Optimization Theory and Algorithm	Homework 4 - 10/06/2021

Homework 4

Lecturer: Xiangyu Chang *Scribe:* 赵敬业

1 HW 1

1.1 问题重述

证明下列定理。

(1) 当且仅当 $f(\mathrm{x})-\frac{\alpha}{2}\|\mathrm{x}\|^2$ 是凸时，f 是 α -强凸的.

(2) 假设 $f\in\mathcal{C}^1$. 那么下面的 mi 是等价的：

1. f 是 α -强凸的.
2. $\langle\nabla f(\mathbf{y})-\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle\geq\alpha\|\mathbf{x}-\mathbf{y}\|^2$
3. 如果 $f\in C^2$，则 $\nabla^2 f\succeq\alpha I$ 处处成立.($\nabla^2 f$ 是正定的.)

1.1.1 证明（1）

必要性：$\forall x,y\in dom(f)$ 满足：

$$f(y)\geq f(x)+<\nabla f(x),y-x>+\frac{\alpha}{2}||y-x||^2 \tag{1}$$

令 $g(x)=f(x)-\frac{\alpha}{2}||x||^2$，则

$$\bigtriangledown g(x)=\nabla f(x)-\alpha||x||$$

$$\begin{aligned}
g(y)&=f(y)-\frac{\alpha}{2}||y||^2\\
&\geq f(x)+<\nabla f(x),y-x>+\frac{\alpha}{2}||y-x||^2-\frac{\alpha}{2}||y||^2\\
&=g(x)+<\nabla f(x)-\alpha x,y-x>+\frac{\alpha}{2}<x,x>+\alpha<x,y-x>+\frac{\alpha}{2}<x-y,x-y>+\frac{\alpha}{2}-\frac{\alpha}{2}<y,y>\\
&=g(x)+<\nabla g(x),y-x>+\alpha<x,y>-\frac{\alpha}{2}<x,y>+\frac{\alpha}{2}<y,-x>\\
&=g(x)+<\nabla g(x),y-x>
\end{aligned}$$

必要性得证

充分性

$g(x)=f(x)-\frac{\alpha}{2}||x||^2$

$$g(x)=f(x)-\frac{\alpha}{2}||x||^2$$

$$g(y)\geq g(x)+<\nabla g(x),y-x>$$

易知，$g(x)+<\nabla g(x),y-x>=f(x)+<\nabla f(x),y-x>+\frac{\alpha}{2}||y-x||^2-\frac{\alpha}{2}||y||^2$ 成立

$$
\begin{aligned}
g(y) &\geq f(x)+<\nabla f(x),y-x>+\frac{\alpha}{2}||y-x||^2-\frac{\alpha}{2}||y||^2 \\
f(y) &\geq f(x)+<\nabla f(x),y-x>+\frac{\alpha}{2}||y-x||^2
\end{aligned}
$$

QED

1.2 证明（2）

1⇒2

$$
\begin{aligned}
f(y) &\geq f(x)+\langle\nabla f(x),y-x\rangle+\frac{\alpha}{2}\|x-y\|^2 \\
f(x) &\geq f(y)+\langle\nabla f(y),x-y\rangle+\frac{\alpha}{2}\|y-x\|^2
\end{aligned}
$$

(1)+(2) 得

$$
\begin{aligned}
f(x)+f(y) &\geq f(x)+f(y)\,\langle\nabla f(x),y-x\rangle-\langle\nabla f(y),y-x\rangle+\alpha\|x-y\|^2 \\
&\Rightarrow \langle\nabla f(y)-\nabla f(x),y-x\rangle\geq\alpha\|y-x\|^2=\alpha\|x-y\|^2
\end{aligned}
$$

2⇒3

取 $y=x+tv$：

$$\langle\nabla f(x+tv)-\nabla f(x),tv\rangle\geq\alpha\|tv\|^2$$

即：

$$(\nabla f(x+tv)-\nabla f(x))^T tv\geq\alpha t^2\|v\|^2$$

$$\lim_{t\to 0}\frac{(\nabla f(x+tv)-\nabla f(x))^T v}{t}\geq\alpha\|v\|^2$$

由 $\frac{\partial\nabla f(x)}{\partial v}=\nabla^2 f(x)v$, 可以知道 $\lim\limits_{t\to 0}\frac{(\nabla f(x+tv)-\nabla f(x))^T}{t}\geq\nabla^2 f(x)*v$, 则

$$v^T\nabla^2 f(x)v\geq\alpha\|v\|^2$$

$$
\begin{aligned}
\left\langle v,\nabla^2 f(x)v\right\rangle &\geq\langle v,\alpha v\rangle \\
\Leftrightarrow\left\langle v,(\nabla^2 f(x)-\alpha I)v\right\rangle &\geq 0 \\
\Leftrightarrow\nabla^2 f(x) &\geq\alpha I
\end{aligned}
$$

3⇒1

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) + o(\|y-x\|^2)$$

$$\Leftrightarrow$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y-x\|^2 + o(\|y-x\|^2)$$

$$\Leftrightarrow$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y-x\|^2$$

2 HW 2

2.1 问题重述

请计算下列函数的次梯度和次微分:

(1)$f(X) = \max(x\ 0)$ 称为 $RelU$，被广泛应用于深度学习模型中。

(2)$f(\mathbf{x}) = \max_{i=1,\ldots,m} \left\{\mathbf{a}_i^\top \mathbf{x} + b_i\right\}$.

2.2 解（1）

当 $x < 0$ 时，$f(x) = 0, \partial f(x) = \{0\}$
当 $x > 0$ 时，$f(x) = x, \partial f(x) = \{1\}$
当 $x = 0$ 时，显然 $\partial f(x) = [0, 1]$
综上：

$$\partial f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ [0,1] & \text{for} \quad x = 0 \\ 1 & \text{for} \quad x > 0 \end{cases}$$

2.3 解（2）

令 $f_i(x) = a_i^{\mathrm{T}} x + b_i, i = 1, 2, \ldots, m$ 则

$$\partial f(x) = convex \left(\bigcup_{i \in I(x^0)} \left\{a_i^{\mathrm{T}}\right\} \right)$$
$$I(x^c) = \left\{i : f(x^0) = f_i(x^0)\right\}$$

3 HW 3

3.1 问题重述

假设：

* data：$\{\mathbf{a}_i, b_i\}_{i=1}^m$，其中 $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}$.

* $b_i = \mathbf{a}_i^\top \mathbf{x} + c_i$, 其中 $\mathbf{x} = (x_1, \dots, x_n)^\top$ 表示回归系数，$c_i \sim \mathcal{N}(0\ 1)$.

* 先验分布：$\mathbf{x} \sim \mathcal{L}\left(0, \frac{1}{\lambda} I_n\right)$，其中

$$\mathbb{P}(\mathbf{x}) = \frac{1}{g(\lambda)} \exp\left\{-\frac{\lambda\|\mathbf{x}\|_1}{2}\right\}$$

* 后验分布：

$$\mathbb{P}(\mathbf{x} \mid A, \mathbf{b}) = \frac{\mathbb{P}(A, \mathbf{b} \mid \mathbf{x})\mathbb{P}(\mathbf{x})}{\mathbb{P}(A, \mathbf{b})}$$

用最大后验法 (MAP) 推导 LASSO 的优化公式：

3.2 证明

由定义可知，

$$P(A, b|x) = \prod_{i=1} p(a_i, b_i|x) = \frac{1}{(2\pi)^{\frac{y}{2}}} \prod_{i=1} e^{-\frac{(b_i - a_i x)^2}{2}}$$

$$\max_x P(x|A, b) = \max_x P(A, b|x)P(x)$$
$$\Leftrightarrow \max_x P(x|A, b) = \max_x \prod_{i=1} P(a_i, b_i|x)P(x)$$
$$\Leftrightarrow \max_x P(x|A, b) = \frac{1}{(2\pi)^{\frac{y}{2}}} \max_x \prod_{i=1} e^{-\frac{(b_i - a_i^T x)^2}{2}} P(x)$$

两边取 log

$$\Rightarrow \max_x \log \prod_{i=1} e^{-\frac{(b_i - a_i^T x)^2}{2}} P(x)$$
$$= \max_X (\sum_{i=1} e^{-\frac{(b_i - a_i^T x)^2}{2}} P(x) + \log \frac{1}{g(\lambda)} e^{-\frac{-\lambda\|x\|_1}{2}})$$

$$= \max_X (\sum_{i=1} e^{-\frac{(b_i - a_i^T x)^2}{2}} + \frac{-\lambda\|x\|_1}{2})$$

两边取负数

$$min_{i=1} \sum_{i=1} (b_i - a_i^T x)^2 + \lambda\|x\|_1$$

即：

$$\min_x \|Ax - b\|^2 + \lambda \|x\|_1$$

QED

4 HW 4

4.1 问题重述

下面的优化问题称为弹性 nel，它是由 [Zou and Haslie, 2005] 提出的:

$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_1+(1-\lambda)\|\mathbf{x}\|_2^2$$

(1) $g(\mathbf{x})=\lambda\|\mathbf{x}\|_1+(1-\lambda)\|\mathbf{x}\|_2^2$. 请计算 ils 近端算子 prox $_{\gamma g}(\mathbf{z})$.
(2) 给出求解弹性 nel 的近似梯度下降算法。

4.2 解（1）

$$prox_{rq}(z)=$$

$$\arg\min_x\left\{\frac{1}{2r}\|x-z\|^2+g(x)\right\}=\arg\min_x\left\{\frac{1}{2r}\|x-z\|^2+\lambda\|x\|_1+(1-\lambda)\|x\|_2^2\right\}$$

$$=\arg\min_x\left\{\sum_{i=1}^{y}(\frac{1}{2r}(x_i-z_i)^2+\lambda|x_i|+(1-\lambda)x_i^2)\right\}$$

则对每个 x_i 都需要

$$prox_{rq}(z)=\arg\min_{x_i}\left\{\frac{1}{2r}(x_i-z_i)^2+\lambda|x|+(1-\lambda)x_i^2\right\}$$

令上面式子 $=\phi(x)$ 当 $x_i\geq 0$ 时

$$\phi(x_i)=\frac{1}{2r}(x_i-z_i)^2+\lambda x_i+(1-\lambda)x_i^2$$

$$\phi^{'}(x_i)=\frac{1}{r}(x_i-z_i)+\lambda+2(1-\lambda)x_i=0$$

$$\Rightarrow x_i=\frac{z_i-r\lambda}{1+2(1-\lambda)r}$$

其中 $z_i>r\lambda$ 当 $x_i=0$ 时

$$\partial\phi(x_i)=\frac{1}{r}(x_i-z_i)+\lambda\partial|x_i|+2(1-\lambda)x_i$$

$$0\in\partial\phi(x_i)\Rightarrow -\frac{z_i}{r\lambda}\in[-1,1]$$

其中 $z_i\in[-r\lambda,r\lambda]$ 当 $x_i\leq 0$ 时

$$\phi(x_i)=\frac{1}{2r}(x_i-z_i)^2-\lambda x_i+(1-\lambda)x_i^2$$

$\phi^{'}(x_i)=\frac{1}{r}(x_i-z_i)-\lambda+2(1-\lambda)x_i\rightarrow x_i=\frac{z_i+r\lambda}{1+2(1-\lambda)r}$
其中满足 $z_i<r\lambda$ 综上：

$$
x_i = \begin{cases} \frac{z_i - r\lambda}{1+2(1-\lambda)r} & z_i > r\lambda \\ 0 & z_i \in [-r\lambda, r\lambda] \\ \frac{z_i + r\lambda}{1+2(1-\lambda)r} & z_i < r\lambda \end{cases}
$$

综上所述：

$$
x = prox_{rq} \frac{1}{1+2(1-\lambda)r} sign(z)(|z| - r\lambda)_+
$$

4.3 解 (2)

1: 输入: 给出初值点 $x^0 \in dom(f)$,t_{max} 和 $t=0$、$\beta = \lambda_{max}\left(A^{\mathrm{T}}A\right)$

2: while $\quad t < t_{max}$do

3:

$$
z^t = x^t - \frac{1}{\lambda_{max}\left(A^{\mathrm{T}}A\right)} A^{\mathrm{T}}\left(Ax^t - b\right) = \left(I - \frac{A^{\mathrm{T}}A}{\lambda_{max}\left(A^{\mathrm{T}}A\right)}\right) x^t + \frac{A^{\mathrm{T}}b}{\lambda_{max}\left(A^{\mathrm{T}}A\right)}
$$

$$
x^{t+1} = prox_{\frac{1}{\lambda_{max}(A^{\mathrm{T}}A)}\left(\lambda||x_1 + (1-\lambda)||x||_2^2\right)} \left(\frac{1}{1 + \frac{2(1-\lambda)}{\lambda_{max}(A^{\mathrm{T}}A)}}\right) sign(z^t) \left(|z| - \frac{\lambda}{\lambda_{max}(A^{\mathrm{T}}A)}\right)
$$

$$
t = t + 1
$$

$$
f(\hat{x^t}) = min(f(\hat{x^{t-1}}), f(x^t))
$$

if $\quad f(\hat{x^t}) = f(\hat{x^{t-1}})$:

$$
\hat{x} = x^t
$$

end if

4: end while

5: 输出: x^T , 即循环最后的 $\hat{x}$

5 问题五

5.1 问题重述

请参阅 lexl 书的第 242 页和 readme 文件:

(1) 通过次梯度下降算法再现 LASSO 问题的结果 >

(2) 使用近端梯度下降算法求解 LASSO.

(3) 比较对不同的 λ 而言 LASSO 算法和岭回归算法.

5.2 代码说明

根据题目要求，第一二问分别采用次梯度下降法和 proximal gradient descent algorithm 解决 Lasso 问题，并绘制迭代图像，下面将展示并描述部分运行结果，在附录部分展示代码及所有结果

5.2.1 解（1）（2）

首先计算可以知道 λ=1 正确结果 $f\star$=3.59 左右

$\star$ 次梯度下降法,alpha=0.0005，最终迭代结果为: 15232，由附录的代码可知，最终发散为 nan

$\star$ 次梯度下降法,alpha=0.0002，最终迭代结果为: 15232, 同上，发散为 nan

$\star$ 次梯度下降法,alpha=0.0001，最终迭代结果为: 3, 收敛

$\star$ 次梯度下降法，消失梯度 alpha=0.02/sqrt(k)，最终迭代结果为: 15232, 同上，发散为 nan

$\star$ 近端梯度下降法，最终迭代结果为: 3.592841659833062，收敛

图片展示如下：

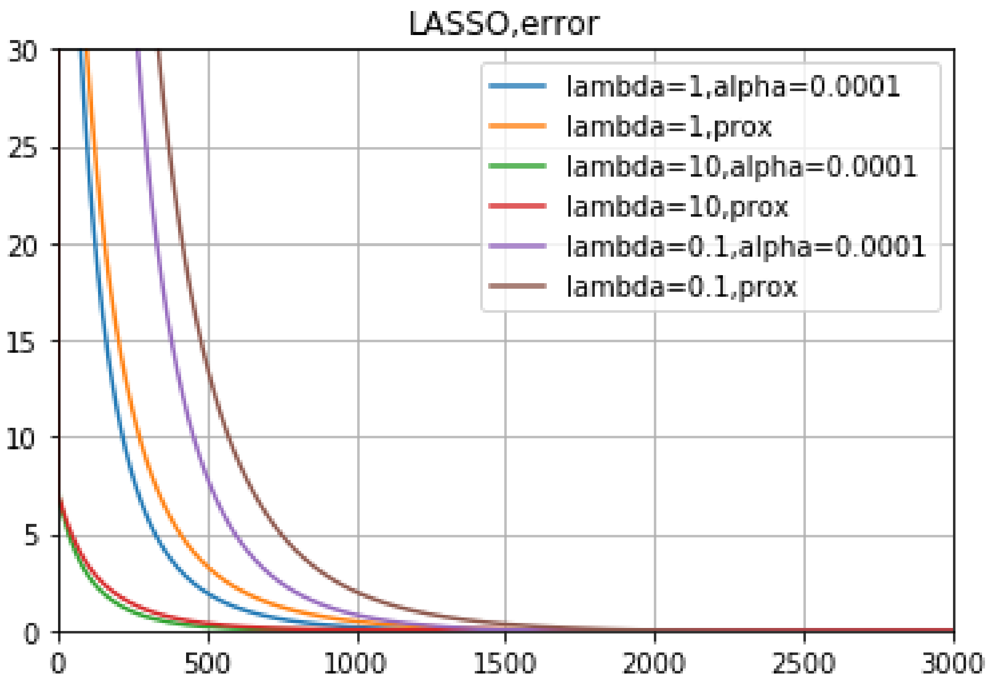

Figure 1: 残差总图像

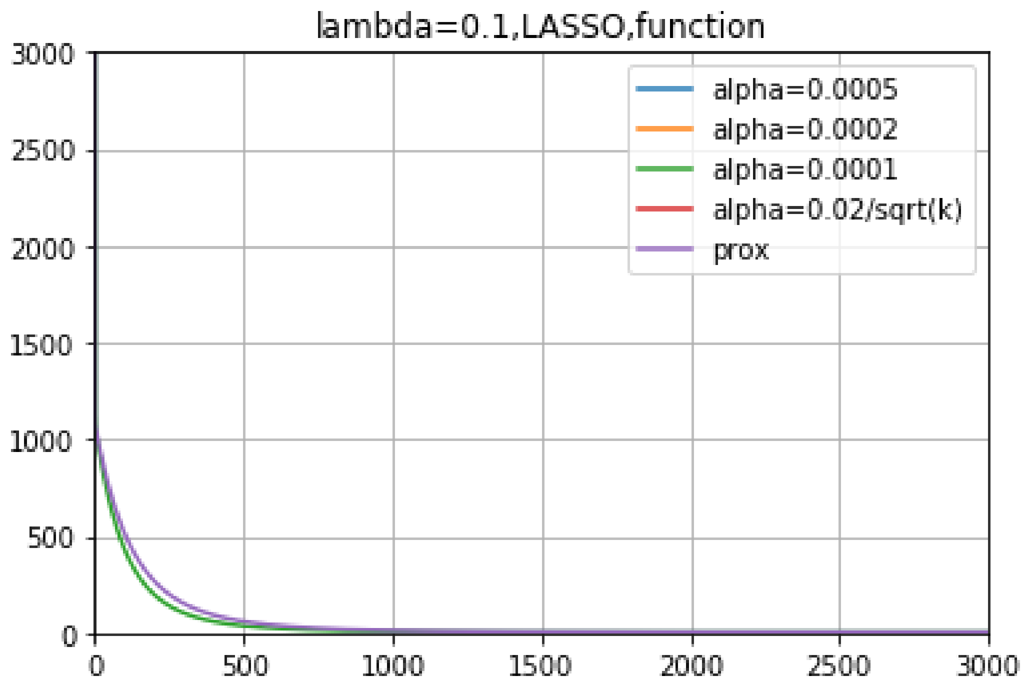

Figure 2: Lambda=0.1 时次梯度下降法和近端梯度下降算法函数图

- ⋆ 首先计算可以知道 λ=1 正确结果 $f\star$=17 左右
- ⋆ 次梯度下降法,alpha=0.0005，最终迭代结果为: 15232, 发散为 nan
- ⋆ 当采用次梯度下降法,alpha=0.0002，最终迭代结果为: 15232, 发散为 nan
- ⋆ 当采用次梯度下降法,alpha=0.0001，最终迭代结果为: 17, 收敛
- ⋆ 次梯度下降法，消失梯度 alpha=0.02/sqrt(k)，最终迭代结果为: 15232, 同上，发散为 nan
- ⋆ 当采用近端梯度下降法，最终迭代结果为: 17, 收敛

- ⋆ 当 λ=10 时，正确结果 $f\star$=152 左右
- ⋆ 当采用次梯度下降法,alpha=0.0005，最终迭代结果为: 15232，事实上发散为 nan
- ⋆ 当采用次梯度下降法,alpha=0.0002, 最终迭代结果为: 15232, 事实上发散为 nan
- ⋆ 当采用次梯度下降法,alpha=0.0001, 最终迭代结果为: 152., 收敛
- ⋆ 次梯度下降法，消失梯度 alpha=0.02/sqrt(k), 最终迭代结果为: 15232, 同上，发散为 nan
- ⋆ 当采用近端梯度下降法，最终迭代结果为:151，收敛

根据图片展示结果，可以看出在 $\alpha = 0.0001$ 或使用近端梯度下降法时，迭代收敛，当然普遍来看选取固定 α 可以得到更加快速收敛的速度，但是并不能收敛到最低，只有 prox 完成了随着迭代步数增加收敛越来越靠近正确结果的任务。当 λ 取值越大，迭代序列与全局最优解的残差 $\frac{f(x^k)-f^*)}{f*}$ 越小，迭代序列越接近全局最优，收敛效果越好。

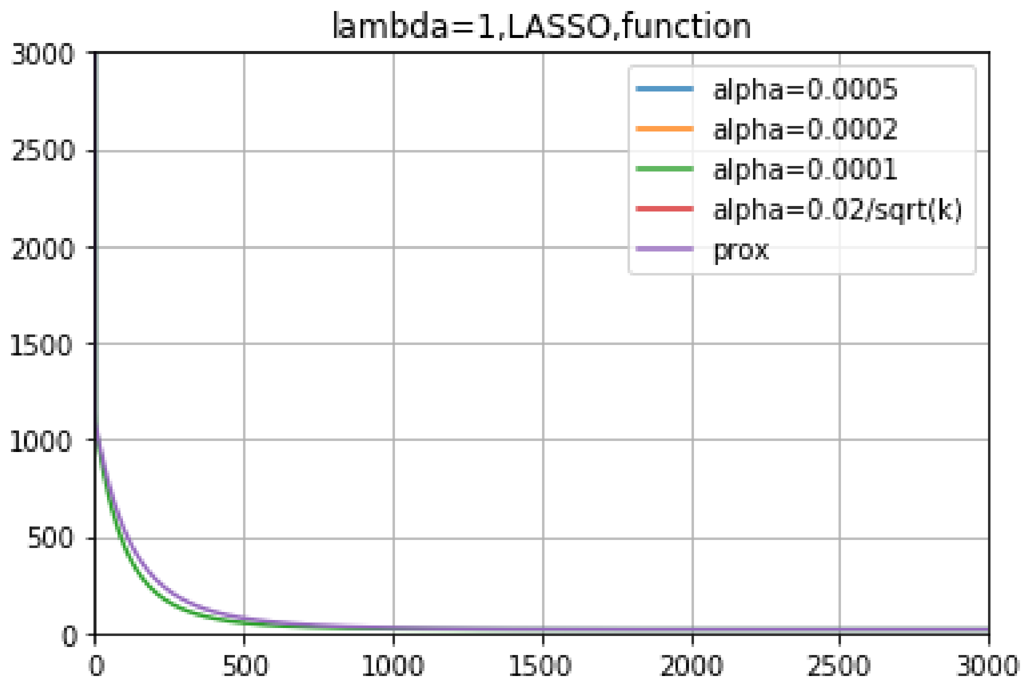

Figure 3: Lambda=1 时次梯度下降法和近端梯度下降算法残差结果图

5.3 解（3）

比较岭回归和 lasso 的不同结果

- ⋆ λ= 0.1, 最终迭代结果为: 9863。
- ⋆ λ= 1, 最终迭代结果为: 9864。
- ⋆ λ= 10, 最终迭代结果为: 9878。

函数迭代相对残差结果见下图：

岭回归相比较 lasso 问题而言，使用梯度下降法收敛比较慢，但是在 3000 步之内都在收敛，而且可以看到当 lambda 越小，收敛速度越慢。

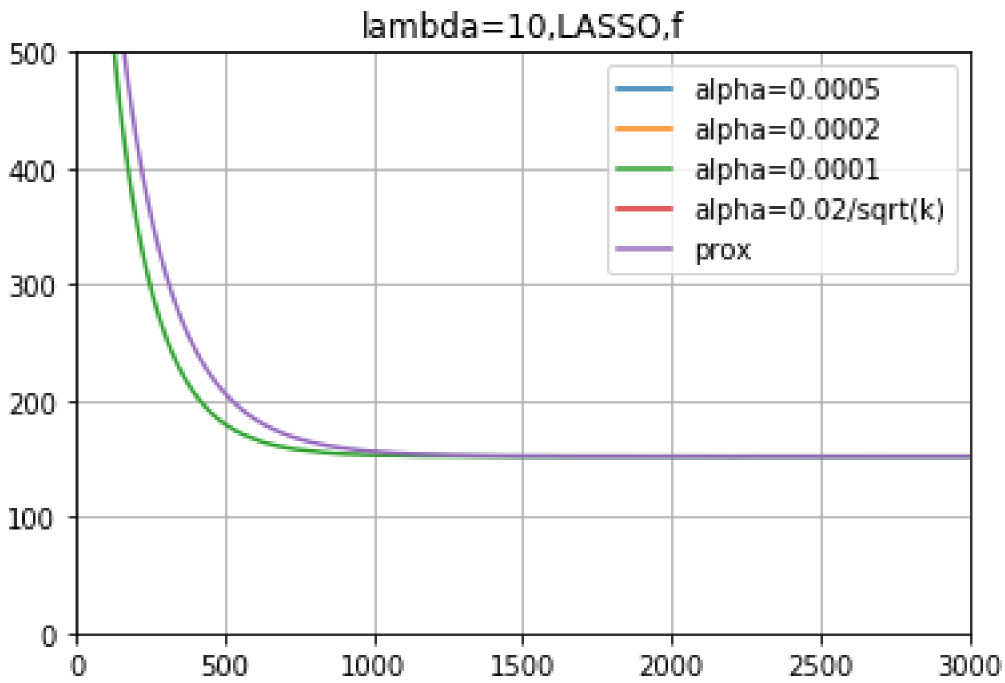

Figure 4: Lambda=10 时次梯度下降法和近端梯度下降算法残差结果图

参考文献

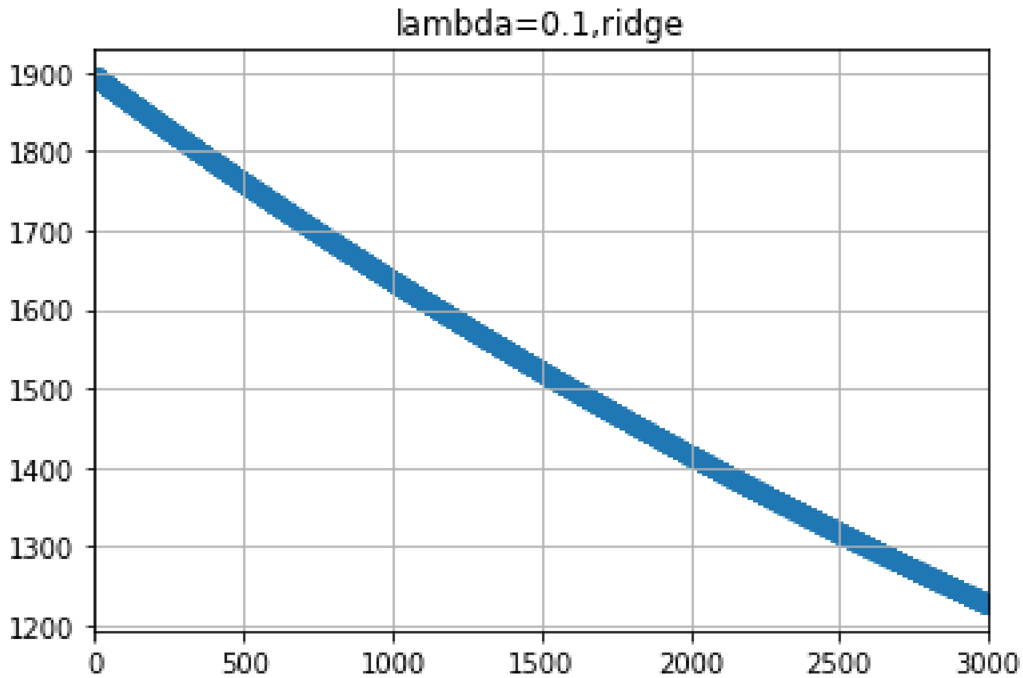

Figure 5: Lambda=0.1 时岭回归算法结果图

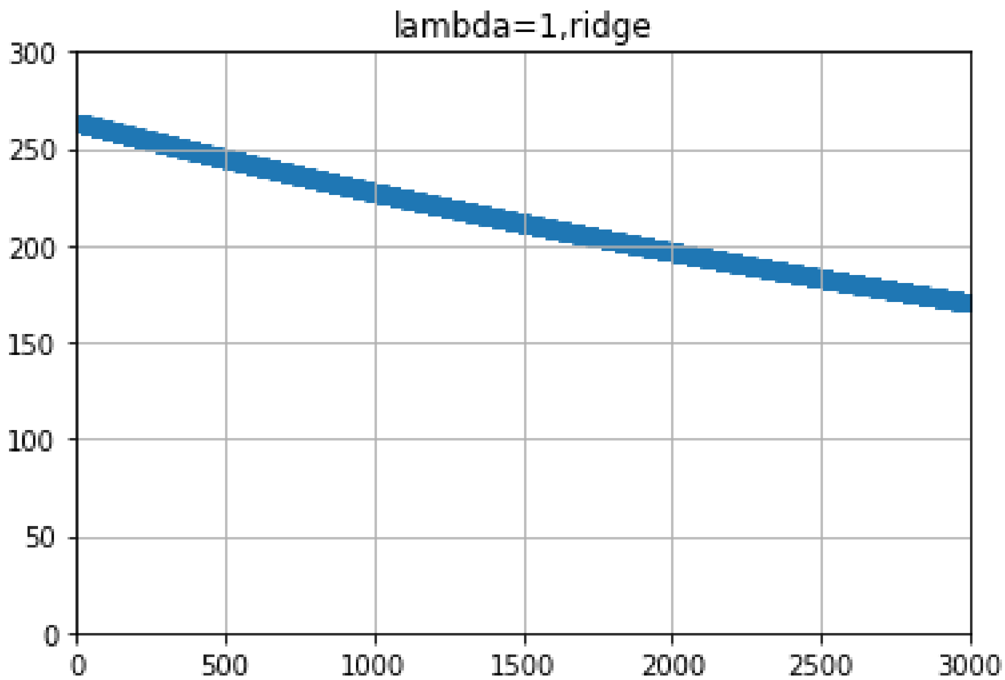

Figure 6: Lambda=1 时岭回归算法结果图

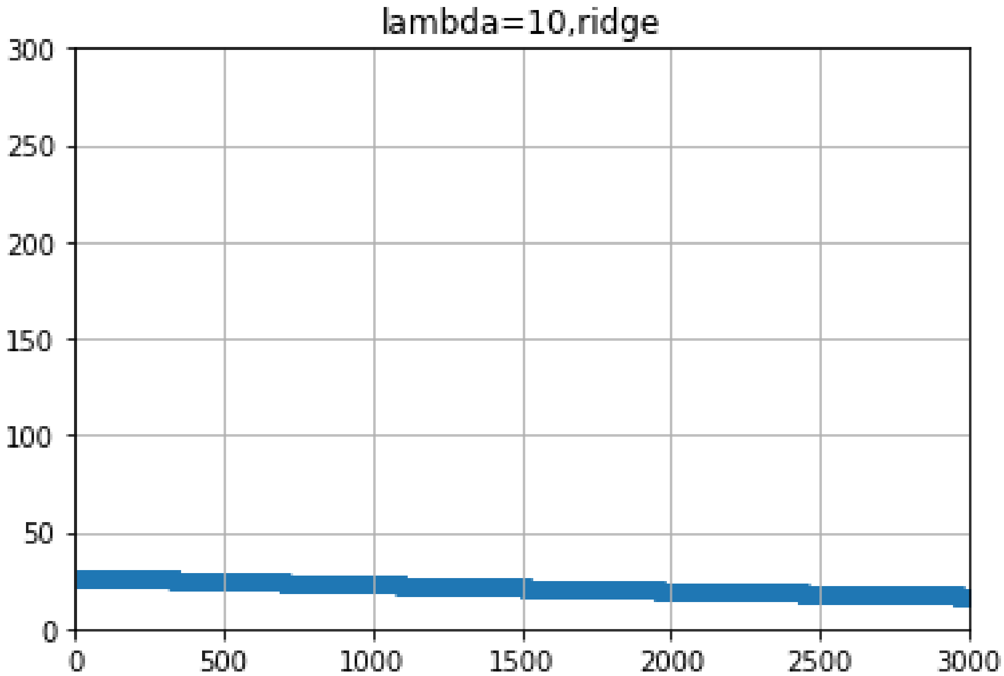

Figure 7: Lambda=10, 岭回归算法