1. The Parallel Stochastic Gradient Method in LIBMF

In LIBMF, we implement a parallel stochastic gradient method to solve RVMF, BMF, as well as OCMF under the framework proposed by Chin et al. (2015a). First, the training matrix is divided into many blocks and some working threads are created. Then, a block scheduler assigns independent blocks to threads that run stochastic gradient method (SG) in parallel. We use the learning rate scheduler in Chin et al. (2015b) and further extend it to cover more loss and regularization functions by following Duchi et al. (2011). In the rest of Section 1, we derive the SG update rules for RVMF and BMF, and the algorithm of OCMF is left in Section 2.

Let $R$ be a training matrix and $P \in \mathbb{R}^{k \times m}$ and $Q \in \mathbb{R}^{k \times n}$ standard for the two factor matrices. Both RVMF and BMF can be formulated as an unconstrained optimization problem.

$$\min_{P, Q} \sum_{(u,v) \in R} \phi_{u,v}(P, Q) + \rho_{u,v}(P, Q), \quad (1)$$

where

$$\phi_{u,v}(P, Q) = f(p_u, q_v; r_{u,v}) + \lambda_p \| p_u \|_2^2 + \lambda_q \| q_v \|_2^2$$

$$\rho_{u,v}(P, Q) = \mu_p \| p_u \|_1 + \mu_q \| q_v \|_1.$$  

In (1), $u$ and $v$ are respectively the row index and the column index of $R$, and $f(\cdot)$ is a non-convex loss function of $p_u$ and $q_v$. In each SG iteration, we sample one entry to construct a subproblem problem combining a first-order approximation of the loss and the L2-regularization, the L1-regularization terms, and the two proximal terms $\| p - p_u \|_2^2$ and $\| q - q_v \|_2^2$. If an entry $(u, v)$ is sampled, the corresponding subproblem is

$$\min_p g_u^T p + \mu_p \| p \|_1 + \frac{\sqrt{G_u}}{2\eta} \| p - p_u \|_2^2$$

$$\min_q h_v^T q + \mu_q \| q \|_1 + \frac{\sqrt{H_v}}{2\eta} \| q - q_v \|_2^2, \quad (2)$$

where

$$g_u = \partial_{p_u} \phi_{u,v} = \partial_{p_u} f + \lambda_p p_u,$$

$$h_v = \partial_{q_v} \phi_{u,v} = \partial_{q_v} f + \lambda_q q_v,$$

and $\eta$ is a pre-specified constant controlling the weight of the proximal terms. Obviously, the existence of the proximal terms encourages a new model close to the current model. See Table 1 for the supported loss functions and their subgradients $\partial_p f$ and $\partial_q f$. The two values $G_u$ and $H_v$ are respectively defined by

$$G_u = \frac{1}{k} \sum_{t=1}^{T_u-1} (g_u^t)^T g_u^t$$

and

$$H_v = \frac{1}{k} \sum_{t=1}^{T_v-1} (h_v^t)^T h_v^t.$$
Table 1: Subgradients of supported loss functions in LIBMF. The first three are for RVMF, and the others are for BMF. Note we denote $p^T q$ by $\hat{r}$.

<table>
<thead>
<tr>
<th>Loss</th>
<th>$f(p, q; r)$</th>
<th>$\frac{\partial f}{\partial p}$</th>
<th>$\frac{\partial f}{\partial q}$</th>
<th>$\kappa$</th>
<th>$\hat{r} - r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2-norm</td>
<td>$(r - \hat{r})^2$</td>
<td>$\kappa q$</td>
<td>$\kappa p$</td>
<td>$\hat{r} - r$</td>
<td></td>
</tr>
<tr>
<td>L1-norm</td>
<td>$</td>
<td>r - \hat{r}</td>
<td>$</td>
<td>$\kappa q$</td>
<td>$\kappa p$</td>
</tr>
<tr>
<td>KL-divergence</td>
<td>$r \log \left( \frac{r}{\hat{r}} \right) - r + \hat{r}$</td>
<td>$\kappa q$</td>
<td>$\kappa p$</td>
<td>$1 - \frac{r}{\hat{r}}$</td>
<td></td>
</tr>
<tr>
<td>Squared hinge</td>
<td>$\max(0, 1 - rr)$</td>
<td>$\kappa q$</td>
<td>$\kappa p$</td>
<td>$-2r \max(0, 1 - r\hat{r})$</td>
<td></td>
</tr>
<tr>
<td>Hinge</td>
<td>$\max(0, 1 - \hat{r})$</td>
<td>$\kappa q$</td>
<td>$\kappa p$</td>
<td>$\left{ \begin{array}{l} -r \text{ if } 1 - r\hat{r} &gt; 0 \ 0 \text{ otherwise} \end{array} \right.$</td>
<td></td>
</tr>
<tr>
<td>Logistic</td>
<td>$\log(1 + \exp(-r\hat{r}))$</td>
<td>$\kappa q$</td>
<td>$\kappa p$</td>
<td>$\frac{-r \exp(-r\hat{r})}{1 + \exp(-r\hat{r})}$</td>
<td></td>
</tr>
</tbody>
</table>

where $T_u$ and $T_v$ are the numbers of updates of $p_u$ and $q_v$, respectively. Notice that $G_u$ and $H_v$ are scalars rather than matrices in [Duchi et al. (2011)]. Moreover, Duchi et al. (2011) use L2-norm instead of squared L2-norm in (2), and they do not provide the update rules considering both L1-regularization and L2-regularization. According to the optimality condition of (2), we can drive the SG update rules

\[
\begin{align*}
(p_u)_i & \leftarrow \text{sign} \left( (p_u)_i - \frac{\eta}{\sqrt{G_u}} (g_u)_i \right) \max \left( 0, |(p_u)_i| - \frac{\eta}{\sqrt{G_u}} (g_u)_i - \frac{\eta}{\sqrt{H_v}} (h_v)_i \right), \\
(q_v)_i & \leftarrow \text{sign} \left( (q_v)_i - \frac{\eta}{\sqrt{H_v}} (h_v)_i \right) \max \left( 0, |(q_v)_i| - \frac{\eta}{\sqrt{H_v}} (h_v)_i - \frac{\eta}{\sqrt{G_u}} (g_u)_i \right),
\end{align*}
\]

which can be reduced to

\[
\begin{align*}
(p_u)_i & \leftarrow (p_u)_i - \frac{\eta}{\sqrt{G_u}} (g_u)_i, \\
(q_v)_i & \leftarrow (q_v)_i - \frac{\eta}{\sqrt{H_v}} (h_v)_i
\end{align*}
\]

if L1-regularization is dropped (i.e., $\mu_p = \mu_q = 0$). The complete procedure of one SG iteration is shown in Algorithm 1. First, some variables are initialized and the value $\kappa$ in Table 1 is computed and cached according to the selected loss function. We then apply (3) on all coordinates of $p_u$ and $q_v$, while $G_u$ and $H_v$ are updated in the end of the iteration. Note that for NMF we add two projections at line 16-19 by following Dror et al. (2012).
Algorithm 1 One SG update for solving RVMF and BMF in LIBMF.

1: $G \leftarrow 0, \quad H \leftarrow 0$
2: $\eta_u \leftarrow \eta_0/\sqrt{G_u}, \quad \eta_v \leftarrow \eta_0/\sqrt{H_v}$
3: $z \leftarrow \kappa(r_{u,v}, p_u, q_v)$
4: for $i \leftarrow 1$ to $k$ do
5: \quad $(g_u)_i \leftarrow \lambda p_u (p_{u})_i + z (q_v)_i$
6: \quad $(h_v)_i \leftarrow \lambda q_v (q_{v})_i + z (p_u)_i$
7: \quad $G \leftarrow G + (g_u)_i^2, \quad H \leftarrow H + (h_v)_i^2$
8: \quad $(p_u)_i \leftarrow (p_u)_i - \eta_u (g_u)_i$
9: \quad $(q_v)_i \leftarrow (q_v)_i - \eta_v (h_v)_i$
10: \quad if $\mu_p > 0$ then
11: \quad \quad $(p_u)_i \leftarrow \text{sign}((p_u)_i) \max(0, |(p_u)_i| - \eta_u \mu_p)$
12: \quad end if
13: \quad if $\mu_q > 0$ then
14: \quad \quad $(q_v)_i \leftarrow \text{sign}((q_v)_i) \max(0, |(q_v)_i| - \eta_v \mu_q)$
15: \quad end if
16: \quad if NMF then
17: \quad \quad $(p_u)_i \leftarrow \max(0, (p_u)_i)$
18: \quad end if
19: \quad end if
20: end for
21: $G_u \leftarrow G_u + G/k, \quad H_v \leftarrow H_v + H/k$

2. One-class Matrix Factorization in LIBMF

Recall the OCMF model (i.e., BPR) in LIBMF. BPR assumes that all positive entries should be ranked above all missing entries (i.e., dummy negative signals), so it is a kind of pairwise ranking problem. Naturally, the pairwise logistic loss can be used to compare the prediction values of a pair of positive and missing entries in the same row of $R$. We call the corresponding optimization problem [4] row-oriented BPR.

$$\min_{P,Q} \sum_{(u,v) \in R} \sum_{(u,w) \notin R_u} \left[ \log(1 + e^{p_u^T (q_w - q_v)}) + \mu_p \|p_u\|_1 + \mu_q (\|q_v\|_1 + \|q_w\|_1) + \frac{\lambda_p}{2} \|p_u\|_2^2 + \frac{\lambda_q}{2} (\|q_v\|_2^2 + \|q_w\|_2^2) \right]$$

(4)

where $R_u$ is the set of positive entries in the $u$th row of $R$. If $u$ is column index and $v$ and $w$ are row indexes, then [4] becomes column-oriented BPR. If a row contains the ratings of a user, row-oriented BPR is suggested to build entry pairs user-wisely. Otherwise, column-oriented BPR should be used. Because we need a pair of entries to construct a subproblem for OCMF, the block scheduler in [Chin et al. (2015a)] is modified so that it returns two blocks in which we can select pairs of positive and a dummy negative entries at random. If a pair of a positive entry $(u,v) \in R$ and a negative entry $(u,w) \notin R$ is sampled, the
Algorithm 2 One SG update for solving OCMF in LIBMF.

1: $G_u \leftarrow 0$, $H_v \leftarrow 0$, $H_w \leftarrow 0$
2: $\eta_u \leftarrow \eta_0/\sqrt{G_u}$, $\eta_v \leftarrow \eta_0/\sqrt{H_v}$, $\eta_{vw} \leftarrow \eta_0/\sqrt{H_w}$
3: $z \leftarrow e^{P_u^T (q_w - q_v)} / (1 + e^{P_u^T (q_w - q_v)})$
4: for $i \leftarrow 1$ to $k$ do
5:   $(g_u)_i \leftarrow \lambda_p (p_u)_i - z((q_w)_i - (q_v)_i)$
6:   $(h_v)_i \leftarrow \lambda_q (q_v)_i - z(p_u)_i$
7:   $(h_w)_i \leftarrow \lambda_q (q_w)_i + z(q_u)_i$
8:   $G_u \leftarrow G_u + (g_u)_i^2$, $H_v \leftarrow H_v + (h_v)_i^2$, $H_w \leftarrow H_w + (h_w)_i^2$
9:   $(p_u)_i \leftarrow (p_u)_i - \eta_u (g_u)_i$
10:  $(q_v)_i \leftarrow (q_v)_i - \eta_v (h_v)_i$
11:  $(q_w)_i \leftarrow (q_w)_i - \eta_{vw} (h_w)_i$
12: if $\mu_p > 0$ then
13:   $(p_u)_i \leftarrow \text{sign}((p_u)_i) \max(0, |(p_u)_i| - \eta_u \mu_p)$
14: end if
15: if $\mu_q > 0$ then
16:   $(q_v)_i \leftarrow \text{sign}((q_v)_i) \max(0, |(q_v)_i| - \eta_v \mu_q)$
17:   $(q_w)_i \leftarrow \text{sign}((q_w)_i) \max(0, |(q_w)_i| - \eta_{vw} \mu_q)$
18: end if
19: if NMF then
20:   $(p_u)_i \leftarrow \max(0, (p_u)_i)$
21:   $(q_v)_i \leftarrow \max(0, (q_v)_i)$
22:   $(q_w)_i \leftarrow \max(0, (q_w)_i)$
23: end if
24: end for
25: $G_u \leftarrow G_u + G_u/k$, $H_v \leftarrow H_v + H_v/k$, $H_w \leftarrow H_w + H_w/k$

subproblem of row-oriented BPR is

$$
\min_p \ g_u^T p + \mu_p \|p\|_1 + \frac{\sqrt{G_u}}{2\eta} \|p - p_u\|_2^2
$$

$$
\min_q \ h_v^T q + \mu_q \|q\|_1 + \frac{\sqrt{H_v}}{2\eta} \|q - q_v\|_2^2
$$

$$
\min_q \ h_w^T q + \mu_q \|q\|_1 + \frac{\sqrt{H_w}}{2\eta} \|q - q_w\|_2^2
$$

where

$$
g_u = z (q_w - q_v) + \lambda_p p_u$$

$$
h_u = -z p_u + \lambda_q q_v$$

$$
h_w = z p_u + \lambda_q q_w$$

and

$$
z = \frac{e^{P_u^T (q_w - q_v)}}{1 + e^{P_u^T (q_w - q_v)}}.$$
We omit the update rules because they are the same as (3). Algorithm 2 shows the SG procedure for solving BPR.

References


