机器学习中的优化算法作业3 (6.2, 6.4, 7.2, 7.6, 7.9, 8.3, 8.4)

谢泽钰: 2020012544

6.2 $f(x^k+\alpha d^k)$ 关于 α 强凸, 由一阶条件知

$$\frac{d}{d\alpha} f(x^k+\alpha_k d^k) = \alpha_k (d^k)^T A d^k + (x^k)^T A d^k + b^T d^k = 0$$

$$\Rightarrow \text{步长为 } \alpha_k = -\frac{(x^k)^T A d^k + b^T d^k}{(d^k)^T A d^k} = -\frac{(Ax+b)^T d^k}{(d^k)^T A d^k}$$

在最速下降法中 $d^k = -\nabla f(x^k) = -(Ax^k+b)$ 代入上式:

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^T A \nabla f(x^k)}$$

6.4 (a) 原问题 $\Leftrightarrow \min\limits_{x,t} \; t + \frac{1}{2}\|x\|^2$, s.t. $x_i \le t$, $i=1,2,\cdots,k$

设 $\lambda_i \ge 0$ 为 $x_i \le t$ 对应的乘子. 其 KKT 条件:

$$x_i = 0, \quad i = k+1, \cdots, n$$

$$x_i + \lambda_i = 0 \qquad \lambda_i(t - x_i) = 0 \qquad i=1,2,\cdots,k$$

$$t \ge x_i \qquad \lambda_i \ge 0 \qquad i=1,2,\cdots,k$$

$$\sum_{i=0}^{k} \lambda_i = 1$$

$\Rightarrow \sum\limits_{i=1}^{K} x_i^2 = -t$, $t \ge -\frac{1}{k}$ 取等 $\Leftrightarrow x_i^* = -\frac{1}{k}$, $i=1,2,\cdots,k$

$x_i^* = 0$, $i = k+1,\cdots,n$, 最小值 $f^* = -\frac{1}{2k}$

(b) $\max_{1\le i\le k} x_i \le \max_{1\le i\le k}(x_i-y_i)+\max_{1\le i\le k} y_i \le \max_{1\le i\le k} y_i + \|x-y\|_2$

当 $\|x\|$ 和 $\|y\| \le \sqrt{k}$:

$$f(x)-f(y)\le \|x-y\|_2+\frac{1}{2}(x+y)^T(x-y)\le \left(1+\frac{1}{\sqrt{k}}\right)\|x-y\|_2$$

(c) 因 $x^0=0$，故 $g^0=e_1$

$\Rightarrow x_1^1<0$， $x_2^1, x_3^1, \cdots$ 均为 $0 \Rightarrow g^1=e_2$

故由归纳法 $x^k\in \text{span}\{e_1,\cdots,e_k\}$

因 $k<K$，故 x^k 的第 K 个元素为 0，此时

$$f^k-f^*\ge x_K^k+\frac{1}{2}\|x^k\|^2+\frac{1}{2(1+R)}\ge \frac{1}{2(1+R)}\ge \frac{GR}{2(1+\sqrt{K})}$$

7.2 二次罚函数： $P_E(x,6)=-x_1x_2x_3+\frac{6}{2}(x_1+2x_2+3x_3-60)^2$

可行解取 $x=(10,10,10)^T$ 时 $P_E(x,6)=-1000<0$

当 x_i 中存在 0 时，最小值大于 $0 \Rightarrow$ 最小值坐标均非 0.

又对 $z>0$ 取 $x=(-z,-z,z+20)$ 为可行解

当 $z\to\infty$ 时，原问题、罚问题均无界 $\Rightarrow$ 无解.

故只需考虑局部极小解的存在性

由罚问题一阶最优条件：

$$x_2x_3=\frac{x_3x_1}{2}=\frac{x_1x_2}{3}$$

因 $x_1,x_2,x_3\neq 0$，故 $2x_2=x_1$，$3x_3=x_1$

$\Rightarrow$ 罚问题最优性条件 $\Leftrightarrow -x_1^2+18\sigma(x_1-20)=0$

$\Delta = 81\sigma^2-360\sigma>0$ 时有极小值点，此时

$$x_1(\sigma)=9\sigma-\sqrt{81\sigma^2-360\sigma},\quad x_2(\sigma)=\frac{1}{2}x_1(\sigma),\quad x_3(\sigma)=\frac{1}{3}x_1(\sigma)$$

故当 $\sigma\to\infty$，一个局部极小解为

$$x_1=\lim_{\sigma\to\infty}\frac{360\sigma}{9\sigma+\sqrt{81\sigma^2-360\sigma}}=20,\quad x_2=10,\quad x_3=\frac{20}{3}$$

Lagrange乘子为

$$\lambda=-\lim_{\sigma\to\infty}\sigma(x_1(\sigma)+2x_2(\sigma)+3x_3(\sigma)-60)$$

$$=-60\lim_{\sigma\to\infty}\sigma\left(\frac{6\sigma}{3\sigma+\sqrt{9\sigma^2-40\sigma}}-1\right)$$

$$=-60\lim_{\sigma\to\infty}\frac{40\sigma^2}{(3\sigma+\sqrt{9\sigma^2-40\sigma})^2}$$

$$=-\frac{200}{3}$$

又由 $\nabla^2_{xx}P_E(x,\sigma)=\sigma(1,2,3)^T(1,2,3)-\begin{pmatrix}0 & x_3 & x_2\\ x_3 & 0 & x_1\\ x_2 & x_1 & 0\end{pmatrix}$

知罚因子的值由 $\nabla^2_{xx}P_E(x(\sigma),\sigma)$ 的正定性所确定

7.6 (a) 反证法：若 $f(x^k) > f(x^*)$. 则 $v(M, x^*) < v(M, x^k)$

与题设矛盾！

(b) $v(M_k, x^k) \le v(M_k, x^*) \Rightarrow M_{k+1} = M_k + \sqrt{v(M_k, x_k)}$

$\le M_k + \sqrt{v(M_k, x^*)} = M_k + |f(x^*) - M_k|$

而 $f(x^*) \ge M_k \Rightarrow M_{k+1} \le M_k + f(x^*) - M_k = f(x^*)$

(c) 考虑 Morrison 方法中 $x^k \to x^*$ $(k \to \infty)$

则 $M_{k+1} = M_k + \sqrt{v(M_k, x^k)}$

$= M_k + \sqrt{(f(x^k) - M_k)^2 + \sum_{i \in E} c_i^2(x^k)}$

$\exists N \in \mathbb{N}_+$, $\forall n \ge N$, 成立

$M_{n+1} = M_n + \sqrt{(f(x^n) - M_n)^2 + \varepsilon_n} \le M_n + |f(x^n) - M_n| + \varepsilon_n$

其中 $\varepsilon_n > 0$ 且 $n \to \infty$ 时 $\varepsilon_n \to 0$

同理 $M_{n+1} \ge M_n + |f(x^n) - M_n| - \varepsilon_n$

令 $n \to \infty$ 且注意 $f(x^n) \to f(x^*)$. 则有

$$\lim_{k \to \infty} M_{k+1} = \lim_{k \to \infty} M_k = f(x^*)$$

(d) Hessen 矩阵的 (i,j) 元为:

$$2\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}+2(f-M)\frac{\partial^2 f}{\partial x_i\partial x_j}+2\sum_{k\in\varepsilon}\frac{\partial c_k}{\partial x_i}\frac{\partial c_k}{\partial x_j}+2\sum_{k\in\varepsilon}c_k\frac{\partial^2 c_k}{\partial x_i\partial x_j}$$

与算法 7.1 的联系: Morrison 方法仍为惩罚方法

但 7.1 通过调节惩罚项 $\sum_{i\in\varepsilon}c_i^2(x)$ 的的系数 σ_k 的大小施加惩罚

而 Morrison 是惩罚自适应方法, 构造 $\min_x v(M_k,x)$ 完成优化

7.9. (a) Lagrange 函数为 $L_\sigma(x,\lambda)=c^Tx+\lambda^T(Ax-b)+\frac{1}{2}\sigma\|Ax-b\|_2^2,\quad x\geq 0$

增广 Lagrange 函数法迭代格式:

$$x^{k+1}=\arg\min_{x\geq 0}\{L_{\sigma_k}(x,\lambda^k)\}$$

$$\lambda^{k+1}=\lambda^k+\sigma_k(Ax^{k+1}-b)$$

$$\sigma_{k+1}=\min\{\rho\sigma_k,\bar{\sigma}\}$$

线性规划的对偶问题:

$$\min_y\ b^Ty\quad s.t.\ A^Ty+c\leq 0$$

引入松弛变量 s 上面等价于.

$$\min_y\ -b^Ty\quad s.t.\ A^Ty+s-c=0,\ s\geq 0$$

对偶问题的增广 Lagrange 函数:

$$L_\sigma(y,s,\lambda)=-b^Ty+\lambda^T(A^Ty+s-c)+\frac{1}{2}\sigma\|A^Ty+s-c\|_2^2 \quad (s\geq 0)$$

$\Rightarrow$ 迭代格式:

$$(y^{k+1},s^{k+1})=\underset{y,s\geq 0}{\arg\min}\{L_{\sigma_k}(y,s,\lambda^k)\}$$

$$\lambda^{k+1}=\lambda^k+\sigma_k(A^Ty^{k+1}+s^{k+1}-c)$$

$$\sigma_{k+1}=\min\{\rho\sigma_k,\bar{\sigma}\}$$

$\Rightarrow$ $s=P_{\mathbb{R}^n_+}(c-A^Ty-\frac{\lambda}{\sigma})$ 代入即可

(b) 仅考虑 $\lambda^k\geq 0$ 情况

首先取 $\lambda^0=\lambda^0=0$, 设 x^{k+1} 为 $L_\sigma(x,\lambda^k)$ 全局极小点则

$$0\in c+\sigma A^T(Ax^{k+1}-b+\frac{\lambda^k}{\sigma})$$

故满足引理1 引理2

由线性规划可行域 $\neq\emptyset \Rightarrow \exists\hat{x}$ s.t $\|A\hat{x}-b\|=0$

由引理2及 c^Tb 为凸 $\Rightarrow \exists K$, s.t. $\forall k\geq K$. 有:

$$\frac{\sigma}{2}\|Ax^k-b\|^2\leq c^T\hat{x}-c^Tx^k+\lambda^kA(\hat{x}-x^k)+\frac{\sigma}{2}\|A\hat{x}-b\|^2\leq 0$$

故对 $\forall k\geq K$, $Ax^k=b \Rightarrow$ 有限终止性

8.3 (a) $u = \arg\min_u \{ I_C(u) + \frac{1}{2}\|u-x\|^2 \}$

$$= \arg\min_{\|u\|_2 \le t} \{ \|u-x\|^2 \}$$

$$= P_{\|x\|_2 \le t}(x)$$

(b) $f(x) = \inf_{y \in C} |x-y|$，先求 $f(x)$ 在 $\hat{x}$ 处次梯度

若 $f(\hat{x}) = 0$ 则 $g = 0$

若 $f(\hat{x}) > 0$. 取 $\hat{y}$ 为 $\hat{x}$ 上的投影，即 $\hat{y} \in P_C(\hat{x})$

次梯度为 $g \in \partial \|x - P_C(\hat{x})\| \subseteq \partial f(x)$

特别地，若 $\|\cdot\| \triangleq \|\cdot\|_2$ 则 $\hat{x} \neq P_C(\hat{x})$ 时：

$$g = \frac{\hat{x} - P_C(\hat{x})}{\|\hat{x} - P_C(\hat{x})\|_2} \in \partial f(\hat{x})$$

故设 $u = \mathrm{prox}_f(x)$. 则 $x - u \in \partial f(u)$

$$\Rightarrow x - u \in \begin{cases} \partial \|u - P_C(u)\| & f(u) > 0 \\ \{0\} & f(u) = 0 \end{cases}$$

$\Rightarrow x = y$ 时 $u = x$；$x \neq y$ 时 u 满足 $x - u \in \|u - P_C(u)\|$

(c) 若 $f(\hat{x}) = 0$ 则 $g = 0$，否则取 $\hat{y} = P_C(\hat{x})$ 则

$$g \in \partial \left(\frac{1}{2} \|\hat{x} - P_C(x)\|^2 \right)$$