在线最优化求解

对于推荐系统这种高维高数据量的场景，模型需要对线上数据实时响应，常见的批量处理方法显得力不从心，需要有在线处理的方法来求解模型参数。本文以模型的稀疏性作主线，逐一介绍几个在线最优化求解算法。

预备知识

凸函数

如果$f(X)$是定义在N维向量空间上的实值函数，对于$f(X)$在定义域C上的任意两点X_1和X_2，以及任意[0,1]之间的值t都有：

$$f(tX_1 + (1-t)X_2) \leq tf(X_1) + (1-t)f(X_2)$$

$$\forall X_1, X_2 \in C, 0 < t < 1$$

那么称$f(X)$是凸函数(Context)，一个函数是凸函数是它存在最优解的充分必要条件。

此外，如果$f(X)$满足：

$$f(tX_1 + (1-t)X_2) < tf(X_1) + (1-t)f(X_2)$$

$$\forall X_1, X_2 \in C, 0 < t < 1$$

则$f(X)$是严格凸函数

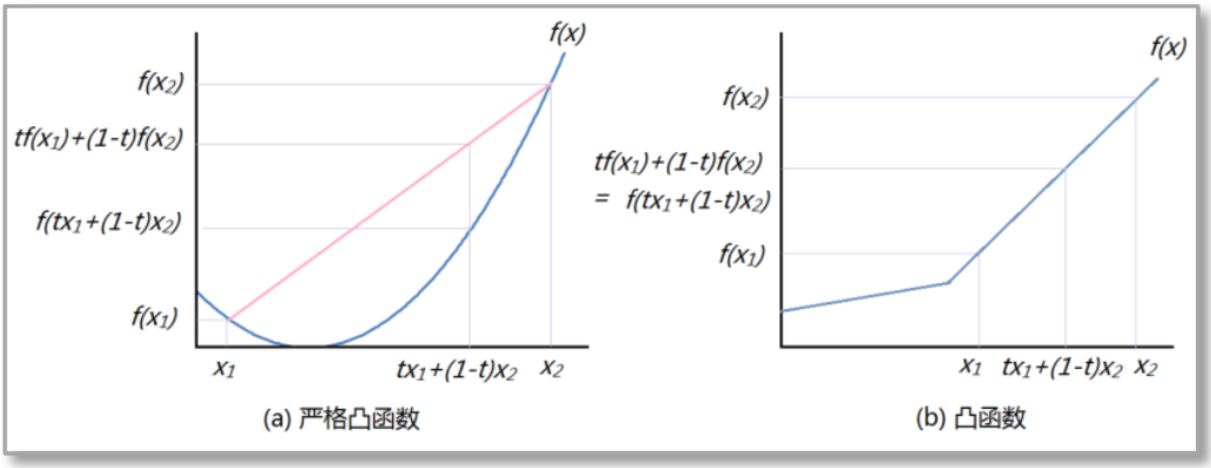

图 1 凸函数

拉格朗日乘数法及KKT条件

常见求解的最优化问题主要有如下三类：

1. 无约束优化问题

$$X = arg, \min_X f(X)$$

2. 有等式约束的优化问题

$$X = arg, \min_X f(X)$$

$$s.t. \quad h_k(X) = 0;\ k = 1, 2, \cdots, n$$

3. 有不等式约束的优化问题

$$X = arg\min_X f(X)$$

$$s.t. \quad \begin{matrix} h_k(X) = 0;\ k = 1,2 \ldots, n \\ g_l(X) \leq 0;\ l = 1,2 \ldots, m \end{matrix}$$

针对无约束优化问题，通常做法是对$f(X)$求导，并令$\frac{\partial}{\partial X} f(X) = 0$，求解得到最优值。如果$f(X)$是凸函数，则保证结果为全局最优。

针对有约束的最优化问题，常用的方法是用**KKT条件（Karush-Kuhn-Tucker Conditions**：将所有的不等式约束、等式约束和目标函数写为一个式子：

$$L(X, A, B) = f(X) + A^T H(X) + B^T G(X)$$

KKT条件是说最优值必须满足以下条件：

$$\frac{\partial}{\partial X} L(X, A, B) = 0$$

$$H(x) = 0$$

$$B^T G(X) = 0$$

KKT是求解最优解X^*的必要条件，要想其成为充分必要条件，还需要$f(X)$为凸函数才行。

在KKT条件中，由于$g_l(X) \leq 0$，所以如果要满足$B^T G(X) = 0$，需要$b_l = 0$或者$g_l(X) = 0$.

在线最优化求解算法

在机器学习中，我们面对的最优化问题都是无约束的最优化问题（有约束最优化问题可以利用拉格朗日乘数法或无约束最优问题），可以描述成：

$$W = \arg, \min_W l(W, Z)$$

$$Z = (X_j, y_j)|j = 1, 2, \cdots, M$$

$$y_j = h(W, X_j)$$

W是模型的特征权重，也是我们需要求解的参数。虽然上文已经给出了无约束最优化问题解析解的求法，但是在实际的数值计算中，通常是采用著名的梯度下降算法（GD）

$$W^{(t+1)} = W^{(t)} - \eta^{(t)} \nabla_W l(W^{(t)}, Z)$$

为了避免模型出现过拟合的情况，我们通常在损失函数的基础上加上一个关于特征权重W的限制，限制它的模不要太大

$$W = \arg, \min_W [l(W, Z) + \lambda \psi(W)]$$

$\psi(W)$称为正则化因子，是一个关于W求模的函，常用的正则化因子有L1和L2正则化

在Batch训练模型下，L1正则化可以产生更加稀疏的模型，这是我们比较关注的，除了特征选择的作用外，稀疏性可以使得预测计算的复杂度降低。

然而在Online模式下，每次W的更新不是沿着全局梯度进行下降，而是沿着某个样本的梯度方向进行下降，整个寻优过程变得像是一个“随机”查找的过程，这样即使采用L1正则化的方式，也很难产生稀疏解。

接下来沿着提升模型稀疏性的主线介绍Online模式下常用的几种优化算法。

TG

为了得到稀疏的特征权重，最简单粗暴的方式就是设定一个阈值，当W的某个纬度上系数小于这个阈值时将其设置为0（称做简单截断）。这种方法可能由于训练不足造成部分特征的丢失。

截断梯度法(TG)是对简单截断的改进，下面进行详细介绍。

L1正则化法

由于L1正则项在0处不可导，往往会造成平滑的凸优化问题变成非平滑的凸优化问题，因此采用次梯度计算L1正则项的梯度

$$W^{(t+1)} = W^{(t)} - \eta^{(t)} G^{(t)} - \eta^{(t)} \lambda sgn(W^{(t)})$$

简单截断法

以k为窗口，当t/k不为整数时采用标准的SGD进行迭代，当t/k为整数时，采用如下权重更新方式：

$$W^{(t+1)} = T_0\left(W^{(t)} - \eta^{(t)} G^{(t)}, \theta\right)$$

$$T_0(v_i, \theta) = \begin{cases} 0 & if\ |v_i| \le \theta \\ v_i & otherwise \end{cases}$$

截断梯度法

上述的简单截断法被TG的作者形容为**too aggressive**，因此TG在此基础上进行改进：

$$W^{(t+1)} = T_1\left(W^{(t)} - \eta^{(t)} G^{(t)}, \eta^{(t)} \lambda^{(t)}, \theta\right)$$

$$T_1(v_i, \alpha, \theta) = \begin{cases} \max(0, v_i - \alpha) & if\ v_i \in [0, \theta] \\ \min(0, v_i + \alpha) & if\ v_i \in [-\theta, 0] \\ v_i & otherwise \end{cases}$$

TG同样以k为窗口，每k步进行一次截断。可以看到，λ和θ决定了W的稀疏性，这两个值越大，则稀疏性越强。算法逻辑:

Algorithm 3. Truncated Gradient

input θ

initial $W \in \mathbb{R}^N$

for $t = 1,2,3\ldots$ *do*

  $G = \nabla_W \ell\left(W, X^{(t)}, y^{(t)}\right)$

  refresh W *according to*

$$w_i = \begin{cases} \max(0, w_i - \eta^{(t)} g_i - \eta^{(t)} \lambda^{(t)}) & if\ \left(w_i - \eta^{(t)} g_i\right) \in [0, \theta] \\ \max(0, w_i - \eta^{(t)} g_i + \eta^{(t)} \lambda^{(t)}) & if\ \left(w_i - \eta^{(t)} g_i\right) \in [-\theta, 0] \\ w_i - \eta^{(t)} g_i & otherwise \end{cases}$$

end

return W

将上式进行改写：

$$w_i^{(t+1)} = \begin{cases} Trnc\left(\left(w_i^{(t)} - \eta^{(t)} g_i^{(t)}\right), \lambda_{TG}^{(t)}, \theta\right) & if\ mod(t, k) = 0 \\ w_i^{(t)} - \eta^{(t)} g_i^{(t)} & otherwise \end{cases}$$

$$\lambda_{TG}^{(t)} = \eta^{(t)} \lambda k$$

$$Trnc\left(w, \lambda_{TG}^{(t)}, \theta\right) = \begin{cases} 0 & if\ |w| \le \lambda_{TG}^{(t)} \\ w - \lambda_{TG}^{(t)} sgn(w) & if\ \lambda_{TG}^{(t)} \le |w| \le \theta \\ w & otherwise \end{cases}$$

如果令$\lambda_{TG}^{(t)} = \theta$，截断公式$Trnc(w, \lambda_{TG}^{(t)}, \theta)$变为

$$Trnc(w, \lambda_{TG}^{(t)}, \theta) = \{ 0 \quad if\ |W| \le \theta\ w \quad otherwise$$

此时，TG退化成简单截断。

如果令$\theta=\infty$，截断公式$Trnc(w,\lambda_{TG}^{(t)},\theta)$变为

$$Trnc(w,\lambda_{TG}^{(t)},\theta)=\{0 \quad if\ |W|\le\lambda_{TG}^{(t)}\ w \quad otherwise$$

如果再令$k=1$，那么权重维度更新公式变为

$$w_i^{(t+1)}=Trnc((w_i^{(t)}-\eta^{(t)}g_i^{(t)}),\eta^{(t)}\lambda,\infty)=w_i^{(t)}-\eta^{(t)}g_i^{(t)}-\eta^{(t)}\lambda sgn(w_i^{(t)})$$

此时，TG退化成L1正则化法。

FOBOS

FOBOS算法原理

前向后向切分(FOBOS，Forward-Backward Spliting)，将权重的更新分为两步：

$$W^{(t+\frac{1}{2})}=W^{(t)}-\eta^{(t)}G^{(t)}$$

$$W^{(t+1)}=argmin_W\left\{\frac{1}{2}\left\|W-W^{(t+\frac{1}{2})}\right\|^2+\eta^{\left(t+\frac{1}{2}\right)}\Psi(W)\right\}$$

第一步是一个标准的梯度下降法，第二步可以看作是对梯度下降的结果进行微调（前一项是保证微调发生在梯度下降结果的附近，后一项用于处理正则化），将两个式子进行合并：

$$W^{(t+1)}=argmin_W\left\{\frac{1}{2}\left\|W-W^{(t)}+\eta^{(t)}G^{(t)}\right\|^2+\eta^{\left(t+\frac{1}{2}\right)}\Psi(W)\right\}$$

令$F(W)=\frac{1}{2}||W-W^{(t)}+\eta^{(t)}G^{(t)}||^2+\eta^{(t+\frac{1}{2})}\psi(W)$，求$\frac{\partial F(W)}{\partial W}=0$

$$W-W^{(t)}+\eta^{(t)}G^{(t)}+\eta^{(t+\frac{1}{2})}\partial\psi(W)=0$$

得到FOBOS的另一种权重更新形式：

$$W^{(t+1)}=W=W^{(t)}-\eta^{(t)}G^{(t)}-\eta^{(t+\frac{1}{2})}\partial\psi(W^{(t+1)})$$

可以看到$W^{(t+1)}$不仅与当前状态$W^{(t)}$有关，还与更新后的$\psi(W^{(t+1)})$有关。

L1-FOBOS

在L1正则化下，有$\psi(W)=\lambda ||W||*1。用向量V=[v_1,v_2,|cdots,V_N] |in R^N来表示W^{(t+|frac 12)}*, 用标量*|tilde |lambda |in R*来表示*|eta^{(t+|frac 12)}|lambda*，则权重更新公式按维度变为：*$W^{(t+1)}=|arg |min_W|sum*{i=1}^N(\frac 12(w_i-v_i)^2+\tilde \lambda |w_i|)$$

因为$\sum$的每一项都是非负的，可以拆解成每一维单独求解

$$w_i^{(t+1)} = \arg\min_{w_i}(\frac{1}{2}(w_i - v_i)^2 + \tilde{\lambda}|w_i|)$$

首先，假设w_i^*是上式的最优解，则有$w_i^* v_i \geq 0$,证明如下:

反证法：
假设：$w_i^* v_i < 0$，那么有：

$$\frac{1}{2}v_i^2 < \frac{1}{2}v_i^2 - w_i^* v_i + \frac{1}{2}(w_i^*)^2 < \frac{1}{2}(w_i^* - v_i)^2 + \tilde{\lambda}|w_i^*|$$

这与w_i^*是$\text{minimize}_{w_i}\left(\frac{1}{2}(w_i - v_i)^2 + \tilde{\lambda}|w_i|\right)$的最优解矛盾，故假设不成立，$w_i^* v_i \geq 0$

既然有$w_i^* v_i \geq 0$，可以分两种情况$v_i \geq 0$和$v_i < 0$来讨论:

(1) 当$v_i \geq 0$ 时：

由于$w_i^* v_i \geq 0$，所以$w_i^* \geq 0$，相当于对$\text{minimize}_{w_i}\left(\frac{1}{2}(w_i - v_i)^2 + \tilde{\lambda}|w_i|\right)$ 引入了不等式约束条件 $-w_i \leq 0$；
为了求解这个含不等式约束的最优化问题，引入拉格朗日乘子 $\beta \geq 0$，由 KKT 条件，有：$\frac{\partial}{\partial w_i}\left(\frac{1}{2}(w_i - v_i)^2 + \tilde{\lambda} w_i - \beta w_i\right)\Big|_{w_i = w_i^*} = 0$ 以及 $\beta w_i^* = 0$；
根据上面的求导等式可得：$w_i^* = v_i - \tilde{\lambda} + \beta$；
分为两种情况：
① $w_i^* > 0$ ：
由于 $\beta w_i^* = 0$ 所以 $\beta = 0$
这时候有：$w_i^* = v_i - \tilde{\lambda}$
又由于$w_i^* > 0$，所以$v_i - \tilde{\lambda} > 0$
② $w_i^* = 0$ ：
这时候有：$v_i - \tilde{\lambda} + \beta = 0$
又由于$\beta \geq 0$，所以$v_i - \tilde{\lambda} \leq 0$
所以，在$v_i \geq 0$时，$w_i^* = \max(0, v_i - \tilde{\lambda})$

(2) 当$v_i < 0$ 时：
采用相同的分析方法，在$v_i < 0$时，有：$w_i^* = -\max(0, -v_i - \tilde{\lambda})$

综上，FOBOS在L1正则化条件下，特征权重的各个纬度更新方式为：

$$w_i^{(t+1)} = sgn(v_i) max(0, |v_i| - \tilde{\lambda})$$

$$= sgn\left(w_i^{(t)} - \eta^{(t)} g_i^{(t)}\right) max\left\{0, \left|w_i^{(t)} - \eta^{(t)} g_i^{(t)}\right| - \eta^{(t+\frac{1}{2})}\lambda\right\}$$

可以看出，L1-FOBOS在每次更新W之前都会判断，当$|w_i^{(t)}-\eta^{(t)}g_i^{(t)}|-\eta^{(t+\frac{1}{2})}\lambda\leq 0$时都会对该纬度进行截断。

直观理解就是：当一条样本产生的梯度令对应纬度的权重值发生足够大的变化$\eta^{(t+\frac{1}{2})}\lambda$时，认为在本次更新过程中该纬度不够重要，应当令其权重为0。

RDA

RDA算法原理

简单截断、TG、FOBOS都是建立在SGD的基础之上，属于梯度下降类算法，这类方法的优点是精度比较高，能在稀疏性上得到提升。

正则对偶平均(RDA，Regularized Dual Averaging)从另一方面来求解Online Optimization，并且更有效地提升了特征权重的稀疏性，其特征权重的更新策略为：

$$W^{(t+1)}=\underset{W}{argmin}\left\{\frac{1}{t}\sum_{r=1}^{t}\langle G^{(r)},W\rangle+\Psi(W)+\frac{\beta^{(t)}}{t}h(W)\right\} \tag{3-3-1}$$

其中，$\langle G^{(r)},W\rangle$表示梯度$G^{(r)}$对W的积分平均值，$\psi(W)$为正则项，$h(W)$为一个辅助的严格凸函数，$\beta^{(t)}|t\geq 1$是一个非负且非自减序列。

本质上，上式包含了三部分

- 线性函数$\frac{1}{2}\sum_{r=1}^{t}\langle G^{(r)},W\rangle$，包含了之前所有梯度（或负梯度）的平均值。
- 正则项$\psi(W)$
- 额外正则项$\beta^{(t)}|t\geq 1$，是一个严格凸函数。

L1-RDA

在L1正则化下，令$\psi(W)=\lambda||W||_1$，$h(W)=\frac{1}{2}||W||_2^2$，$\beta^{(t)}|t\geq 1$定义为$\beta^{(t)}=\gamma\sqrt{t}$，则权重更新公式为：

$$W^{(t+1)}=\underset{W}{argmin}\left\{\frac{1}{t}\sum_{r=1}^{t}\langle G^{(r)},W\rangle+\lambda\|W\|_1+\frac{\gamma}{2\sqrt{t}}\|W\|_2^2\right\} \tag{3-3-2}$$

拆分成N个独立的标量最小化问题：

$$\underset{w_i\in\mathbb{R}}{minimize}\left\{\bar{g}_i^{(t)}w_i+\lambda|w_i|+\frac{\gamma}{2\sqrt{t}}w_i^2\right\} \tag{3-3-3}$$

其中，$\lambda>0$，$\frac{\gamma}{\sqrt{t}}>0$，$\bar{g}_i^{(t)}=\frac{1}{t}\sum_{r=1}^{t}g_i^{(r)}$

假设$w_i^是最优解，并且定义|xi |in |partial|w_i^为|w_i^|在w_i^$的次倒数，则有

$$\partial|w_i^*| = \begin{cases} \{-1 < \xi < 1\} & if\ w_i^* = 0 \\ \{1\} & if\ w_i^* > 0 \\ \{-1\} & if\ w_i^* < 0 \end{cases} \tag{3-3-4}$$

对公式3-3-3进行求导并等于0，有

$$\bar{g}_i^{(t)} + \lambda\xi + \frac{\gamma}{\sqrt{t}} w_i = 0$$

由于$\lambda > 0$，针对上式分三种情况$|\bar{g}_i^{(t)}| < \lambda$、$\bar{g}_i^{(t)} > \lambda$和$\bar{g}_i^{(t)} < \lambda$讨论：

(1) 当$\left|\bar{g}_i^{(t)}\right| < \lambda$ 时：

还可以分为三种情况：

① 如果$w_i^* = 0$，由(3-3-5)得$\xi = -\bar{g}_i^{(t)}/\lambda \in \partial|0|$，满足(3-3-4)

② 如果$w_i^* > 0$，由(3-3-4)得$\xi = 1$，则$\bar{g}_i^{(t)} + \lambda + \frac{\gamma}{\sqrt{t}} w_i > \bar{g}_i^{(t)} + \lambda \geq 0$，不满足 (3-3-5)

③ 如果$w_i^* < 0$，由(3-3-4)得$\xi = -1$，则$\bar{g}_i^{(t)} - \lambda + \frac{\gamma}{\sqrt{t}} w_i < \bar{g}_i^{(t)} - \lambda \leq 0$，不满足 (3-3-5)

所以，当$\left|\bar{g}_i^{(t)}\right| < \lambda$时，$w_i^* = 0$

(2) 当$\bar{g}_i^{(t)} > \lambda$ 时：

采用相同分析方法得到$w_i^* < 0$，有$\xi = -1$满足条件，即：$w_i^* = -\frac{\sqrt{t}}{\gamma}\left(\bar{g}_i^{(t)} - \lambda\right)$

(3) 当$\bar{g}_i^{(t)} < -\lambda$ 时：

采用相同分析方法得到$w_i^* > 0$，有$\xi = 1$满足条件，即：$w_i^* = -\frac{\sqrt{t}}{\gamma}\left(\bar{g}_i^{(t)} + \lambda\right)$

综合上面的分析可以得到L1-RDA特征权重的各个纬度更新方式为：

$$w_i^{(t+1)} = \begin{cases} 0 & \text{if } \left|\bar{g}_i^{(t)}\right| < \lambda \\ -\frac{\sqrt{t}}{\gamma}\left(\bar{g}_i^{(t)} - \lambda sgn(\bar{g}_i^{(t)})\right) & otherwise \end{cases} \tag{3-3-6}$$

直观理解：当某个纬度上累计梯度平均值的绝对值$|\bar{g}_i^{(t)}|$小于阈值λ时，该纬度权重将被置为0.

FTRL

L1-FOBOS和L1-LDA形式统一

L1-FOBOS这一类基于梯度下降的方法有比较高的精度，L1-RDA却能够在损失一定精度的情况下产生更好的稀疏性。 FTRL(Follow the Regularized Leader)结合了L1-FOBOS和L1-RDA的优点。

下面将先对L1-FOBOS和L1-RDA的形式进行统一。

L1-FOBOS的迭代形式，这里令$\eta^{(t+\frac{1}{2})}=\eta^{(t)}=\Theta(\frac{1}{\sqrt{t}})$

$$W^{(t+\frac{1}{2})}=W^{(t)}-\eta^{(t)}G^{(t)}$$

$$W^{(t+1)}=\arg\min_W \frac{1}{2}||W-W^{(t+\frac{1}{2})}||^2+\eta^{(t)}\lambda||W||_1$$

把上面两个公式合在一起，有：

$$W^{(t+1)}=\arg\min_W \frac{1}{2}||W-W^{(t)}+\eta^{(t)}G^{(t)}||^2+\eta^{(t)}\lambda||W||_1$$

将上式按维度拆分成N个独立的最优化步骤:

$$ \begin{aligned}

&\min_{w_i \in R}{\frac 12(w_i-w_i^{(t)}+\eta^{(t)}g_i^{(t)})^2+\eta^{(t)}\lambda|w_i|} \ &\min_{w_i \in R}{\frac 12(w_i-w_i^{(t)})^2+\frac 12(\eta^{(t)}g_i^{(t)})^2+w_i\eta^{(t)}g_i^{(t)}-w_i^{(t)}\eta^{(t)}g_i^{(t)}+\eta^{(t)}\lambda|w_i|} \ &\min_{w_i \in R}{w_ig_i^{(t)}+\lambda|w_i|+\frac 1{2\eta^{(t)}}(w_i-w_i^{(t)})^2+[\frac {\eta^{(t)}}2(g_i^{(t)})^2-w_i^{(t)}g_i^{(t)}]} \end{aligned} $$

变量$\frac{\eta^{(t)}}{2}(g_i^{(t)})^2-w_i^{(t)}g_i^{(t)}$与$w_i$无关，因此上式等价于

$$\min_{w_i\in R} w_ig_i^{(t)}+\lambda|w_i|+\frac{1}{2\eta^{(t)}}(w_i-w_i^{(t)})^2$$

再将这N个独立最优化子步骤合并，那么L1-FOBOS可以写作

$$W^{(t+1)}=\underset{W}{argmin}\left\{G^{(t)}\cdot W+\lambda\|W\|_1+\frac{1}{2\eta^{(t)}}\|W-W^{(t)}\|_2^2\right\}$$

而L1-RD的公式可以写作：

$$W^{(t+1)}=\underset{W}{argmin}\left\{G^{(1:t)}\cdot W+t\lambda\|W\|_1+\frac{1}{2\eta^{(t)}}\|W-0\|_2^2\right\}$$

这里$G^{(1:t)}=\sum_{s=1}^t G^{(s)}$；令$\sigma^{(s)}=\frac{1}{\eta^{(s)}}-\frac{1}{\eta^{(s-1)}}$，$\sigma^{(1:t)}=\frac{1}{\eta^{(t)}}$，则上面两个式子可以写作：

$$W^{(t+1)} = \underset{W}{argmin}\left\{G^{(t)} \cdot W + \lambda\|W\|_1 + \frac{1}{2}\sigma^{(1:t)}\left\|W - W^{(t)}\right\|_2^2\right\} \tag{3-4-1}$$

$$W^{(t+1)} = \underset{W}{argmin}\left\{G^{(1:t)} \cdot W + t\lambda\|W\|_1 + \frac{1}{2}\sigma^{(1:t)}\|W - 0\|_2^2\right\} \tag{3-4-2}$$

可以看出，L1-FOBOS和L1-RDA的区别在于：

1. 前者计算的是梯度以及L1正则项对当前模的影响，后者采用了累加的处理方式。
2. 前者的第三项限制W的变化不能离已迭代过的解太远，而后者则限制W不能离0点太远。

FTRL算法原理

FTRL综合考虑了FOBOS和RDA对于正则项和W限制的区别，其特征权重的更新公式为：

$$W^{(t+1)} = \underset{W}{argmin}\left\{G^{(1:t)} \cdot W + \lambda_1\|W\|_1 + \lambda_2\frac{1}{2}\|W\|_2^2 + \frac{1}{2}\sum_{s=1}^{t}\sigma^{(s)}\left\|W - W^{(s)}\right\|_2^2\right\} \tag{3-4-3}$$

L2正则项的引入相当于在求解过程中加了一个约束，使得结果更加平滑。

（3-4-3）式展开

$$W^{(t+1)} = \underset{W}{argmin}\left\{\left(G^{(1:t)} - \sum_{s=1}^{t}\sigma^{(s)}W^{(s)}\right) \cdot W + \lambda_1\|W\|_1 + \frac{1}{2}\left(\lambda_2 + \sum_{s=1}^{t}\sigma^{(s)}\right)\|W\|_2^2 + \frac{1}{2}\sum_{s=1}^{t}\sigma^{(s)}\left\|W^{(s)}\right\|_2^2\right\}$$

其中，$\frac 12\sum_{s=1}^t\sigma^{(s)}||W^{(s)}||*2^2对于W^{(t+1)}*
来说是个常数，可以省略。令*Z^{(t)}=G^{(1:t)}-|sum*{s=1}^t\sigma^{(t)}W^{(t)}$

上式等价于

$$W^{(t+1)} = \underset{W}{argmin}\left\{Z^{(t)} \cdot W + \lambda_1\|W\|_1 + \frac{1}{2}\left(\lambda_2 + \sum_{s=1}^{t}\sigma^{(s)}\right)\|W\|_2^2\right\}$$

针对每个维度将其拆分成N个独立的标量最小化问题

$$\underset{w_i \in \mathbb{R}}{minimize}\left\{z_i^{(t)}w_i + \lambda_1|w_i| + \frac{1}{2}\left(\lambda_2 + \sum_{s=1}^{t}\sigma^{(s)}\right)w_i^2\right\}$$

求导解析得到

$$w_i^{(t+1)} = \begin{cases} 0 & \text{if } \left|z_i^{(t)}\right| < \lambda_1 \\ -\left(\lambda_2 + \sum_{s=1}^{t} \sigma^{(s)}\right)^{-1} \left(z_i^{(t)} - \lambda_1 sgn(z_i^{(t)})\right) & otherwise \end{cases}$$

Per-Coordinate Learning Rates

在FTRL中，针对每个维度的学习率$\eta^{(t)}$的选择和计算都是单独考虑的，标准的OGD里面使用的是一个全局的学习旅$\eta^{(t)} = \frac{1}{\sqrt{t}}$。

FTRL中，维度i上的学习率计算方式：

$$\eta_i^{(t)} = \frac{\alpha}{\beta + \sqrt{\sum_{s=1}^{t} \left(g_i^{(s)}\right)^2}} \quad \textbf{(3-4-5)}$$

由于$\sigma^{(1:t)} = \frac{1}{\eta^{(t)}}$，所以$\sum_{s=1}^{t} \sigma^{(s)} = \frac{1}{\eta^{(t)}} = \frac{\beta+\sqrt{\sum_{s=1}^{t}(g_i^{(s)})^2}}{\alpha}$，这里的$\alpha, , \beta$是需要输入的参数。

FTRL算法逻辑

Algorithm 6. FTRL-Proximal with L1 & L2 Regularization

input $\alpha,\ \beta,\ \lambda_1,\ \lambda_2$
initialize $W \in \mathbb{R}^N,\ Z = 0 \in \mathbb{R}^N,\ Q = 0 \in \mathbb{R}^N$
for t =1,2,3… do
  $G = \nabla_W \ell\left(W, X^{(t)}, y^{(t)}\right)$ *# gradient of loss function*
  for i in 1,2,…,N do *# for each coordinate*
    $\sigma_i = \frac{1}{\alpha}\sqrt{q_i + g_i^2} - \sqrt{q_i}\ \ \&\ \ q_i = q_i + g_i^2$ *# equals* $\frac{1}{\eta^{(t)}} - \frac{1}{\eta^{(t-1)}}$
    $z_i = z_i + g_i - \sigma_i w_i$
    $w_i = \begin{cases} 0 & \text{if } \left|z_i^{(t)}\right| < \lambda_1 \\ -\left(\lambda_2 + \frac{\beta+\sqrt{q_i}}{\alpha}\right)^{-1} (z_i - \lambda_1 sgn(z_i)) & otherwise \end{cases}$
  end
end
return W

程序中第六行更改

\sigma_i=\frac 1{\alpha}(\sqrt {q_i+g_i^2}- \sqrt {q_i}) ;;&;; q_i=qi+g_i^2