the optimization algorithm is derived.

Low-rank matrix completion

The optimization problem is

$$\begin{aligned} \min_X. \quad & f(X) = \tau||X||_* + \frac{1}{2}||X||_F^2 \\ \text{s.t.} \quad & P_\Omega(X) = P_\Omega(M) \end{aligned}$$

The lagrangian is

$$\begin{aligned} \mathcal{L}(X,Y) = {} & \tau||X||_* + \frac{1}{2}||X||_F^2 \\ & + < Y, P_\Omega(X) - P_\Omega(M) > \end{aligned}$$

Calculate the X that minimizes the lagrangian at Y by singular value shrinkage

$$X = \mathcal{D}_\tau(Y) = \text{argmin}_X \mathcal{L}(X,Y)$$

The dual function is

$$\begin{aligned} g(Y) &= \inf_X \mathcal{L}(X,Y) \\ &= \tau||\hat{X}||_* + \frac{1}{2}||\hat{X}||_F^2 \\ & \quad + < Y, P_\Omega(\hat{X}) - P_\Omega(M) > \end{aligned}$$

where $\hat{X} = \mathcal{D}_\tau(Y)$
Calculate the gradient of dual function w.r.t Y.

$$\partial_Y g(Y) = P_\Omega(\hat{X} - M)$$

The updating rule is

$$\begin{cases} \hat{X} = \mathcal{D}_\tau(Y) \\ \partial_Y g(Y) = P_\Omega(\hat{X} - M) \\ Y := Y + \alpha \partial_Y g(Y) \end{cases}$$

Produce a feasible point

$$X_{ij} = \begin{cases} \hat{X}_{ij} & (i,j) \notin \Omega \\ M_{ij} & (i,j) \in \Omega \end{cases}$$

Produce a upper bound by evaluating the objective $u = f(X)$ and a lower bound by evaluating the dual function $l = \mathcal{L}(\mathcal{D}_\tau(Y), Y)$.

Quadratic programming over unit box

The optimization problem is

$$\begin{aligned} \min_x. \quad & f(x) = \frac{1}{2}x^TPx + q^Tx \quad (P \geq_{S_{++}} 0) \\ \text{s.t.} \quad & x_i^2 \leq 1 \end{aligned}$$

The lagrangian is

$$\begin{aligned} \mathcal{L}(x,\lambda) = {} & \frac{1}{2}x^TPx + q^Tx \\ & + x^T\text{diag}(\lambda)x - 1^T\lambda \end{aligned}$$

Calculate the X that minimizes the lagrangian at λ by setting gradient to zero.

$$x = [P + 2\text{diag}(\lambda)]^{-1}q = \text{argmin}_x \mathcal{L}(x,\lambda)$$

The dual function is

$$\begin{aligned} g(Y) &= \inf_x \mathcal{L}(x,\lambda) \\ &= \frac{1}{2}\hat{x}^TP\hat{x} + q^T\hat{x} \\ & \quad + \hat{x}^T\text{diag}(\lambda)\hat{x} - 1^T\lambda \end{aligned}$$

where $\hat{x} = [P + 2\text{diag}(\lambda)]^{-1}q$.
Calculate the gradient of dual function w.r.t λ

$$\partial_{\lambda_i} g(\lambda) = \hat{x}_i^2 - 1$$

The updating rule is

$$\begin{cases} \hat{x} = [P + 2\text{diag}(\lambda)]^{-1}q \\ \partial_{\lambda_i} g(\lambda) = \hat{x}_i^2 - 1 \\ \lambda_i := \max(\lambda_i + \alpha\partial_{\lambda_i} g(\lambda), 0) \end{cases}$$

Produce a feasible point

$$\begin{aligned} x_i &= \min(x_i, 1) \\ x_i &= \max(x_i, -1) \end{aligned}$$

Produce a upper bound by evaluating the objective $u = f(x)$ and a lower bound by evaluating the dual function $l = \mathcal{L}([P + 2\text{diag}(\lambda)]^{-1}q, \lambda)$.

As you can see, the left part is a gradient method on dual and the right part is a projected gradient method on dual since λ's should be positive or zero in dual problem. They are basicly the same.