最优化算法作业 1

陈文轩

更新：June 15, 2025

1. 判断以下集合是否是凸集

- 考虑这样点的集合，这些点离给定点 x_0 比离给定集合 S 中的任何点都更近，即集合 $\{x \mid ||x-x_0||_2 \le ||x-y||_2, \forall y \in S\}, S \subset \mathbb{R}^n$。

  记 $A = \{x \mid ||x-x_0||_2 \le ||x-y||_2, \forall y \in S\}, A_y = \{x \mid ||x-x_0||_2 \le ||x-y||_2\}$，由于

  $$\begin{aligned} &||x-x_0||_2 \le ||x-y||_2 \Longleftrightarrow ||x-x_0||_2^2 \le ||x-y||_2^2 \\ \Longleftrightarrow &||x||_2^2 - 2\langle x, x_0\rangle + ||x_0||_2^2 \le ||x||_2^2 - 2\langle x, y\rangle + ||y||_2^2 \\ \Longleftrightarrow &\langle x, y-x_0\rangle \le \frac{1}{2}||y||_2^2 - ||x_0||_2^2 \end{aligned}$$

  因此 A_y 是闭的半空间，显然是凸集。因此 $A = \bigcap\limits_{y\in S} A_y$ 也是凸集。

- 记 $n \times n$ 的对称矩阵集合为 $\mathbb{S}^n$，集合 $\{X \in \mathbb{S}^n \mid \lambda_{\min}(X) \ge 1\}$。

  $\{X \in \mathbb{S}^n \mid \lambda_{\min}(X) \ge 1\} = \{X \in \mathbb{S}^n \mid X - I_n \succeq 0\} = \mathbb{S}_+^n + I_n$，这是一个凸锥的平移，因此也是凸集。

2. 判断以下函数是否是凸函数

- 函数 $f(x) = \sum\limits_{i=1}^{r} |x|_{[i]}$ 在 $\mathbb{R}^n$ 上定义，其中向量 $|x|$ 的分量满足 $|x|_i = |x_i|$(即 $|x|$ 是 x 的每个分量的绝对值)，而 $|x|_{[i]}$ 是 $|x|$ 中第 i 大的分量。换句话说，$|x|_{[1]} \ge |x|_{[2]} \ge \cdots \ge |x|_{[n]}$ 是 x 的分量的绝对值按非增序排序。

  显然 $\forall i$，$g_i(x) = |x_i|$ 是凸函数，因此 $g_i(x)$ 的任意非负系数线性组合也是凸函数。又 $f(x) = \max\limits_{\substack{I\subset\{1,\cdots,n\}\\|I|=r}} \sum\limits_{i\in I} g_i(x)$ 是有限个凸函数逐点取最大值，故 $f(x)$ 是凸函数。

- 若 f, g 都是凸函数，并且都非递减，而且 f, g 函数值都是正的。那么他们的乘积函数 $h = fg$ 是否为凸函数？

由 f,g 凸，有$\forall x,y\in\mathbb{R},\lambda\in[0,1],f(\lambda x+(1-\lambda)y)\leq\lambda f(x)+(1-\lambda)f(y)$,
$g(\lambda x+(1-\lambda)y)\leq\lambda g(x)+(1-\lambda)g(y)$。又由 f,g 函数值为正，有
$f(\lambda x+(1-\lambda)y)g(\lambda x+(1-\lambda)y)\leq(\lambda f(x)+(1-\lambda)f(y))(\lambda g(x)+(1-\lambda)g(y))$,
即$h(\lambda x+(1-\lambda)y)\leq(\lambda f(x)+(1-\lambda)f(y))(\lambda g(x)+(1-\lambda)g(y))$.

又由 f,g 非递减，有

$$
\begin{aligned}
&(\lambda f(x)+(1-\lambda)f(y))(\lambda g(x)+(1-\lambda)g(y))-(\lambda h(x)+(1-\lambda)h(y))\\
=&(\lambda f(x)+(1-\lambda)f(y))(\lambda g(x)+(1-\lambda)g(y))-(\lambda f(x)g(x)+(1-\lambda)f(y)g(y))\\
=&\lambda^2 f(x)g(x)+\lambda(1-\lambda)(f(y)g(x)+f(x)g(y))+(1-\lambda)^2 f(y)g(y)-\lambda f(x)g(x)-\\
&(1-\lambda)f(y)g(y)\\
=&\lambda(1-\lambda)(f(x)g(y)+f(y)g(x)-f(x)g(x)-f(y)g(y))\\
=&-\lambda(1-\lambda)(f(x)-f(y))(g(x)-g(y))\leq 0
\end{aligned}
$$

故 $h(\lambda x+(1-\lambda)y)\leq\lambda h(x)+(1-\lambda)h(y)$，即 $h(x)$ 是凸函数。

3. 对于最大分量函数 $f(x)=\max\limits_{1\leq i\leq n}x_i,x\in\mathbb{R}^n$，证明其共轭函数为

$$
f^*(y)=\begin{cases}0 & y\geq 0,\sum_i y_i=1\\ \infty & \text{otherwise}\end{cases}
$$

$f^*(y)=\sup\limits_{x\in\mathbb{R}^n}(y^Tx-\max\limits_{1\leq i\leq n}x_i)$。考虑以下情形：

- $y\geq 0,\sum\limits_{i=1}^n y_i>1$，此时取 $x=t\mathbb{1},t\to+\infty$ 时 $f^*(y)=t\left(\sum\limits_{i=1}^n y_i-1\right)\to\infty$。
- $y\geq 0,\sum\limits_{i=1}^n y_i<1$，此时取 $x=t\mathbb{1},t\to-\infty$ 时 $f^*(y)=-t\left(1-\sum\limits_{i=1}^n y_i\right)\to\infty$。
- y 的某个分量 $y_i<0$，此时取 $x=te_i,t\to-\infty$ 时 $f^*(y)=ty_i\to\infty$。
- $y\geq 0,\sum\limits_{i=1}^n y_i=1$，此时取 $x=t\mathbb{1},f^*(y)=t\left(\sum\limits_{i=1}^n y_i-1\right)=0$。又有

$$
\begin{aligned}
&y^Tx-\max_{1\leq i\leq n}x_i=\sum_{j=1}^n x_jy_j-\max_{1\leq i\leq n}x_i\leq\sum_{j=1}^n y_j\max_{1\leq i\leq n}x_i-\max_{1\leq i\leq n}x_i\\
&=\max_{1\leq i\leq n}x_i\left(\sum_{i=1}^n y_i-1\right)=0
\end{aligned}
$$

故 $f^*(y)=\sup\limits_{x\in\mathbb{R}^n}(y^Tx-\max\limits_{1\leq i\leq n}x_i)=0$

由上，即有 $f^*(y)=\begin{cases}0 & y\geq 0,\sum_i y_i=1\\ \infty & \text{otherwise}\end{cases}$

4. 对于分式线性问题

$$
\begin{aligned}
\min\quad & f_0(x)\\
s.t.\quad & Gx\leq h,Ax=b
\end{aligned}
$$

其中分式线性函数：

$$f_0(x)=\frac{c^Tx+d}{e^Tx+f},\mathrm{dom}\, f_0(x)=\{x\mid e^Tx+f>0\}$$

证明该问题等价于一个线性规划问题：

$$\begin{aligned}\min\quad & c^Ty+dz\\ s.t.\quad & Gy\le hz\\ & Ay=bz\\ & e^Ty+fz=1\\ & z\ge 0\end{aligned}$$

令 $z=\frac{1}{e^Tx+f},y=xz$，此时显然有 $z>0,x=\frac{y}{z},f_0(x)=\frac{c^T\frac{y}{z}+d}{\frac{1}{z}}=c^Ty+dz$。此时有

$$Gy\le hz\Longleftrightarrow G\frac{y}{z}\le h\Longleftrightarrow Gx\le h,\quad Ay=bz\Longleftrightarrow A\frac{y}{z}=b\Longleftrightarrow Ax=b$$

$$e^Ty+fz=1\Longleftrightarrow e^Txz+fz=1\Longleftrightarrow e^Tx+f=\frac{1}{z}\Longleftrightarrow z=\frac{1}{e^Tx+f}$$

因此两个问题等价。

5. 对于 $i=1,\cdots,m$，令 B_i 是 $\mathbb{R}^n$ 中的球体，它的球心和半径分别是 x_i 和 ρ_i。我们希望找到 $B_i,i=1,\cdots,m$ 的最小外接球，即找到一个球 B，使得 B 包含所有 B_i，并且 B 的半径最小。将这个问题写为一个 SOCP 问题。

直接写出 SOCP 问题即可：

$$\begin{aligned}\min_{c\in\mathbb{R}^n,R\in\mathbb{R}}\quad & R\\ s.t.\quad & R-\rho_i\ge 0,i=1,\cdots,m\\ & ||x_i-c||_2\le R-\rho_i,i=1,\cdots,m\end{aligned}$$

最优化算法作业 2

陈文轩

更新：May 10, 2025

1. 考虑以下问题：

$$\begin{aligned} \min_{x\in\mathbb{R}^2} \quad & x_1^2 + x_2^2 \\ s.t. \quad & (x_1-1)^2 + (x_2-1)^2 \le 2 \\ & (x_1-1)^2 + (x_2+1)^2 \le 2 \end{aligned}$$

其中 $x = (x_1, x_2)^\top \in \mathbb{R}^2$。

(a). 这是一个凸优化问题吗？

(b). 写出此问题的拉格朗日函数。使用 Slater 条件验证在这个问题中是否存在强对偶性。

(c). 写出这个优化问题的 KKT 条件。求出 KKT 点和最优点。

解

(a). 由于目标函数与约束函数均为二次函数，且 Hessian 矩阵为 $2I_2$，故是凸优化问题。

(b). 引入非负 Lagrange 乘子 $\lambda_1, \lambda_2 \ge 0$，可以得到问题的 Lagrange 函数为：

$$L(x, \lambda_1, \lambda_2) = x_1^2 + x_2^2 + \lambda_1((x_1-1)^2 + (x_2-1)^2 - 2) + \lambda_2(x_1-1)^2 + (x_2+1)^2 - 2$$

取 $x = (1, 0)^\top$，显然这个 x 满足严格不等式，故 Slater 条件成立，即具有强对偶性。

(c). KKT 条件为：

$$\begin{cases} \nabla_x L(x, \lambda_1, \lambda_2) = 2(x_1 + (\lambda_1+\lambda_2)(x_1-1), x_2 + (\lambda_1+\lambda_2)(x_2-1)) = 0 \\ \lambda_1((x_1-1)^2 + (x_2-1)^2 - 2) = 0, \lambda_2((x_1-1)^2 + (x_2+1)^2 - 2) = 0 \\ (x_1-1)^2 + (x_2-1)^2 \le 2, (x_1-1)^2 + (x_2+1)^2 \le 2 \\ \lambda_1 \ge 0, \lambda_2 \ge 0 \end{cases}$$

解得 $x_1 = x_2 = \lambda_1 = \lambda_2 = 0$，即 $(0, 0)^\top$ 是 KKT 点，也是最优点。

2. 考虑一个凸的分段线性最小化问题，变量为 $x \in \mathbb{R}^n$

$$\min \max_{i=1,\cdots,m} (a_i^\top x + b_i)$$

其中 $a_i \in \mathbb{R}^n, n_i \in \mathbb{R}$。

(a). 考虑原问题的如下等价问题2.1，推导2.1的对偶问题

$$\begin{aligned} \min \max_{i=1,\cdots,m} \quad & y_i \\ s.t. \quad & a_i^\top x + b_i \le y_i, i = 1, \cdots, m \end{aligned} \tag{2.1}$$

变量为 $x \in \mathbb{R}^n, y \in \mathbb{R}^m$。

(b). 假设我们通过平滑函数逼近目标函数, 逼近函数为：

$$f_0(x) = \log \sum_{i=1}^m \exp(a_i^\top x + b_i)$$

现在我们考虑无约束问题，即 $\min f_0(x)$。证明该问题的对偶问题如下：

$$\begin{aligned} \max \quad & b^\top z - \sum_{i=1}^m z_i \log z_i \\ s.t. \quad & A^\top z = 0, \mathbb{1}^\top z = 1, z \succeq 0. \end{aligned} \tag{2.2}$$

其中 $\mathbb{1}$ 表示全是 1 的向量。

(c). 设问题2.1的最优函数值是 p_1^*，问题2.2的最优函数值是 p_2^*，证明 $0 \le p_2^* - p_1^* \le \log m$。

解

(a). 先转化为线性规划问题：

$$\begin{aligned} \min \quad & t \\ s.t. \quad & a_i^\top x + b_i \le t, i = 1, \cdots, m \end{aligned}$$

Lagrange 函数为 $L(x,t,\lambda) = t + \sum_{i=1}^m \lambda_i (a_i^\top x + b_i - t), \lambda_i \ge 0$，目标是 $\max_\lambda \min_{x,t} L(x,t,\lambda)$。

$\nabla_t L(x,t,\lambda) = 1 - \sum_{i=1}^m \lambda_i = 0 \Rightarrow \mathbb{1}^\top \lambda = 1, \nabla_x L(x,t,\lambda) = \sum_{i=1}^m \lambda_i a_i = 0$，

此时 $L(x,y,\lambda) = \sum_{i=1}^m \lambda_i b_i$。因此对偶问题是：

$$\begin{aligned} \max \quad & \sum_{i=1}^m \lambda_i b_i \\ s.t. \quad & \mathbb{1}^\top \lambda = 1, \sum_{i=1}^m \lambda_i a_i = 0 \\ & \lambda_i \ge 0, i = 1, \cdots, m \end{aligned}$$

(b). 记 $A = (a_i, \cdots, a_m), z_j = \dfrac{\exp(a_j^\top x + b_j)}{\sum_{i=1}^m \exp(a_i^\top x + b_i)}$，则 $\sum_{j=1}^m z_j = 1, z_j \ge 0$。

此时 $f_0(x) = \log \sum_{i=1}^m \exp(a_i^\top x + b_i) = \sum_{i=1}^m z_i (a_i^\top x + b_i) - \sum_{i=1}^m z_i \log z_i := L(x,z)$。

$\nabla_x L(x,z) = \sum_{i=1}^m z_i a_i = 0 \Rightarrow A^\top z = 0$，此时 $L(x,z) = b^\top z - \sum_{i=1}^m z_i \log z_i$。

因此对偶问题是：

$$\begin{aligned} \max \quad & b^\top z - \sum_{i=1}^m z_i \log z_i \\ s.t. \quad & A^\top z = 0, \mathbb{1}^\top z = 1, z \succeq 0. \end{aligned}$$

(c). 记 $M=\max_i(a_i^\top x+b_i)$，则 $\exp(M)\le\sum_{i=1}^m\exp(a_i^\top x+b_i)\le m\exp(M)$。

取对数，即有 $M\le\log\sum_{i=1}^m\exp(a_i^\top x+b_i)\le M+\log m,\forall x$。

对问题2.2的最优解 x_2^*，有 $p_2^*=\log\sum_{i=1}^m\exp(a_i^\top x_2^*+b_i)$，由上有 $\max_i(a_i^\top x_2^*+b_i)\le p_2^*$。

又 $p_1^*=\min_x\max_i(a_i^\top x+b_i)\le\max_i(a_i^\top x_2^*+b_i)\le p_2^*$，故 $p_1\le p_2$。

对问题2.1的最优解 x_1^*，有 $\max_i(a_i^\top x_1^*+b_i)=p_1^*$。由上有 $\log\sum_{i=1}^m\exp(a_i^\top x_1^*+b_i)\le p_1^*+\log m$。又 $p_2^*=\min_x\log\sum_{i=1}^m\exp(a_i^\top x+b_i)\le p_1^*+\log m$，故 $p_2^*\le p_1^*+\log m$。

综上所述有 $0\le p_2^*-p_1^*\le\log m$。

3. 若 $f(x)=\|x\|$ 表示任意范数，$x\in\mathbb{R}^n$，证明次微分集合如下：

$$\partial f(x)=\{g\in\mathbb{R}^n:\|g\|_*\le 1,\langle g,x\rangle=\|x\|\},$$

其中 $\|g\|_*$ 表示对偶范数。

解

由次微分定义，有 $g\in\partial f(x)\Rightarrow f(y)\ge f(x)+\langle g,y-x\rangle,\forall y$，即 $\|y\|\ge\|x\|+\langle g,y-x\rangle,\forall y$。

分以下情况讨论：

- 若 $x=0$，则 $\langle g,x\rangle=\|x\|$ 显然成立；
- 对 $x\ne 0$，取 $y=0$，则 $0\ge\|x\|+\langle g,-x\rangle\Leftrightarrow\langle g,x\rangle\ge\|x\|$；

  取 $y=2x$，则 $2\|x\|\ge\|x\|+\langle g,x\rangle\Leftrightarrow\langle g,x\rangle\le\|x\|$。故 $\langle g,x\rangle=\|x\|$。

而 $g\in\partial f(x)\Rightarrow\|y\|\ge\langle g,y\rangle,\forall y$。令 $v=\dfrac{y}{\|y\|},t=\|y\|$，则条件变为：

$t\|v\|\ge t\langle g,v\rangle\Leftrightarrow\langle g,v\rangle\le 1,\forall\|v\|=1$。由对偶范数定义，这等价于 $\|g\|_*=\sup_{\|v\|=1}\langle g,v\rangle\le 1$。

故 $\{g\in\mathbb{R}^n:\|g\|_*\le 1,\langle g,x\rangle=\|x\|\}\subset\partial f(x)$。

反之，若 $g\in\mathbb{R}^n$ 满足 $\|g\|_*\le 1,\langle g,x\rangle=\|x\|$，

则 $\|y\|\ge\|y\|\cdot\|g\|_*\ge\langle g,y\rangle=\langle g,y-x\rangle+\langle g,x\rangle=\|x\|+\langle g,y-x\rangle,\forall y\Rightarrow g\in\partial f(x)$。

故 $\partial f(x)\subset\{g\in\mathbb{R}^n:\|g\|_*\le 1,\langle g,x\rangle=\|x\|\}\Rightarrow\partial f(x)=\{g\in\mathbb{R}^n:\|g\|_*\le 1,\langle g,x\rangle=\|x\|\}$。

4. 对于 $y\in\mathbb{R}^m$，给定 $\mu-$ 强凸，$L-$ 光滑函数 $g(y)$，即 $\nabla g(y)$ 的 Lipchitz 常数为 L。若 $A\in\mathbb{R}^{m\times n},m\le n$ 是行满秩矩阵，证明：

(a). $f(x)=g(Ax)$ 是 $\bar{L}-$ 光滑的，其中

$$\bar{L}=L\|A\|^2$$

$\|A\|$ 表示矩阵 A 的谱范数。

(b). $f(x)=g(Ax)$ 满足规则性条件，即

$$\langle\nabla f(x),x-x_{\text{proj}}\rangle\ge\bar{\mu}\|x-x_{\text{proj}}\|^2$$

其中 x_{proj} 表示 x 到函数 $f(x)$ 最小值解集合的正交投影点，$\bar{\mu}=\mu\lambda_{\min}(AA^\top)$。

解

(a). 由于 $\nabla f(x) = A^\top \nabla g(Ax)$，有：

$$\begin{aligned}\|\nabla f(x_1) - \nabla f(x_2)\| &= \|A^\top \nabla g(Ax_1) - A^\top \nabla g(Ax_2)\| \le \|A^\top\| \cdot \|\nabla g(x_1) - \nabla g(x_2)\| \\ &\le \|A^\top\| \cdot L \cdot \|Ax_1 - Ax_2\| \le \|A^\top\| \cdot L \cdot \|A\| \cdot \|x_1 - x_2\| \\ &= \|A\|^2 \cdot L \cdot \|x_1 - x_2\| \coloneqq \bar{L}\|x_1 - x_2\|\end{aligned}$$

故 $f(x)$ 是 $\bar{L}-$ 光滑的。

(b). 由 $g(y)\mu-$ 强凸，有 $g(y)$ 有唯一的最小值点 y^*，且 $\nabla f(y^*) = 0$。A 行满秩 $\Rightarrow AA^\top \succ 0$。
此时 $\mu\|y - y^*\|^2 \le \langle \nabla g(y) - \nabla g(y^*), y - y^* \rangle = \langle \nabla g(Ax), Ax - y^* \rangle$。
又 $\forall v \in \mathrm{Row}(A), \|Av\|^2 \ge \lambda_{\min}(AA^\top)\|v\|^2$，$x - x_{\text{proj}} \in \mathrm{Col}(A^\top) = \mathrm{Row}(A)$

$$\begin{aligned}\langle \nabla f(x), x - x_{\text{proj}} \rangle &= \langle A^\top \nabla g(Ax), x - x_{\text{proj}} \rangle = \langle \nabla g(Ax), A(x - x_{\text{proj}}) \rangle \\ &= \langle \nabla g(Ax), Ax - y^* \rangle \ge \mu\|Ax - y^*\|^2 = \mu\|A(x - x_{\text{proj}})\|^2 \\ &\ge \mu\lambda_{\min}(AA^\top)\|x - x_{\text{proj}}\|^2 \coloneqq \bar{\mu}\|x - x_{\text{proj}}\|^2\end{aligned}$$

故 $\langle \nabla f(x), x - x_{\text{proj}} \rangle \ge \bar{\mu}\|x - x_{\text{proj}}\|^2$。

5. 考虑凸函数 $f(x)$ 的共轭函数

$$f^*(y) = \sup_x (x^\top y - f(x))$$

证明：

(a). 若 $x \in \partial f(y)$，则 $y \in \partial f^*(x)$；

(b). 若 $f(x)$ 是闭凸函数，利用 $f = f^{**}$ 证明 $x \in \partial f(y)$ 等价于 $y \in \partial f^*(x)$。

解

(a). 对 $x \in \partial f(y), \forall z, f(z) \ge f(y) + x^\top(z - y) \Rightarrow x^\top z - f(z) \le x^\top y - f(y), \forall z$
显然 $y = z$ 时取等，故 $f^*(x) = \sup_z(z^\top x - f(z)) = x^\top y - f(y)$。$\forall w$，
$f^*(x) + y^\top(w - x) = x^\top y - f(y) + y^\top(w - x) = w^\top y - f(y) \le \sup_z(w^\top z - f(z)) = f^*(w)$。
故 $\forall w, f^*(w) \ge f^*(x) + y^\top(w - x)$，即 $y \in \partial f^*(x)$。

(b). 由上，$x \in \partial f(y) \Rightarrow y \in \partial f^*(x)$，$y \in \partial f^*(x) \Rightarrow x \in \partial(f^*)^*(y) = \partial f^{**}(y) = \partial f(y)$。
故 $x \in \partial f(y) \Leftrightarrow y \in \partial f^*(x)$。

最优化算法作业 3

陈文轩

更新：April 14, 2025

1. • 若 $f(x)$ 是二阶连续可微，证明 $\nabla f(x)$ 是 L–Lipschitz 连续等价于 $LI \succeq \nabla^2 f(x) \succeq -LI$。

解

$\Longrightarrow$: $\forall x, v \in \mathbb{R}^n, \|\nabla^2 f(x)v\| = \lim\limits_{t\to 0} \dfrac{\|\nabla f(x+tv) - \nabla f(x)\|}{t} \le \lim\limits_{t\to 0} \dfrac{L\|tv\|}{t} = L\|v\|$,

故 $\|\nabla^2 f(x))\| \le L$，又 $\nabla^2 f(x)$ 是对称矩阵，故 $\rho(\nabla^2 f(x)) = \|\nabla^2 f(x)\|_2 \le L$，

即 $\nabla^2 f(x)$ 的特征值模长均不大于 L，即 $LI \succeq \nabla^2 f(x) \succeq -LI$。

$\Longleftarrow$: $\forall x, y \in \mathbb{R}^n, \nabla f(y) - \nabla f(x) = \displaystyle\int_0^1 \nabla^2 f(x + t(y-x))(y-x)\,\mathrm{d}t, \|\nabla^2 f(x)\| \le L$,

$\|\nabla f(y) - \nabla f(x)\| \le \displaystyle\int_0^1 \|\nabla^2 f(x+t(y-x))\| \cdot \|y-x\|\,\mathrm{d}t \le \int_0^1 L\|y-x\|\,\mathrm{d}t = L\|y-x\|$

即 $\nabla f(x)$ 是 L–Lipschitz 连续的。

• 估计逻辑回归函数的梯度的 Lipschitz 常数：

$$\min_x l(x) \coloneqq \frac{1}{m}\sum_{i=1}^m \log(1 + \exp(-b_i a_i^\top x))$$

其中 $b_i \in \{-1, 1\}, a_i \in \mathbb{R}^n, i = 1, \cdots, m$ 是给定的数据。

解

记 $f_i(x) = \log(1+e^{-b_i a_i^\top x}), \sigma(x) = \dfrac{1}{1+e^{-x}}, \nabla f_i(x) = \dfrac{-b_i a_i}{1+e^{-b_i a^\top x}} = -b_i a_i \sigma(-b_i a_i^\top x)$。

记 $\sigma_i = \sigma(-b_i a^\top x) \in [0,1]$，则 $\nabla^2 f_i(x) = -b_i a_i(-b_i a_i^\top)\sigma_i(1-\sigma_i) = a_i a_i^\top \sigma_i(1-\sigma_i)$,

$$\|\nabla^2 l(x)\| = \left\|\frac{1}{m}\sum_{i=1}^m \nabla^2 f_i(x)\right\| = \frac{1}{m}\sum_{i=1}^m \|a_i a_i^\top \sigma_i(1-\sigma_i)\| \le \frac{1}{4m}\sum_{i=1}^m \|a_i\|^2$$

故梯度的一个 Lipchitz 常数为 $\dfrac{1}{4m}\displaystyle\sum_{i=1}^m \|a_i\|^2$。

2. 对于次梯度算法，请构造一个非光滑函数例子，说明常数步长不收敛。

解

取 $f(x) = \|x\|$，则 $\partial f(x) = \begin{cases} \{\mathrm{sgn}(x)\}, & x \ne 0 \\ [-1,1], & x = 0 \end{cases}$。对任意步长 α，取初值 $x_0 = \dfrac{\alpha}{2}$，

则迭代序列为 $x_n = (-1)^n \dfrac{\alpha}{2}$ 不收敛。

3. 计算下面函数的邻近点映射，即 $\text{prox}_h(x)=\arg\min\limits_y\left(h(y)+\frac{1}{2}\|y-x\|^2\right)$

- $h(x)=\|x\|_\infty$（需要求解一个一维子问题）。

  解

  对 $h(x)=\|x\|_\infty, h^*(z)=\sup\limits_x(z^\top x-h(x))=I_{\{z\mid\|z\|_1\le 1\}}(z)$。由 $x=\text{prox}_h(x)+\text{prox}_{h^*}(x)$

  考虑 $\text{prox}_{h^*}(x)=\arg\min\limits_z\left(I_{\{z\mid\|z\|_1\le 1\}}(z)+\frac{1}{2}\|z-x\|^2\right)$ 是 x 关于 L^1 范数球的投影。

  而对 $C=\{z\mid\|z\|_1\le 1\}, P_C(x)_k=\begin{cases}x_k-\lambda, & x_k>\lambda\\ 0, & -\lambda\le x_k\le\lambda\\ x_k+\lambda, & x_k<-\lambda\end{cases}$，

  其中对 $\|x\|\le 1,\lambda=0$，否则 λ 是 $\sum\limits_{k=1}^n\max\{|x_k|+-\lambda,0\}=1$ 的解。

  综上所述，$\text{prox}_h(x)=x-P_C(x), P_C(x)$ 如上定义。

- $h(x)=\max\{0,\|x\|_2-1\}$。

  解

  对 $\|y\|_2\le 1, h(y)=0$，此时原问题化为 x 关于 L^1 范数球的投影问题。此时有

  $\text{prox}_h(x)=\arg\min\limits_{\|y\|_2\le 1}\frac{1}{2}\|y-x\|^2=\begin{cases}x, & \|x\|_2\le 1\\ \frac{x}{\|x\|_2}, & \|x\|_2>1\end{cases}$。

  对 $\|y\|_2>1$，问题化为 $\arg\min\limits_{\|y\|_2>1}\left(\|y\|-1+\frac{1}{2}\|y-x\|^2\right)$，由对称性，取 x,y 共线。

  $$\arg\min_{\|y\|_2>1}\left(\|y\|-1+\frac{1}{2}\|y-x\|^2\right)\xrightarrow[y=tu]{u=\frac{x}{\|x\|_2}}u\arg\min_{t>1}\left(t-1+\frac{1}{2}(t-\|x\|_2^2)\right)$$

  $$=\frac{x}{\|x\|_2}\arg\min_{t>1}\left(\frac{1}{2}t^2+(1-\|x\|)t+\frac{1}{4}\|x\|^2-1\right)=\begin{cases}\frac{x}{\|x\|_2}, & \|x\|_2<2\\ \frac{x}{\|x\|_2}(\|x\|_2-1), & \|x\|_2\ge 2\end{cases}$$

  比较两种情况下 $f(y)=\max\{0,\|y\|_2-1\}+\frac{1}{2}\|y-x\|^2$ 的函数值：

  对 $\|x\|_2<1, f(y_1)=0$，$f(y_2)=\frac{1}{2}(\|x\|_2-1)^2$，故使用第一种情况的解 $y=x$。

  对 $1<\|x\|<2$，两种情况下解都为 $y=\frac{x}{\|x\|_2}$，即为最终答案。

  对 $\|x\|>2, f(y_1)=\frac{1}{2}(\|x\|_2-1)^2, f(y_2)=\|x\|_2-\frac{3}{2}, f(y_1)-f(y_2)=\frac{1}{2}(\|x\|_2-2)^2\ge 0$，

  故使用第二种情况的解 $y=\frac{x}{\|x\|_2}(\|x\|_2-1)$。

  综上所述，$\text{prox}_h(x)=\begin{cases}x, & \|x\|_2\le 1\\ \frac{x}{\|x\|_2}, & 1<\|x\|_2\le 2\\ \frac{x}{\|x\|_2}(\|x\|_2-1), & \|x\|_2>2\end{cases}$

4. 考虑 D-最优实验设计（D-optimal experimental design），其目标是最大化估计量的信息内容，通过差分香农熵测量，具体到最大化 $\det V(m_1,\cdots,m_n)$，具体背景参考《convex optimization：7.5 节》。

该问题需要求解下述约束问题：

$$\min_{x\in\Delta_n} -\log\det V(x)$$

其中 $V(x)=\sum_{i=1}^{n} x_i a_i a_i^\top, a_i\in\mathbb{R}^d, i=1,\cdots,d$ 是给定的数据，$\Delta_n=\left\{x:\sum_{i=1}^{n} x_1=1, x\geq 0\right\}$

请使用条件梯度法求解该问题，写出迭代公式，并且给出子问题的解。

解

$\frac{\partial}{\partial x_i}-\log\det V(x)=-\operatorname{tr}\left(\nabla_V \log\det V(x)\frac{\partial V(x)}{\partial x_i}\right)=-\operatorname{tr}(V(x)^{-1}a_i a_i^\top)=-a_i^\top V(x)^{-1}a_i$

$\Rightarrow \nabla(-\log\det V(x))=\left(-a_1^\top V(x)^{-1}a_1,\cdots,-a_n^\top V(x)^{-1}a_n\right)^\top$，记 $f(x)=-\log\det V(x))$，需要求解子问题 $x_k=\arg\min_{x\in\Delta_n}\langle\nabla f(y_{k-1}),x\rangle$，这是一个线性问题，最优解在顶点 e_j 上取。

其中 $j=\arg\min_{x\in\Delta_n}\langle\nabla f(y_{k-1}),e_i\rangle=\arg\min_{i\in\{1,\cdots,n\}}(-a_i^\top V(y_{k-1})a_i)=\arg\max_{i\in\{1,\cdots,n\}} a_i^\top V(y_{k-1})a_i$。

$y_k=(1-\alpha_k)y_{k-1}+\alpha_k y_k$，$\alpha_k$ 取消失步长或通过精确线搜索得到。

5. 求解问题

$$\min f(x)\quad \text{s.t. } x\in\Delta$$

其中，$\Delta_n=\left\{x\in\mathbb{R}^n:\sum_{i=1}^{n} x_1=1, x_n\geq 0\right\}$。使用镜像梯度法，迭代公式为

$$x^{k+1}=\arg\min_{x\in\Delta}\left(\nabla f(x^k)^\top(x-x^k)+\frac{1}{\alpha_k}\sum_{i=1}^{n} x_i\log\frac{x_i}{x_i^k}\right)$$

证明：

$$x_i^{k+1}=\frac{x_i^k\exp\left(-\alpha_k\nabla f(x^k)_i\right)}{\sum_{j=1}^{n} x_j^k\exp\left(-\alpha_k\nabla f(x^k)_j\right)}$$

解

令 $L(x,\lambda)=\nabla f(x^k)^\top(x-x^k)+\frac{1}{\alpha_k}\sum_{i=1}^{n} x_i\log\frac{x_i}{x_i^k}+\lambda\left(\sum_{i=1}^{n} x_i-1\right)$，对 x 求导，

$\frac{\partial L(x,\lambda)}{\partial x_i}=\nabla f(x^k)_i+\frac{1}{\alpha_k}\left(\log\frac{x_i}{x_i^k}+1\right)+\lambda$，求解 $\frac{\partial L(x,\lambda)}{\partial x_i}=0$，得到

$x_i=x_i^k\exp\left(-\alpha_k\nabla(f^k)_i-\alpha_k\lambda-1\right)\xrightarrow{C=\exp(-\alpha_k\lambda-1)}Cx_i^k\exp\left(-\alpha_k\nabla f(x^k)_i\right)\geq 0$。

需要调整 λ 使得 C 满足 $\sum_{i=1}^{n} x_i=1\Rightarrow C\sum_{j=1}^{n} x_j^k\exp\left(-\alpha_k\nabla f(x^k)_j\right)=1$

故 $C=\frac{1}{\sum_{j=1}^{n} x_j^k\exp\left(-\alpha_k\nabla f(x^k)_j\right)}, x_i=\frac{x_i^k\exp\left(-\alpha_k\nabla f(x^k)_i\right)}{\sum_{j=1}^{n} x_j^k\exp\left(-\alpha_k\nabla f(x^k)_j\right)}$ 即为所求。

最优化算法作业 4

陈文轩

更新：June 8, 2025

1. 对于可逆矩阵 A，求解如下方程的零点

$$F(X) = X^{-1} - A = 0$$

可以得到 A^{-1}。

(a). 使用牛顿法，写出迭代公式。

(b). 实现该算法，随机生成 100×100 维可逆矩阵 A，作出误差随着迭代数的收敛图像。

提示：根据 $DF(X)[B] = -X^{-1}BX^{-1}$，计算 $DF(X)^{-1}[B]$

解

(a). 记 Newton 法迭代格式为 $X_{k+1} = X_k + \Delta X$，则 ΔX 满足 $DF(X_k)[\Delta X] = -F(X_k)$，即 $X_k^{-1}\Delta X X_k^{-1} = X_K^{-1} - A, \Delta X = X_k - X_k A X_k$，迭代公式为 $X_{k+1} = 2X_k - X_k A X_k$。

(b). 算法代码如下：

Newton 法代码

```python
import numpy as np
import matplotlib.pyplot as plt
np.random.seed(42); epsilon = 0.09
A = np.random.randn(100, 100)
while np.linalg.det(A) == 0:
    A = np.random.randn(100, 100)
A_inv = np.linalg.inv(A)
X = A_inv + np.eye(100) * epsilon
max_iter = 50; errors = []
for k in range(max_iter):
    X_new = 2 * X - np.dot(np.dot(X, A), X)
    error = np.linalg.norm(X_new - A_inv, 'fro')
    errors.append(error); X = X_new
plt.plot(range(max_iter), errors)
plt.yscale('log'); plt.xlabel('Iteration')
plt.ylabel('Frobenius Norm of Error')
plt.title('Convergence of Newton\'s Method'); plt.show()
```

测试时发现算法对初值极其敏感，`epsilon`=0.1 时便无法收敛。以下是 `epsilon`=0.09 时的收敛图像：

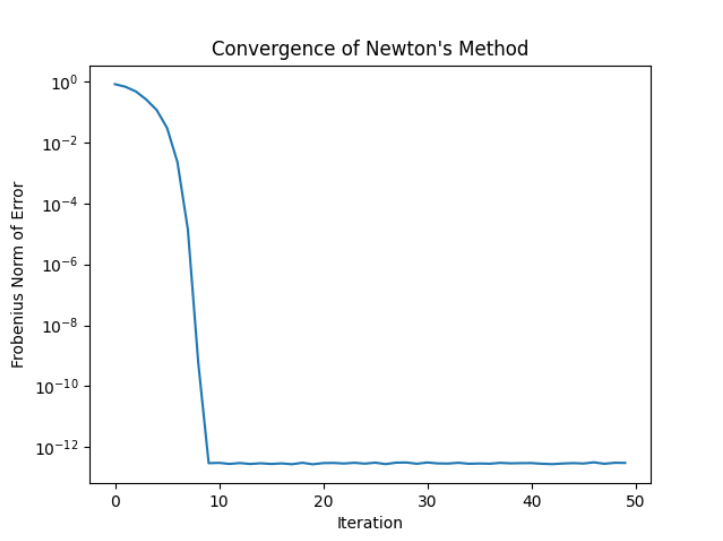

图 1: 牛顿法收敛图像

2. 给定集合 $C_i, i = 1, 2, \cdots, m$ 为闭凸集，且易于计算投影，考虑投影问题：

$$\begin{aligned} \min \quad & \frac{1}{2}\|x - c\|^2 \\ \text{s.t.} \quad & x \in C_1 \cap C_2 \cap \cdots \cap C_m \end{aligned}$$

请使用 ADMM 算法求解此问题，说明是否收敛（无需证明收敛）。

解

把问题重写为以下形式：

$$\begin{aligned} \min_{x,z} \quad & \frac{1}{2}\|z - c\|^2 + \sum_{i=1}^{m} \mathcal{I}_{C_i}(x_i) \\ \text{s.t.} \quad & x_i - z = 0, i = 1, 2, \cdots, m \end{aligned}$$

增广 Lagrange 函数为 $L(x, z, \lambda) = \frac{1}{2}\|z-c\|^2 + \sum_{i=1}^{m} \mathcal{I}_{C_i}(x_i) + \sum_{i=1}^{m} \lambda_i^\top (x_i - z) + \frac{\rho}{2} \sum_{i=1}^{m} \|x_i - z\|^2$，其中 λ_i 是 Lagrange 乘子，ρ 是正的罚参数。对 x, z 做交替极小化：

$$\begin{aligned} x_i^{k+1} &= \arg\min_{x_i} \left(\mathcal{I}_{C_i}(x_i) + \lambda_i^{k\top} (x_i - z^k) + \frac{\rho}{2}\|x_i - z^k\|^2 \right) \\ &= \arg\min_{x_i} \left(\mathcal{I}_{C_i}(x_i) + \lambda_i^{k\top} x_i - \lambda_i^{k\top} z^k + \frac{\rho}{2} \left(\|x_i\|^2 + \|z^k\|^2 - 2x_i^\top z^k \right) \right) \\ &= \arg\min_{x_i} \left(\mathcal{I}_{C_i}(x_i) + \frac{\rho}{2} \left(\|x_i\|^2 + \left\| z^k - \frac{1}{\rho}\lambda_i^{k\top} \right\|^2 - 2x_i^\top \left(z^k - \frac{1}{\rho}\lambda_i^{k\top} \right) \right) \right) \\ &= \arg\min_{x_i} \left(\mathcal{I}_{C_i}(x_i) + \frac{\rho}{2} \left\| x_i - \left(z^k - \frac{1}{\rho}\lambda_i^{k\top} \right) \right\| \right) \\ &= \mathcal{P}_{C_i} \left(z^k - \frac{1}{\rho}\lambda_i^{k\top} \right) \end{aligned}$$

$$z^{k+1}=\arg\min_z\left(\frac{1}{2}\|z-c\|^2+\sum_{i=1}^m\lambda_i^{k\top}(x_i^{k+1}-z)+\frac{\rho}{2}\sum_{i=1}^m\|x_i^{k+1}-z\|^2\right):=\arg\min_z\varphi(z)$$

$$\nabla\varphi(z)=(z-c)-\sum_{i=1}^m\lambda_i^{k\top}+\rho\sum_{i=1}^m(z-x_i^{k+1})=(1+m\rho)z-c-\sum_{i=1}^k\left(\lambda_i^k+\rho x_i^{k+1}\right)=0$$

$$\Longrightarrow z^{k+1}=\frac{c+\sum\limits_{i=1}^k\left(\lambda_i^k+\rho x_i^{k+1}\right)}{1+m\rho},\lambda_i^{k+1}=\lambda_i^k+\tau\rho(x_i^{k+1}-z_i^{k+1}),\tau\in\left(0,\frac{1+\sqrt{5}}{2}\right]$$

这个算法是收敛的。

3. 相关系数矩阵的逼近问题的定义为：

$$\begin{aligned}\min\quad&\frac{1}{2}\|X-G\|_F^2\\ \text{s.t.}\quad&X_{ii}=1,i=1,2,\cdots,n\\&X\succeq 0\end{aligned}$$

其中自变量 X 取值于对称矩阵空间 S_n，G 为给定的实对称矩阵。这个问题在金融领域中有重要的应用。由于误差等因素，根据实际观测得到的相关系数矩阵的估计 G 往往不具有相关系数矩阵的性质（如对角线为 1，正定性），我们的最终目标是找到一个和 G 最接近的相关系数矩阵 X。试给出满足如下要求的算法：

(a). 对偶近似点梯度法，并给出化简后的迭代公式；

(b). 针对原始问题的 ADMM，并给出每个子问题的显式解。

解

(a). Lagrange 函数为 $L(X,y,Z)=\frac{1}{2}\|X-G\|_F^2-y^\top(\operatorname{diag}(X)-\mathbb{1})-\operatorname{tr}(ZX)$。

又 $y^\top\operatorname{diag}(X)=\operatorname{tr}(\operatorname{Diag}(Y)X)$，其中 $Diag(y)$ 是对角元素为 y 的元素的对角矩阵，

故 $L(X,y,Z)=\frac{1}{2}\|X-G\|_F^2-\operatorname{tr}(\operatorname{Diag}(Y)X)-\mathbb{1}^\top y-\operatorname{tr}(ZX)$。

$\nabla_X L(X,y,Z)=X-G-\operatorname{Diag}(y)-Z=0\Rightarrow X=G+Z+\operatorname{Diag}(Y)$，

代回 $L(X,y,Z)$ 得到对偶问题目标函数：

$$\begin{aligned}g(y,Z)=&\frac{1}{2}\|G+\operatorname{Diag}(y)+Z-G\|_F^2-\operatorname{tr}(\operatorname{Diag}(y)(G+\operatorname{Diag}(y)+Z))+y^\top\mathbb{1}\\&-\operatorname{tr}(Z(G+\operatorname{Diag}(y)+Z))\\=&\frac{1}{2}\|\operatorname{Diag}(y)+Z\|_F^2-\operatorname{tr}(\operatorname{Diag}(y)G)-\operatorname{tr}(\operatorname{Diag}(y)^2)-\operatorname{tr}(\operatorname{Diag}(y)Z)+y^\top\mathbb{1}\\&-\operatorname{tr}(ZG)-\operatorname{tr}(Z\operatorname{Diag}(y))-\operatorname{tr}(Z^2)\\=&-\frac{1}{2}\|\operatorname{Diag}(y)\|_F^2-\frac{1}{2}\|Z\|_F^2-\operatorname{tr}(\operatorname{Diag}(y)Z)-\operatorname{tr}(\operatorname{Diag}(y)G)-\operatorname{tr}(ZG)\\&+y^\top\mathbb{1}\\=&-\frac{1}{2}\|\operatorname{Diag}(y)+Z\|_F^2-\operatorname{tr}(G(\operatorname{Diag}(y)+Z))+y^\top\mathbb{1}\\=&-\frac{1}{2}\|\operatorname{Diag}(y)+Z+G\|_F^2+\frac{1}{2}\|G\|_F^2+y^\top\mathbb{1}\end{aligned}$$

故对偶问题是 $\max\limits_{Z\succeq 0,y}\left(-\frac{1}{2}\|\operatorname{Diag}(y)+Z+G\|_F^2+\frac{1}{2}\|G\|_F^2+y^\top\mathbb{1}\right)$。

等价写为 $\min\limits_{y,Z}(f(y,Z)+h(Z)):=\min\limits_{y,Z}\left(\frac{1}{2}\|\operatorname{Diag}(y)+Z+G\|_F^2-\mathbb{1}^\top y+\mathcal{I}_{S_+^n}(Z)\right)$

此时迭代可写为 $(y^{k+1},Z^{k+1})=\operatorname{prox}_{\alpha_k h}\left((y^k,Z^k)-\alpha_k\nabla f(y^k,Z^k)\right)$，$\alpha_k$ 是步长。

先求 ∇f，记 $\Lambda^k=G+\operatorname{Diag}(y^k)+Z^k,\nabla_y f(y^k,Z^k)=\operatorname{diag}(\Lambda^k)-\mathbb{1},\nabla_Z f(y^k,Z^k)=\Lambda^k$。

记 $y_{mid}^{k+1}=y^k-\alpha_k(\operatorname{diag}(\Lambda^k)-\mathbb{1}),Z_{mid}^{k+1}=Z^k-\alpha_k\Lambda^k$，考虑临近点算子的作用：

由于 h 只和 Z 有关，故 $y^{k+1}=y_{mid}^{k+1},Z^{k+1}=\operatorname{prox}_{\alpha_k h}(Z_{mid}^{k+1})=\mathcal{P}_{S_+^n}(Z_{mid}^{k+1})$。

即 $Z_{mid}^{k+1}=V\operatorname{diag}(\lambda_1,\cdots,\lambda_n)V^\top$ 时 $Z^{k+1}=V\operatorname{diag}(\max\{0,\lambda_1\},\cdots,\max\{0,\lambda_n\})V^\top$。

以上给出了一个完整的对偶近似点梯度迭代公式。

(b). 把问题重写为以下形式：

$$\begin{aligned}\min_{X,Z}\quad&\frac{1}{2}\|X-G\|_F^2+\mathcal{I}_{\operatorname{diag}(X)=\mathbb{1}}(X)+\mathcal{I}_{S_+^n}(Z)\\ \text{s.t.}\quad&X-Z=0\end{aligned}$$

此时增广 Lagrange 函数为

$$L(X,Z,\Lambda)=\frac{1}{2}\|X-G\|_F^2+\mathcal{I}_{\operatorname{diag}(X)=\mathbb{1}}(X)+\mathcal{I}_{S_+^n}(Z)+\operatorname{tr}(\Lambda^\top(X-Z))+\frac{\rho}{2}\|X-Z\|_F^2$$

其中 Λ 是 Lagrange 乘子，ρ 是正的罚参数。对 X,Z 做交替极小化：

X_{ii}^{k+1} 直接取为 1，只需 $\arg\min\limits_X\left(\frac{1}{2}\|X-G\|_F^2+\operatorname{tr}(\Lambda^{k\top}(X-Z^k))+\frac{\rho}{2}\|X-Z^k\|_F^2\right)$。

又由于展开式中所有 X_{ij} 均为变量分离形式，故只需按以下方式更新：

$$X_{ij}^{k+1}=\arg\min_{X_{ij}}\left(\frac{1}{2}(X_{ij}-G_{ij})^2+\Lambda_{ij}^k X_{ij}+\frac{\rho}{2}(X_{ij}-Z_{ij}^k)^2\right)=\frac{G_{ij}+\rho Z_{ij}^k-\Lambda_{ij}^k}{1+\rho}$$

$$\begin{aligned}Z^{k+1}=&\arg\min_Z\left(\mathcal{I}_{S_+^n}(Z)+\operatorname{tr}(\Lambda^\top(X^{k+1}-Z))+\frac{\rho}{2}\|X^{k+1}-Z\|_F^2\right)\\ =&\arg\min_Z\left(\mathcal{I}_{S_+^n}(Z)+\frac{\rho}{2}\left\|X^{k+1}-Z+\frac{\Lambda^k}{\rho}\right\|_F^2\right)=\mathcal{P}_{S_+^n}\left(X^{k+1}+\frac{\Lambda^k}{\rho}\right)\end{aligned}$$

若有特征值分解 $X^{k+1}+\frac{\Lambda^k}{\rho}=V\operatorname{diag}(\lambda_1,\cdots,\lambda_n)V^\top$，则 Z 的更新为：

$Z^{k+1}=V\operatorname{diag}(\max\{0,\lambda_1\},\cdots,\max\{0,\lambda_n\})V^\top,\Lambda^{k+1}=\Lambda^k+\rho(X^{k+1}-Z^{k+1})$。

4. 给定算子 $A:\mathbb{R}^n\to\mathbb{R}^n$，称 A 是单调算子，如果其满足：

$$\langle Ax-Ay,x-y\rangle\geq 0,\forall x,y$$

证明：

(a). 给定闭凸函数 $f(x)$，次微分算子 ∂f 是单调算子。

(b). 给定闭凸函数 $f(x)$，若 $f(x)$ 是 μ-强凸的，那么 ∂f 是强单调算子，即：

$$\langle\partial f(x)-\partial f(y),x-y\rangle\geq\mu\|x-y\|^2,\forall x,y$$

(c). A 是非扩张的，等价于 $\frac{1}{2}(I+A)$ 是固定非扩张的。

(d). 给定凸函数 $f(x)$，且 $f(x)$ 一阶光滑，$\nabla f(x)$ 是 L-Lipchitz 连续的。定义 $G = I - t\nabla f$，$t \in \left(0, \frac{1}{L}\right]$

- 验证 G 是固定非扩张的。
- 若 $f(x)$ 是 μ-强凸的，那么 G 是压缩算子，即存在 $\rho \in (0,1)$ 使得

$$\|G(x) - G(y)\| \leq \rho \|x - y\|$$

解

(a). $\forall x, y \in \operatorname{dom} f, g_x \in \partial f(x), g_y \in \partial f(y)$，由次微分定义，有
$f(y) \geq f(x) + \langle g_x, y - x\rangle, f(x) \geq f(y) + \langle g_y, x - y\rangle$，
相加，$0 \geq \langle g_x, y - x\rangle + \langle g_y, x - y\rangle = \langle g_x - g_y, y - x\rangle$，两侧乘 -1 即为单调算子定义。

(b). 此时 $f(y) \geq f(x) + \langle g_x, y - x\rangle + \frac{\mu}{2}\|y - x\|^2, f(x) \geq f(y) + \langle g_y, x - y\rangle + \frac{\mu}{2}\|x - y\|^2$
仍然相加，$0 \geq \langle g_x, y - x\rangle + \langle g_y, x - y\rangle + \mu\|x - y\|^2 = -\langle g_y - g_x, y - x\rangle + \mu\|x - y\|^2$，
$\Rightarrow \langle g_y - g_x, y - x\rangle \geq \mu\|x - y\|^2$，即为强单调算子定义。

(c). $\Longrightarrow$：记 $T = \frac{1}{2}(I + A), u = Tx, v = Ty, A = 2T - I, \|(2T - I)x - (2T - I)y\|^2 \leq \|x - y\|^2$，
即 $\|2(u - v) - (x - y)\|^2 \leq \|x - y\|^2 \Rightarrow 4\|u - v\|^2 - 4\langle u - v, x - y\rangle + \|x - y\|^2 \leq \|x - y\|^2$，
$\Rightarrow \|u - v\|^2 \leq \langle u - v, x - y\rangle$，即 $\|Tx - Ty\|^2 \leq \langle Tx - Ty, x - y\rangle$，故 T 固定非扩张。

$\Longleftarrow$：由 $\|u - v\|^2 \leq \langle u - v, x - y\rangle$ 有 $4\|u - v\|^2 \leq 4\langle u - v, x - y\rangle$，此时
$\Rightarrow 4\|u - v\|^2 - 4\langle u - v, x - y\rangle + \|x - y\|^2 \leq \|x - y\|^2 \Rightarrow \|2(u - v) - (x - y)\|^2 \leq \|x - y\|^2$
$\Rightarrow \|(2T - I)x - (2T - I)y\|^2 = \|Ax - Ay\|^2 \leq \|x - y\|^2$，即 A 非扩张。

(d).

- 由 ∇f 是 L-Lipchitz 连续的，有 $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$。
记 $H = 2G - I = I - 2t\nabla f$，则有

$$\begin{aligned}
\|H(x) - H(y)\|^2 =& \|x - y - 2t(\nabla f(x) - \nabla f(y))\|^2 \\
=& \|x - y\|^2 - 4t\langle \nabla f(x) - \nabla f(y), x - y\rangle + 4t^2\|\nabla f(x) - \nabla f(y)\|^2 \\
\leq& \|x - y\|^2 - \frac{4t}{L}\|\nabla f(x) - \nabla f(y)\|^2 + 4t^2\|\nabla f(x) - \nabla f(y)\|^2 \\
=& \|x - y\|^2 + 4(t^2 - \frac{t}{L})\|\nabla f(x) - \nabla f(y)\|^2 \leq \|x - y\|^2
\end{aligned}$$

故 H 是非扩张的，进而 G 是固定非扩张的。

- 由 f 是 μ-强凸的，有 $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2$。

$$\begin{aligned}
\|G(x) - G(y)\| =& \|x - y - t(\nabla f(x) - \nabla f(y))\|^2 \\
=& \|x - y\|^2 - 2t\langle \nabla f(x) - \nabla f(y), x - y\rangle + t^2\|\nabla f(x) - \nabla f(y)\|^2 \\
\leq& \|x - y\|^2 - 2t\mu\|x - y\|^2 + t^2L^2\|x - y\|^2 \\
=& (L^2t^2 - 2\mu t + 1)\|x - y\|^2 \leq \|x - y\|^2
\end{aligned}$$

故 G 是压缩算子。

最优化算法作业 5

陈文轩

更新：May 21, 2025

1. 对于 group lasso 问题，请使用交替线性极小化方法求解，写出迭代公式。Group lasso 问题如下：

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{m} \sqrt{n_g}\|\beta_g\|_2$$

where：

- $y \in \mathbb{R}^n$ 和 $X \in \mathbb{R}^{n\times p}$ 是给定的数据，
- 变量 $\beta \in \mathbb{R}^p, \lambda > 0$ 是给定的参数，
- 下标集 $\{1, 2, \ldots, p\}$ 被分为不相交的 m 组，g 是分组下标，β_g 是对应分组 g 的分量，n_g 是分组 g 的分量维度。

解

首先进行外推操作：$\hat{\beta}_g^{k-1} = \beta_g^{k-1} + w_g^{k-1}(\beta_g^{k-1} - \beta_g^{k-2}), w_g^{k-1} > 0$ 是外推参数。

然后计算梯度：$\hat{g}_g^k = \nabla_{\beta_g} f(\hat{\beta}^{k-1}) = X_g^\top (X\hat{\beta}_{g^*}^{k-1} - y)$，其中 X_g 是 X 的组 g 对应列子矩阵，$\hat{\beta}_{g^*}^{k-1} = (\beta_1^k, \ldots, \beta_{g-1}^k, \hat{\beta}_g^{k-1}, \beta_{g+1}^{k-1}, \ldots, \beta_m^{k-1})$。

记 $z_g^k = \hat{\beta}_g{}^{k-1} - \frac{1}{L_g^{k-1}}\hat{g}_g^k, (t)_+ = \max\{t, 0\}$，其中 $L_g^{k-1} > 0$ 为常数，需要求解子问题如下：

$$\begin{aligned}
\beta_g^k &= \arg\min_{\beta_g} \left\{ \langle \hat{g}_g^k, \beta_g - \hat{\beta}_g^{k-1}\rangle + \frac{L_g^{k-1}}{2}\|\beta_g - \hat{\beta}_g^{k-1}\|_2^2 + \lambda\sqrt{n_g}\|\beta_g\|_2 \right\} \\
&= \arg\min_{\beta_g} \left\{ \frac{L_g^{k-1}}{2} \left\| \beta_g - \left(\hat{\beta}_g^{k-1} - \frac{1}{L_g^{k-1}}\hat{g}_g^k \right) \right\|_2^2 + \lambda\sqrt{n_g}\|\beta_g\|_2 \right\} \\
&= \arg\min_{\beta_g} \left\{ \frac{\lambda\sqrt{n_g}}{L_g^{k-1}}\|\beta_g\|_2 + \frac{1}{2}\|\beta_g - z_g^k\|_2^2 \right\} = \operatorname{prox}_{\frac{\lambda\sqrt{n_g}}{L_g^{k-1}}\|\cdot\|_2}(z_g^k) \\
&= \left(1 - \frac{\lambda\sqrt{n_g}}{L_g^{k-1}}\right)_+ z_g^k
\end{aligned}$$

以上即为完整的交替近似线性极小化算法更新。

2. 给定原问题

$$\min_{w\in\mathbb{R}^d} P(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(w^\top x_i) + \frac{\lambda}{2}\|w\|^2$$

其中 ϕ_i 为闭凸函数。

证明：对偶问题表述为：

$$\max_{\alpha\in\mathbb{R}^n} D(\alpha) = \frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) - \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2$$

其中 $\phi_i^*(u) = \max_z(zu - \phi_i(z))$ 是 ϕ_i 的共轭函数。

解

把问题重写为 $\min_{w\in\mathbb{R}^d, u\in\mathbb{R}^n}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i(u_i) + \frac{\lambda}{2}\|w\|^2\right), \text{s.t. } u_i = x_i^\top w, i = 1,\dots,n,$

则 Lagrange 函数为 $L(w,u,\mu) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(u_i) + \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n}\mu_i(x_i^\top w - u_i)$。以下记 $n\mu_i = \alpha_i$，

此时 $L(w,u,\mu) = \mathcal{L}(w,u,\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(\phi_i(u_i) + \alpha_i(x_i^\top w - u_i)\right) + \frac{\lambda}{2}\|w\|^2$，对 u, w 求极小：

考虑 u_i，即 $\min_{u_i} f_i(u_i) \coloneqq \min_{u_i}\frac{1}{n}\left(\phi_i(u_i) - \alpha_i u_i\right) = -\max_{u_i}\frac{1}{n}\left(u_i(-\alpha_i) - \phi_i(u_i)\right) = -\frac{1}{n}\phi_i^*(-\alpha_i)$。

考虑 w，即 $\min_w g(w) \coloneqq \min_w\left(\frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\alpha_i x_i^\top w\right), \nabla g = \lambda w + \frac{1}{n}\sum_{i=1}^{n}\alpha_i x_i = 0$

$$\Rightarrow w = -\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i, g(w) = \frac{\lambda}{2}\left\|-\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2 + \frac{1}{\lambda n^2}\left\|\sum_{i=1}^{n}\alpha_i x_i\right\|^2 = -\frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2。$$

相加以上函数，对偶问题为 $\max_\alpha D(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(-\alpha_i) - \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2$。

3. 给定矩阵 $A\in\mathbb{R}^{m\times n}$，使用随机梯度法求解如下问题，写出迭代具体公式

$$\min_{x\in\mathbb{R}^{n\times r}} \operatorname{tr}(x^\top A^\top A x).$$

解

记 $f(x) = \operatorname{tr}(x^\top A^\top A x) = \|Ax\|_F^2 = \sum_{i=1}^{m}\|a_i^\top x\|_2^2$，其中 $a_i^\top$ 是 A 的第 i 行，则 $\nabla f(x) = 2A^\top A x$。

均匀抽取一行 $a_{i_k}^\top$ 来近似梯度 $g_k = 2m a_{i_k} a_{i_k}^\top x_k$，由 $\mathbb{E}(g_k) = 2m\left(\frac{1}{m}\sum_{i=1}^{m} a_i a_i^\top\right)x = 2A^\top A x$，

有 g_k 是 $\nabla f(x)$ 的无偏估计。假设 η_k 是步长序列，则迭代为 $x_{k+1} = x_k - \eta_k g_k$。

4. 在最小二乘问题中出现的平方误差损失函数非常适合使用随机梯度方法进行最小化。我们的问题是：

$$\min_{\theta\in\mathbb{R}^n}\frac{1}{2}\|X\theta - y\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(x_i^\top\theta - y_i)^2,$$

其中 $x_i^\top$ 是 $X\in\mathbb{R}^{m\times n}$ 的第 i 行，$y\in\mathbb{R}^m$。我们可以将这个目标函数写为：

$$f(\theta) = \sum_{i=1}^{m} f(\theta),$$

其中

$$f_i(\theta) = \frac{1}{2}(x_i^\top \theta - y_i)^2，对于 i = 1, \ldots, m.$$

然后随机梯度方法给出更新规则：

$$\theta_{k+1} = \theta_k - \eta_k \nabla f_{s[k]}(\theta_k)$$

其中 η_k 是步长（也称为学习率），$s[k] \in \{1, \ldots, m\}$，通常是通过从集合 $\{1, \ldots, m\}$ 中随机抽取一个数字。

(a). 假设 $\{x_i\}_{i=1}^m$ 是一组相互正交的向量。找到一个固定的步长 η，使得随机梯度方法收敛到最小二乘问题的解。

(b). 如果没有条件上述条件，即 $\{x_i\}_{i=1}^m$ 并不相互正交，那么随机梯度法的收敛需要什么条件？

解

(a). 记最小二乘解为 $\theta^* = (X^\top X)^{-1} X^\top y$，误差 $e_k = \theta_k = \theta^*$，梯度 $\nabla f_i(\theta) = (x_i^\top \theta - y_i) x_i$.
设 $s[k] = i$，则 $e_{k+1} = e_k - \eta(x_i^\top \theta - y_i)x_i = e_k - \eta(x_i^\top e_k)x_i$。由于 $\{x_i\}$ 正交，
有分解 $e_k = \sum_{j=1}^{m} \frac{x_j^\top e_k}{\|x_i\|_2^2} x_j + e^\perp$，其中 $e^\perp$ 与所有 x_i 正交，更新时不变。更新后结果为
$e_{k+1} = \sum_{j \neq i} \frac{x_j^\top e_k}{\|x_i\|_2^2} x_j + \frac{x_i^\top e_k}{\|x_i\|_2^2}(1 - \|x_i\|_2^2 \eta) x_j + e^\perp$，若误差减少，则 $(1 - \|x_i\|_2^2 \eta) < 1, \forall i$。
此时需要 $\eta \in \left(0, \frac{2}{\max\limits_i \|x_i\|_2^2}\right)$，此时取固定步长 $\eta = \frac{1}{\max\limits_i \|x_i\|_2^2}$ 即可。

(b). 此时需要 $X^\top X$ 非奇异，即 $X^\top X$ 正定，此时目标函数强凸，解唯一。
随机梯度均匀采样时其无偏且二阶矩有界。记 $L = \lambda_{\max}(X^\top X)$，则以下学习率收敛：

- 满足 $\eta \in \left(0, \frac{2}{L}\right)$ 的常数学习率；
- 满足 $\sum_{k=0}^{\infty} \eta_k = \infty, \sum_{k=0}^{\infty} \eta_k^2 < \infty$ 的递减学习率。