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：

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

   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 $\nabla^2 f_i(x) = -b_i a_i(-b_i a_i^\top)\sigma_i(1-\sigma_i) = a_i a_i^\top \sigma_i(1-\sigma_i)$，
   $\|\nabla^2 l(x)\| = \left\|\dfrac{1}{m}\displaystyle\sum_{i=1}^m \nabla^2 f_i(x)\right\| = \dfrac{1}{m}\displaystyle\sum_{i=1}^m \|a_i a_i^\top \sigma_i(1-\sigma_i)\| \le \dfrac{1}{4m}\displaystyle\sum_{i=1}^m \|a_i\|^2$
   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 $\partial f(x) = \begin{cases} \{\mathrm{sgn}(x)\}, & x \neq 0 \\ [-1, 1], & x = 0 \end{cases}$。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 $\mathrm{prox}_h(x) = \arg\min_y \left(h(y) + \frac{1}{2}\|y-x\|^2 \right)$

- $h(x) = \|x\|_\infty$（需要求解一个一维子问题）。

解

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

考虑 $\mathrm{prox}_{h^*}(x) = \arg\min_z \left(I_{\{z \mid \|z\|_1 \le 1\}}(z) + \frac{1}{2}\|z-x\|^2 \right)$ 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 $\sum_{k=1}^{n} \max\{|x_k| + -\lambda, 0\} = 1$ 的解。

综上所述，$\mathrm{prox}_h(x) = x - P_C(x), P_C(x)$ 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

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

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

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

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

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 $y = \frac{x}{\|x\|_2}(\|x\|_2 - 1)$。

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

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

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

$\Rightarrow \nabla(-\log\det V(x))=\left(-a_1^\top V(x)^{-1}a_1,\cdots,-a_n^\top V(x)^{-1}a_n\right)^\top$，记 $f(x)=-\log\det V(x))$，

需要求解子问题 $x_k=\arg\min\limits_{x\in\Delta_n}\langle\nabla f(y_{k-1}),x\rangle$，这是一个线性问题，最优解在顶点 e_j 上取。

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

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

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

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

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