

第三章 典型优化问题

修贤超

<https://xianchaoxiu.github.io>

- 3.1 线性规划
- 3.2 最小二乘问题
- 3.3 复合优化问题
- 3.4 随机优化问题
- 3.5 半定规划
- 3.6 矩阵优化
- 3.7 优化模型语言

凸优化问题

■ 标准形式的凸优化问题

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & a_i^\top x = b_i, \quad i = 1, \dots, p \end{aligned}$$

□ f_0, f_1, \dots, f_m 为凸函数

□ $a_i^\top x = b_i$ 为线性等式约束

■ 经常写成

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & Ax = b \end{aligned}$$

■ 考虑

$$\begin{array}{ll}\min_{x_1, x_2} & f_0(x) = x_1^2 + x_2^2 \\ \text{s.t.} & f_1(x) = x_1/(1 + x_2^2) \leq 0 \\ & h_1(x) = (x_1 + x_2)^2 = 0\end{array}$$

□ f_0 为凸函数, 可行集 $\{(x_1, x_2) \mid x_1 = -x_2 \leq 0\}$ 为凸集

□ f_1 非凸, h_1 不是线性函数

■ 不是凸问题, 但可转化为凸优化问题

$$\begin{array}{ll}\min_{x_1, x_2} & x_1^2 + x_2^2 \\ \text{s.t.} & x_1 \leq 0 \\ & x_1 + x_2 = 0\end{array}$$

局部和全局极小

■ 凸优化问题的任意局部极小点都是全局最优

证明 设 x 是局部极小解, y 是全局最优解且 $f_0(y) < f_0(x)$. 存在 $R > 0$ 使

$$z \text{可行}, \quad \|z - x\|_2 \leq R \quad \Rightarrow \quad f_0(z) \geq f_0(x)$$

考虑 $z = \theta y + (1 - \theta)x$ 且 $\theta = R/(2\|y - x\|_2)$

□ $\|y - x\|_2 > R$, 则 $0 < \theta < 1/2$

□ z 是两个可行点的凸组合, 则也可行

□ $\|z - x\|_2 = R/2$, 并且

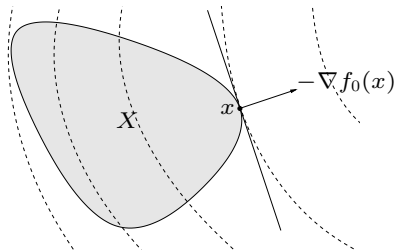
$$f_0(z) \leq \theta f_0(y) + (1 - \theta)f_0(x) < f_0(x)$$

这与 x 是局部极小的假设矛盾

可微凸优化问题的最优性条件

- 设 x 是凸优化问题 $\min_{x \in X} f_0(x)$ 的最优解当且仅当 x 可行且满足

$$\nabla f_0(x)^\top (y - x) \geq 0, \quad \forall y \in X$$



- 如果 $\nabla f_0(x)$ 非零, 它定义了可行集 X 在 x 处的支撑超平面

具体含义

- 无约束优化 x 是最优解当且仅当

$$x \in \text{dom } f_0, \quad \nabla f_0(x) = 0$$

- 等式约束优化问题

$$\min f_0(x) \quad Ax = b$$

x 是最优解当且仅当存在 v 使得

$$x \in \text{dom } f_0, \quad Ax = b, \quad \nabla f_0(x) + A^\top v = 0$$

- 非负约束优化问题

$$\min f_0(x) \quad x \geq 0$$

x 是最优解当且仅当

$$x \in \text{dom } f_0, \quad x \geq 0, \quad \begin{cases} \nabla f_0(x)_i \geq 0, & x_i = 0 \\ \nabla f_0(x)_i = 0, & x_i > 0 \end{cases}$$

线性规划基本形式

■ 线性规划问题的一般形式

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & Gx \leq e \end{aligned}$$

■ 线性规划问题的标准形式

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned}$$

■ 线性规划问题的不等式形式

$$\begin{aligned} \max_{y \in \mathbb{R}^n} \quad & b^\top y \\ \text{s.t.} \quad & A^\top y \leq c \end{aligned}$$

应用举例：基追踪问题

- 基追踪问题是压缩感知中的一个基本问题，可以写为

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \|x\|_1 \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

- 对每个 $|x_i|$ 引入一个新的变量 z_i ，可以转化为

$$\begin{aligned} \min_{z \in \mathbb{R}^n} \quad & \sum_{i=1}^n z_i \\ \text{s.t.} \quad & Ax = b \\ & -z_i \leq x_i \leq z_i, \quad i = 1, \dots, n \end{aligned}$$

应用举例: 数据拟合

- 最小 ℓ_∞ 范数模型

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_\infty$$

- 令 $t = \|Ax - b\|_\infty$, 得到等价问题

$$\begin{aligned} \min_{x \in \mathbb{R}^n, t \in \mathbb{R}} \quad & t \\ \text{s.t.} \quad & \|Ax - b\|_\infty \leq t \end{aligned}$$

- 利用 ℓ_∞ 范数的定义, 可以进一步写为

$$\begin{aligned} \min_{x \in \mathbb{R}^n, t \in \mathbb{R}} \quad & t \\ \text{s.t.} \quad & -t\mathbf{1} \leq Ax - b \leq t\mathbf{1} \end{aligned}$$

- 3.1 线性规划
- 3.2 最小二乘问题
- 3.3 复合优化问题
- 3.4 随机优化问题
- 3.5 半定规划
- 3.6 矩阵优化
- 3.7 优化模型语言

最小二乘问题

- 最小二乘问题的一般形式如下

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^m r_i^2(x)$$

- 如果所有的 $r_i : \mathbb{R}^n \rightarrow \mathbb{R}$ 都是线性函数，则称线性最小二乘问题，否则称为非线性最小二乘问题
- 如果噪声服从高斯分布，最小二乘问题的解对应于原问题的最大似然解
- 1801 年，24 岁的高斯计算出小行星的运动轨道

应用举例：线性最小二乘问题

- 线性最小二乘问题是回归分析中的一个基本模型，它可以表示为

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^m (a_i^\top x - b_i)^2$$

- 记 $A = [a_1, a_2, \dots, a_m]^\top$ ，上式可以等价地写成

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} \|Ax - b\|_2^2$$

- $x \in \mathbb{R}^n$ 为其全局极小解当且仅当 x 满足

$$\nabla f(x) = A^\top (Ax - b) = 0$$

应用举例：数据插值

- 给定数据集 $\{a_i \in \mathbb{R}^p, b_i \in \mathbb{R}^q, i = 1, 2, \dots, m\}$, **插值**是求一个映射 f , 使得

$$b_i = f(a_i), \quad i = 1, 2, \dots, m$$

- 利用线性函数 $f(a) = Xa + y$ 逼近, 可以建立如下最小二乘问题

$$\min_{X \in \mathbb{R}^{q \times p}} \sum_{i=1}^m \|Xa_i + y - b_i\|^2$$

- 设 $\{\phi_i(a)\}_{i=1}^n (n \leq m)$ 为插值空间的一组基, 数据插值可以写成

$$b_j = f(a_j) = \sum_{i=1}^n x_i \phi_i(a_j), \quad j = 1, 2, \dots, m$$

应用举例：数据插值

- 设非线性向量函数 $\phi_i(\theta) : \mathbb{R}^q \rightarrow \mathbb{R}^q$ ，并构造如下复合函数

$$f(\theta) = \phi_n(X_n \phi_{n-1}(X_{n-1} \cdots \phi_1(X_1 \theta + y_1) \cdots + y_{n-1}) + y_n)$$

- 常用的有 ReLU，即

$$\phi_i(\theta) = (\text{ReLU}(\theta_1), \text{ReLU}(\theta_2), \cdots, \text{ReLU}(\theta_q))^{\top}, \quad i = 1, 2, \cdots, n$$

$$\text{ReLU}(t) = \begin{cases} t, & t \geq 0 \\ 0, & \text{其他} \end{cases}$$

- 更多未知的非线性，可能在更大的函数空间中得到一个更好的逼近

应用举例：带微分方程约束优化问题

- 当约束中含微分方程时，称为带微分方程约束的优化问题
- 考虑瓦斯油催化裂解生成气体和其他副产物的反应过程

$$\begin{cases} \dot{y}_1 = -(\theta_1 + \theta_3)y_1^2 \\ \dot{y}_2 = \theta_1 y_1^2 - \theta_2 y_2 \end{cases}$$

- 转化为最小二乘问题

$$\begin{aligned} \min_{\theta \in \mathbb{R}^3} \quad & \sum_{j=1}^n \|y(\tau_j; \theta) - z_j\|^2 \\ \text{s.t.} \quad & y(\tau; \theta) \text{ 满足上述方程组} \end{aligned}$$

其中 z_j 是在时刻 τ_j 的 y 的测量值， n 为测量的时刻数量

- 3.1 线性规划
- 3.2 最小二乘问题
- 3.3 复合优化问题
- 3.4 随机优化问题
- 3.5 半定规划
- 3.6 矩阵优化
- 3.7 优化模型语言

- 复合优化问题一般可以表示为

$$\min_{x \in \mathbb{R}^n} \quad \psi(x) = f(x) + h(x)$$

- $f(x)$ 是光滑函数, 如数据拟合项
- $h(x)$ 可能是非光滑的, 如 ℓ_1 范数正则项, 约束集合的示性函数

- 常用的优化算法有

- 次梯度法
- 近似点梯度法
- Nesterov 加速法
- 交替方向乘子法

■ ℓ_1 范数正则化回归分析问题

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2} \|Ax - b\|_2^2 + \mu \|x\|_1$$

■ 矩阵分离问题

$$\begin{aligned} \min_{X, S \in \mathbb{R}^{m \times n}} \quad & \|X\|_* + \mu \|S\|_1 \\ \text{s.t.} \quad & X + S = M \end{aligned}$$

■ 字典学习问题

$$\begin{aligned} \min_{X, D \in \mathbb{R}^{m \times n}} \quad & \frac{1}{2n} \|DX - A\|_F^2 + \lambda \|X\|_1 \\ \text{s.t.} \quad & \|D\|_F \leq 1 \end{aligned}$$

应用举例：图像去噪

- 图像去噪是指从一个带噪声的图像中恢复出不带噪声的原图
- 由全变差模型，去噪问题可表示为

$$\min_{x \in \mathbb{R}^{n \times n}} \frac{1}{2} \|x - y\|_F^2 + \lambda \|x\|_{TV}$$



应用举例：盲反卷积

- 反卷积是从一个模糊的图像恢复出原来清晰的图像，也称为去模糊
- 反卷积问题的模型

$$y = a * x + \varepsilon$$

- 设噪声为高斯噪声，可转化为

$$\min_{a,x} \quad \frac{1}{2} \|y - a * x\|_2^2$$

- 设原始图像信号在小波变换下是稀疏的，进一步得到

$$\min_{a,x} \quad \frac{1}{2} \|y - a * x\|_2^2 + \|\lambda \odot (Wx)\|_1$$

其中 W 是小波框架, $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)^\top$ 用来控制稀疏度

- 3.1 线性规划
- 3.2 最小二乘问题
- 3.3 复合优化问题
- 3.4 随机优化问题
- 3.5 半定规划
- 3.6 矩阵优化
- 3.7 优化模型语言

- 随机优化问题可以表示为

$$\min_{x \in \mathcal{X}} \quad \mathbb{E}_{\xi}[F(x, \xi)] + h(x)$$

- $F(x, \xi)$ 表示样本 ξ 上的损失或奖励
- $h(x)$ 用来保证解的某种性质

- 设有 N 个样本 $\xi_1, \xi_2, \dots, \xi_N$, 令 $f_i(x) = F(x, \xi_i)$, 得到**经验风险极小化问题**

$$\min_{x \in \mathcal{X}} \quad f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x) + h(x)$$

- 样本数 N 比较多, 可行域所在空间维数 n 比较大, 导致计算困难

应用举例：随机主成分分析

- 如果样本点 ξ 服从某个零均值分布 \mathcal{D} ，则随机主成分分析可以写成

$$\max_{X \in \mathbb{R}^{p \times d}} \text{Tr}(X^\top A A^\top X) \quad \text{s.t.} \quad X^\top X = I$$

\Downarrow

$$\max_{X \in \mathbb{R}^{p \times d}} \text{Tr}(X^\top \mathbb{E}_{\xi \sim \mathcal{D}}[\xi \xi^\top] X) \quad \text{s.t.} \quad X^\top X = I$$

\Downarrow

$$\max_{X \in \mathbb{R}^{p \times d}} \frac{1}{N} \sum_{i=1}^N \text{Tr}(X^\top A_i A_i^\top X) \quad \text{s.t.} \quad X^\top X = I$$

应用举例：分布式鲁棒优化

- 为了提高深度学习预测器的泛化能力，考虑

$$\begin{aligned} \min_h \quad & \mathbb{E}_z[F(h, z)] \\ & \Downarrow \\ \min_h \quad & \max_{\hat{z} \in \Gamma} \mathbb{E}_{\hat{z}}[F(h, \hat{z})] \end{aligned}$$

- 集合 Γ 中随机变量的分布与真实数据的分布在一定意义下非常接近
- Wasserstein 距离可以改变原来经验分布的支撑集

Generative Adversarial Nets

**Ian J. Goodfellow^{*}, Jean Pouget-Abadie[†], Mehdi Mirza, Bing Xu, David Warde-Farley,
Sherjil Ozair[‡], Aaron Courville, Yoshua Bengio[§]**

The adversarial modeling framework is most straightforward to apply when the models are both multilayer perceptrons. To learn the generator's distribution p_g over data x , we define a prior on input noise variables $p_z(z)$, then represent a mapping to data space as $G(z; \theta_g)$, where G is a differentiable function represented by a multilayer perceptron with parameters θ_g . We also define a second multilayer perceptron $D(x; \theta_d)$ that outputs a single scalar. $D(x)$ represents the probability that x came from the data rather than p_g . We train D to maximize the probability of assigning the correct label to both training examples and samples from G . We simultaneously train G to minimize $\log(1 - D(G(z)))$. In other words, D and G play the following two-player minimax game with value function $V(G, D)$:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]. \quad (1)$$

Q&A

Thank you!

感谢您的聆听和反馈