# Optimizing the MMMF Primal

Jason D. M. Rennie
jrennie@gmail.com

March 17, 2005

Let $Y \in \{1, \ldots, l\}^{n \times m}$ be the rating matrix; $Y_{ij}$ is the $i^{\text{th}}$ user's rating of item $j$. Let $S \in \{-1, +1\}^{n \times m}$, $S_{ij}(k) = \begin{cases} +1 & \text{if } k \geq Y_{ij} \\ -1 & \text{if } k < Y_{ij} \end{cases}$. The all-threshold/trace-norm version of MMMF tries to find $X \in \mathbb{R}^{n \times m}$ and $\vec{\theta} \in \mathbb{R}^{n \times l-1}$ (one set of $\theta$'s per row of $X$) so as to minimize

$$J(X) = J(UV^T) = \lambda \|X\|_{\text{tr}} + \sum_{k=1}^{l-1} \sum_{ij} h\Big(S_{ij}(k)(\theta_{ik} - X_{ij})\Big), \tag{1}$$

where $h(z) = (1 - z)_+$ is the hinge loss function[1] and $\|\cdot\|_{\text{tr}}$ is the trace norm,

$$\|X\|_{\text{tr}} = \min_{X = UV'} \frac{1}{2}(\|U\|_{\text{Fro}}^2 + \|V\|_{\text{Fro}}^2), \tag{2}$$

where $\|\cdot\|_{\text{Fro}}$ is the Frobenius norm. Let $p$ be the number of columns of $U$ and $V$; a conservative setting is $p = \max(n, m)$. We can equivalently find $U \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{m \times p}$ and $\theta \in \mathbb{R}^{n \times (l-1)}$ so as to minimize

$$J(U, V) = \frac{\lambda}{2}(\|U\|_{\text{Fro}}^2 + \|V\|_{\text{Fro}}^2) + \sum_{k=1}^{l-1} \sum_{ij} h\Big(S_{ij}(k)(\theta_{ik} - U_i V_j^T)\Big) \tag{3}$$

where $U_i$ is the $i^{\text{th}}$ row of $U$ and $V_j$ is the $j^{\text{th}}$ row of $V$. Note that $J(UV^T) \leq J(U, V)$ and $\min_{UV^T} J(UV^T) = \min_{U,V} J(U, V)$. That is, the second objective is an upper bound for the first objective with the special property that their global minima have the same value (and settings for $U$ and $V$). However, while $J(UV^T)$ is a convex objective (in $UV^T$), $J(U, V)$ is not convex (in $U$,$V$) and we cannot guarantee that gradient descent will not get stuck in local minima.

In practice, we choose random initial values for $U$, $V$ and $\theta$ and use the gradient to find successive improvements. While there is no gurantee we will

[1] Note that $h(\cdot)$ (and $h'(\cdot)$) are a scalar functions: $\mathbb{R} \to \mathbb{R}$. When applied to non-scalar arguments (such as vectors and matrices), they are applied component-wise. I.e. $h'(\vec{z}) = (h'(z_1), h'(z_2), \ldots, h'(z_n))$.

converge to the global minimum, we have had success in practice. The partial derivative for each element of $U$ is

$$\frac{\partial J}{\partial U_{ia}} = \lambda U_{ia} - \sum_{k=1}^{l-1} \sum_{j} S_{ij}(k) h'\Big(S_{ij}(k)(\theta_{ik} - U_i V_j^T)\Big) V_{ja} \tag{4}$$

The partial for $V_{ja}$ is analogous. The partial for each $\theta_{ik}$ is

$$\frac{\partial J}{\partial \theta_{ik}} = \sum_{j} S_{ij}(k) h'\Big(S_{ij}(k)(\theta_{ik} - U_i V_j^T)\Big). \tag{5}$$

Note that the partials lend themselves to matrix notation,

$$\frac{\partial J}{\partial U_i} = \lambda U_i - \sum_{k=1}^{l-1} \sum_{j} S_{ij}(k) h'\Big(S_{ij}(k)(\theta_{ik} - U_i V_j^T)\Big) V_j, \tag{6}$$

$$= \lambda U_i - \sum_{k=1}^{l-1} \Big[S_i(k) * h'\Big(S_i(k) * (\theta_{ik} - U_i V^T)\Big)\Big] V, \tag{7}$$

$$\frac{\partial J}{\partial U} = \lambda U - \sum_{k=1}^{l-1} \Big[S(k) * h'\Big(S(k) * (\theta_{\cdot k} \vec{1}^T - U V^T)\Big)\Big] V. \tag{8}$$

Similarly,

$$\frac{\partial J}{\partial V} = \lambda V - \sum_{k=1}^{l-1} \Big[S(k) * h'\Big(S(k) * (\theta_{\cdot k} \vec{1}^T - U V^T)\Big)\Big]^T U. \tag{9}$$

We can also write the partials with respect to $\theta$ more compactly,

$$\frac{\partial J}{\partial \theta_{\cdot k}} = [S(k) * h'(S(k) * (\theta_{\cdot k} \vec{1}^T - U V^T))] \vec{1}. \tag{10}$$

Finally, we give a compact form of the objective:

$$J = \frac{\lambda}{2}(\|U\|_{\text{Fro}}^2 + \|V\|_{\text{Fro}}^2) + \sum_{k=1}^{l-1} \text{sum}(h(S(k) * (\theta_{\cdot k} \vec{1}^T - U V^T))). \tag{11}$$

We use $*$ to denote element-wise product.