1. 引理

目标函数为$F(\theta)=g(\theta)+h(\theta)$, 其中$g(\theta)$是可微的凸函数, $h(\theta)$是不可微的凸函数, 如果能找到一个辅助函数据$G(\theta,\theta_t)$, 使得:

$$F(\theta)\le G(\theta,\theta_t);\ F(\theta_t)=G(\theta_t,\theta_t)$$

如果按$\theta_{t+1}=argmin_\theta G(\theta,\theta_t)$迭代, 则有:

$$F(\theta_{t+1})\le G(\theta_{t+1},\theta_t)\le G(\theta_t,\theta_t)=F(\theta_t)$$

又因为$F(\theta)$是凸的, 所以按上述方式迭入可以找到$F(\theta)$的最小值.

2. 简单应用

在机器学习中通常有$g(\theta)$为:

$$g(\theta)=\frac{1}{N}\sum_i^N loss(y_i,f_\theta(x_i))+\frac{\lambda_2}{2}\|\theta\|_2^2$$

其中第一部分是经验风险(平均损失/期望损失), 第二部分是L_2正则项. 一般而言, 不可微项是L_1正则, 即

$$h(\theta)=\lambda_1\|\theta\|_1$$

如记$g(\theta)$中的第一项(经验风险)为$\phi(\theta)$, 则可定义$G(\theta,\theta_t)$为:

$$G(\theta,\theta_t)=\phi(\theta_t)+\nabla\phi(\theta_t)^T(\theta-\theta_t)+\frac{1}{2\xi}\|\theta-\theta_t\|_2^2+\frac{\lambda_2}{2}\|\theta\|_2^2+\lambda_1\|\theta\|_1$$

只要ξ足够小, 总有ξ使得$F(\theta)\le G(\theta,\theta_t);\ F(\theta_t)=G(\theta_t,\theta_t)$成立. $G(\theta,\theta_t)$是二次函数, 所以有解析解:

$$G(\theta,\theta_t)=\frac{1}{2}(\frac{1}{\xi}+\lambda_2)\theta^T\theta+(\nabla\phi(\theta_t)-\frac{1}{\xi}\theta_t)^T\theta+\lambda_1\|\theta\|_1+C$$

其中C为常数项, 去除常数项, 两边乘以ξ有:

$$G(\theta,\theta_t)\sim\frac{1}{2}(1+\lambda_2\xi)\theta^T\theta+(\xi\nabla\phi(\theta_t)-\theta_t)^T\theta+\xi\lambda_1\|\theta\|_1$$

两边除以$1+\lambda_2\xi$, 使用配方法有:

$$G(\theta,\theta_t) \sim \frac{1}{2}\|\theta - [\theta_t - \frac{\xi}{1+\lambda_2\xi}(\nabla\phi(\theta_t)+\lambda_2\theta_t)]\|_2^2 + \frac{\xi\lambda_1}{1+\lambda_2\xi}\|\theta\|_1$$

如果记, 则上述方程有解析解

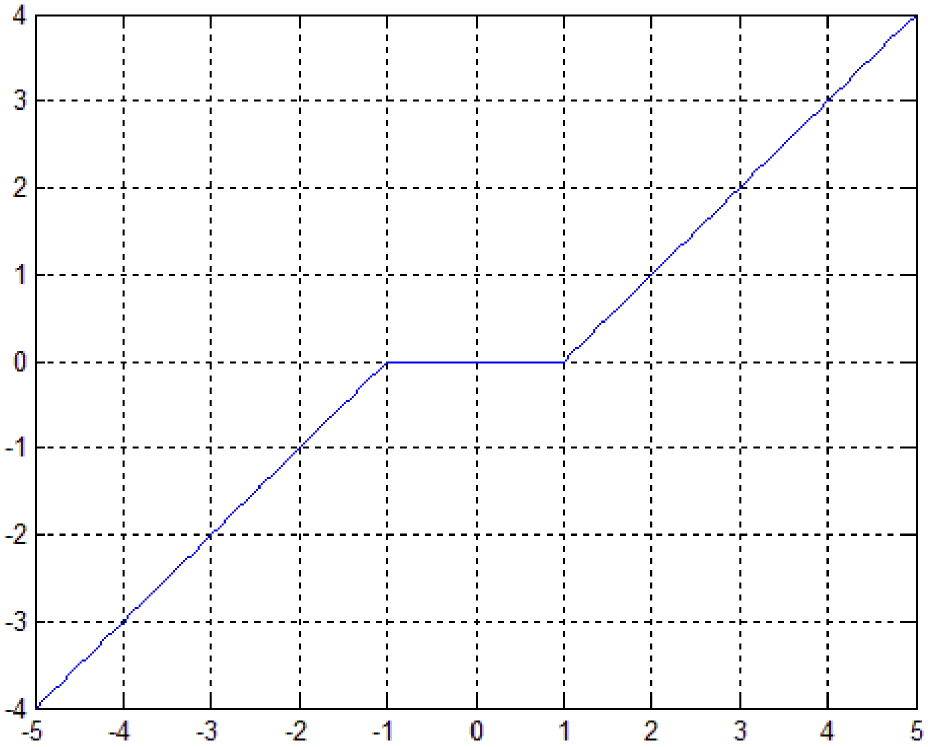
$$z_{t+1} = \theta_t - \frac{\xi}{1+\lambda_2\xi}(\nabla\phi(\theta_t)+\lambda_2\theta_t)$$

- 如果$\lambda_1 = 0, \lambda_2 = 0$, 此时有$\theta_{t+1} = z_{t+1} = \theta_t - \xi\nabla\phi(\theta_t)$, 迭代退化为梯度下降, ξ 为学习率.
- 如果$\lambda_1 \neq 0, \lambda_2 = 0$, 此时有$z_{t+1} = \theta_t - \xi\nabla\phi(\theta_t)$, 下一轮迭代值为 $\theta_{t+1} = S(z_{t+1}, \lambda_1\xi)_+$
- 如果$\lambda_1 \neq 0, \lambda_2 \neq 0$, 此时有$z_{t+1} = \theta_t - \frac{\xi}{1+\lambda_2\xi}(\nabla\phi(\theta_t)+\lambda_2\theta_t)$, 下一轮迭代值为$\theta_{t+1} = S(z_{t+1}, \frac{\xi\lambda_1}{1+\lambda_2\xi})_+$

其中$S(z,\lambda)_+$的定义为:

$$S(z,\lambda)_+ = \begin{cases} z-\lambda, & z>\lambda \\ 0, & |z|\leq\lambda \\ z+\lambda, & z<-\lambda \end{cases}$$

上式称为迭代软阈值, 用图形表示为:



3. 进一步应用

考虑更复杂一点的$G(\theta, \theta_t)$, 即使用Hessian矩阵的对解近似, 如下:

$$G(\theta, \theta_t) = \phi(\theta_t) + \nabla\phi(\theta_t)^T(\theta - \theta_t) + \frac{1}{2\xi}(\theta - \theta_t)^T \wedge_t (\theta - \theta_t) + \frac{\lambda_2}{2}\|\theta\|_2^2 + \lambda_1\|\theta\|_1$$

其中$\wedge_t$是一个对角矩阵, 它是Hessian矩阵的对角近似, 一般为正定矩阵. 同理$G(\theta, \theta_t)$也满足: $F(\theta) \leq G(\theta, \theta_t)$; $F(\theta_t) = G(\theta_t, \theta_t)$, 所以有:

$$G(\theta, \theta_t) \sim \frac{1}{2}\theta^T(\frac{1}{\xi} \wedge_t + \lambda_2 I)\theta + (\nabla\phi(\theta_t) - \frac{1}{\xi} \wedge_t \theta_t)^T\theta + \lambda_1\|\theta\|_1$$

$$G(\theta, \theta_t) \sim \frac{1}{2}\theta^T(\wedge_t + \lambda_2\xi I)\theta + (\xi\nabla\phi(\theta_t) - \wedge_t\theta_t)^T\theta + \lambda_1\xi\|\theta\|_1$$

上述方程式是可以分离变量的, 下面对单个维度计算:

$$\frac{1}{2}(\wedge_t^{(i)} + \lambda_2\xi){\theta^{(i)}}^2 + (\xi\nabla\phi(\theta_t)^{(i)} - \wedge_t^{(i)}\theta_t^{(i)})\theta^{(i)} + \lambda_1\xi|\theta^{(i)}|$$

这个方程有解析解:

$$z_{t+1}^{(i)} = \theta_t^{(i)} - \frac{\xi}{\wedge_t^{(i)} + \lambda_2 \xi}(\nabla \phi(\theta_t)^{(i)} + \lambda_2 \theta_t^{(i)})$$

而下一轮的$\theta_{t+1}^{(i)}$为:

$$\theta_{t+1}^{(i)} = S(z_{t+1}, \frac{\lambda_1 \xi}{\wedge_t^{(i)} + \lambda_2 \xi})_+$$

关于Hession的对角近似

方案1: 用累积梯度平方和开根号

这是AdaGrad的做法, 即有:

$$n_t = n_{t-1} + \nabla f(\theta_t)^2;\ \wedge_t = diag(\sqrt{n_t})$$

这种做法过了激进, $\wedge_t$会因为n_t没有衰减而迅速增大, 实际使用中效果并不好.

方案2: 用指数平滑梯度平方和开根号

这是Adadelta, RMSprop的做法, 即有:

$$n_0 = 0;\ n_{t+1} = \beta n_t + (1-\beta)\nabla f(\theta_t)^2;\ \wedge_t = diag(\sqrt{n_{t+1}})$$

与上面的累加相比, 做了指数平滑, 使较早的梯度分量以指数形式快速衰减, 如下:

$$n_{t+1} = (1-\beta)(\beta^{t-1}\nabla f(\theta_1)^2 + \beta^{t-2}\nabla f(\theta_2)^2 + \cdots + \beta \nabla f(\theta_{t-1})^2 + \nabla f(\theta_t)^2)$$

从上面的公式看出较早的梯度分量被指数衰减了, 这样有效地防止了AdaGrad中因梯度累积造成的问题, 所以实际效果更好.

下面说明原因, 系数是等比数列, 求积公式为:

$$a_n = a_1 q^{n-1} \text{ where q} < 1;\ sum = a_1 \frac{1-q^n}{1-q}$$

对应上式有$a_1 = 1-\beta, q = \beta$, 所以系数之和为$1-\beta^t$, 当$t \to \infty$, 有:

$$n_t^i < \max\left(\nabla f(\theta_1)^{2^{(i)}}, \nabla f(\theta_2)^{2^{(i)}}, \nabla f(\theta_3)^{2^{(i)}}, \cdots\right)$$

这就说明了指数平滑能很好地防御因梯度累积造成的问题.

方案3: 用修正的指数平滑梯度平方和开根号

这是Adam的做法. 指数平滑在极限上是有界的, 但在局部还是会有一定的梯度累积效应带来的误差, 可以用事下方式修正:

$$n_0 = 0;\ n_{t+1} = \beta n_t + (1-\beta)\nabla f(\theta_t)^2;\ \wedge_t = diag(\sqrt{\frac{n_{t+1}}{1-\beta^t}})$$

分母上除的$1-\beta^t$就是分子中所有系数之和. 即变成了指数平滑加权平均. 这样它的量级就等同于梯度了.