预备知识

凸集

对集合C内的任意两点 $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$,若它们之间连线上的所有点仍属于集合 C，即

$$\theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2 \in C \quad (\forall\, 0 \leqslant \theta \leqslant 1)\,, \tag{1.1}$$

则我们称集合C 为“凸”的，即 C 是一个凸集.

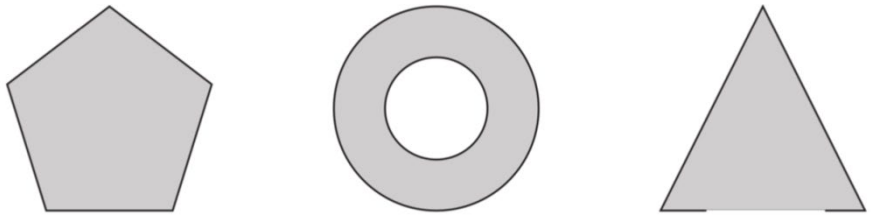


图 1.1 仅有第一个集合是凸的

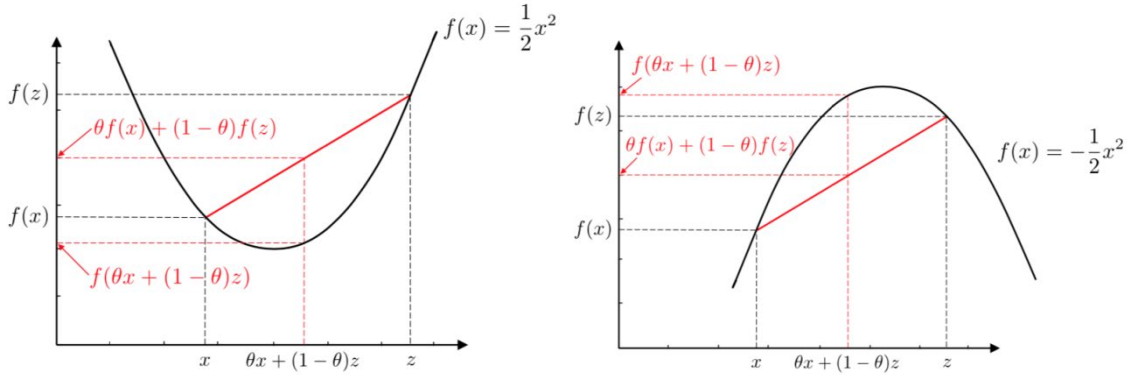
凸函数和凹函数

对定义在凸集上的函数 $f : \mathbb{R}^d \mapsto \mathbb{R}$，令 dom_f 表示其定义域，若 $\forall \boldsymbol{x}, \boldsymbol{z} \in \mathrm{dom}_f$ 均满足

$$f\left(\theta \boldsymbol{x} + (1-\theta)\,\boldsymbol{z}\right) \leqslant \theta f\left(\boldsymbol{x}\right) + (1-\theta)\, f\left(\boldsymbol{z}\right) \quad \left(\forall\, 0 \leqslant \theta \leqslant 1\right), \tag{1.2}$$

则我们称函数 $f(\cdot)$ 为凸的，即 $f(\cdot)$ 是一个凸函数.

若将式1.2中的不等号反向，则函数 $f(\cdot)$ 是凹函数.



(a) 凸函数 $f(x) = \frac{1}{2}x^2$　　(b) 凹函数 $f(x) = -\frac{1}{2}x^2$

图 1.2　典型的凸函数和凹函数

常见的凸函数

表1.1列出了一些常见的凸函数.

表 1.1 常见凸函数

名称	函数形式	定义域	参数
1 维仿射函数	$ax+b$	$x \in \mathbb{R}$	$a, b \in \mathbb{R}$
1 维指数函数	e^{ax}	$x \in \mathbb{R}$	$a \in \mathbb{R}$
1 维幂函数	x^a	$x \in \mathbb{R}_+$	$a \geqslant 1$ 或 $a \leqslant 0$
1 维绝对值幂函数	$\lvert x\rvert^p$	$x \in \mathbb{R}$	$p \geqslant 1$
1 维负熵函数	$x \log x$	$x \in \mathbb{R}_+$	——
d 维仿射函数	$\mathbf{a}^\top \boldsymbol{x} + b$	$x \in \mathbb{R}^d$	$\mathbf{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$
d 维范数	$\lVert\boldsymbol{x}\rVert_p = (\sum_{i=1}^d \lvert x_i\rvert^p)^{1/p}$	$x \in \mathbb{R}^d$	$p \geqslant 1$

梯度与凸函数

函数 $f : \mathbb{R}^d \mapsto \mathbb{R}$ 的梯度(gradient)记为 $\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}; \cdots; \frac{\partial f(\boldsymbol{x})}{\partial x_d}\right) \in \mathbb{R}^d$. 若函数 $f(\cdot)$ 可微，则它是凸函数当且仅当其定义域 dom_f 是凸集且 $\forall \boldsymbol{x}, \boldsymbol{z} \in \mathrm{dom}_f$ 都有

$$f(\boldsymbol{z}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{z} - \boldsymbol{x}) \ . \tag{1.4}$$

上式意味着 $f(\cdot)$ 在定义域中任意点的一阶泰勒展开是其下界. 例如，图1.3显示的凸函数 $f(x) = \frac{1}{2}x^2$ 及其在 $(1, f(1))$ 处的一阶泰勒展开.

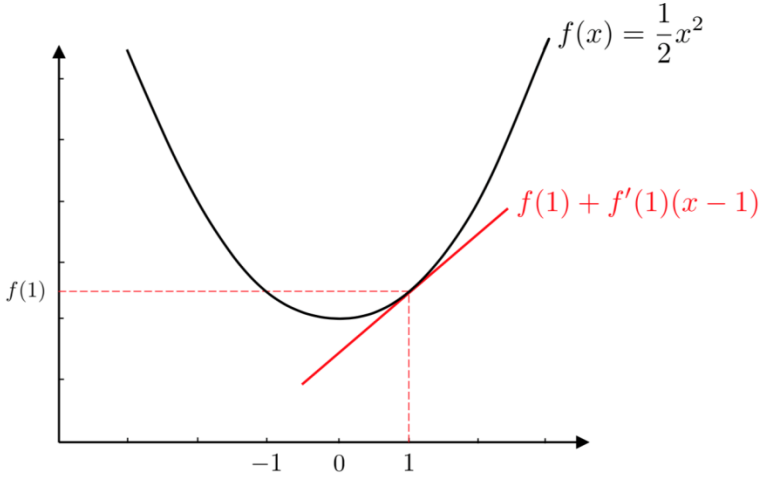


图 1.3 凸函数 $f(x) = \frac{1}{2}x^2$ 及其在 $(1, f(1))$ 处的一阶泰勒展开

Hessian矩阵和凸函数

除了使用一阶信息，我们还可以基于二阶信息来判断函数的凸性.

函数 $f: \mathbb{R}^d \mapsto \mathbb{R}$ 在定义域 dom_f 中 $\boldsymbol{x}$ 处的二阶导数矩阵(即Hessian矩阵)记为 $\nabla^2 f(\boldsymbol{x}) \in \mathbb{R}^{d \times d}$，其中 $\nabla^2 f(\boldsymbol{x})_{ij} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}$.

若函数 $f(\cdot)$ 二阶可微，则它是凸函数当且仅当 dom_f 是凸集且 $\nabla^2 f(\boldsymbol{x}) \succeq 0$，即 $\forall \boldsymbol{x} \in \mathrm{dom}_f$ 的Hessian矩阵都是半正定的.

例如：二次函数 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + c$ 是凸函数当且仅当 $\mathbf{A} \succeq 0$.

数学变换与函数的凸性

一些数学变换能够保持函数的凸性，例如

- f 是凸函数，则$g(x) = f(\mathbf{A}\boldsymbol{x} + \boldsymbol{b})$也是凸函数；
- $f_1, \cdots, f_n$ 是凸函数，$w_1, \cdots, w_n \geq 0$，则$f(\boldsymbol{x}) = \sum_{i=1}^{n} w_i f_i(\boldsymbol{x})$ 也是凸函数；
- $f_1, \cdots, f_n$ 是凸函数，则$f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x})\}$也是凸函数；
- $\forall \boldsymbol{z} \in \mathcal{X}$ $f(\boldsymbol{x}, \boldsymbol{z})$是关于 x 的凸函数，则 $g(\boldsymbol{x}) = \sup_{\boldsymbol{z} \in \mathcal{X}} f(\boldsymbol{x}, \boldsymbol{z})$也是关于$\boldsymbol{x}$ 的凸函数.

共轭函数

函数 $f : \mathbb{R}^d \mapsto \mathbb{R}$ 的共轭函数定义为

$$f_*(\boldsymbol{z}) = \sup_{\boldsymbol{x} \in \mathrm{dom}\, f} \left(\boldsymbol{z}^{\mathrm{T}} \boldsymbol{x} - f(\boldsymbol{x}) \right) , \tag{1.5}$$

其定义域

$$\mathrm{dom}\, f_* = \left\{ \boldsymbol{z} \,\middle|\, \sup_{\boldsymbol{x} \in \mathrm{dom}\, f} \left(\boldsymbol{z}^{\mathrm{T}} \boldsymbol{x} - f(\boldsymbol{x}) \right) < \infty \right\} . \tag{1.6}$$

直观来看，共轭函数 $f_*(z)$ 反映的是线性函数 $z^{\mathrm{T}}x$ 与 $f(x)$ 之间的最大差值. 图1.4为我们显示了一个实例.

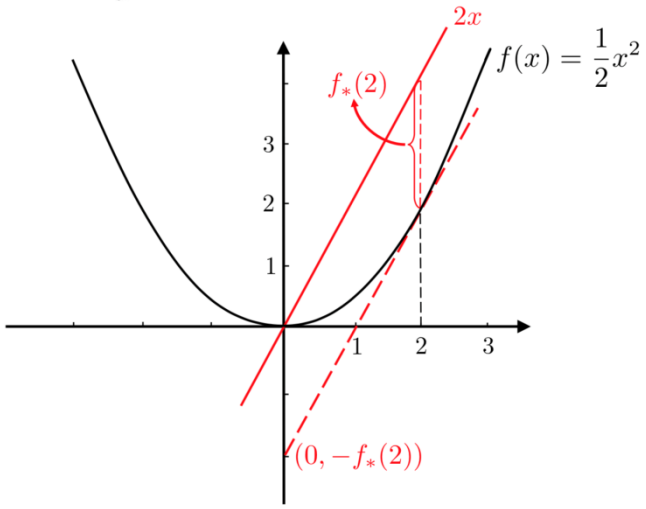


图 1.4 函数 $f(x) = \frac{1}{2}x^2$ 的共轭函数在 $z = 2$ 处的值的计算方法示意图

共轭函数

共轭函数有一些很好的性质：

- 无论原函数 f 是否是凸函数，共轭函数 f_* 一定是凸函数.
- 若函数 f 可微，则：

$$f_*(\nabla f(\boldsymbol{x})) = \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{x} - f(\boldsymbol{x}) = -\left[f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(0 - \boldsymbol{x})\right] \ . \quad (1.7)$$

重要不等式

【Jensen 不等式】 对任意凸函数 $f(\cdot)$ 有

$$f(\mathbb{E}[X]) \leqslant \mathbb{E}[f(X)] \ .$$

由Jensen不等式可知$(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$.

【Hölder 不等式】 对 $p, q \in \mathbb{R}_+$ 且 $\frac{1}{p} + \frac{1}{q} = 1$, 有

$$\mathbb{E}\left[|XY|\right] \leqslant \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \left(\mathbb{E}\left[|Y|^q\right]\right)^{\frac{1}{q}} \ .$$

【Cauchy-Schwartz 不等式】

$$\mathbb{E}[|XY|] \leqslant \sqrt{\mathbb{E}[X^2]E[Y^2]} \ .$$

重要不等式

【Liapounov 不等式】 对 $0 < r \leqslant s$ 有

$$\sqrt[r]{(\mathbb{E}[|X|^r])} \leqslant \sqrt[s]{(\mathbb{E}[|X|^s])} \ .$$

【Minkowski 不等式】 对 $1 \leqslant p$ 有

$$\sqrt[p]{\mathbb{E}[|X+Y|^p]} \leqslant \sqrt[p]{(\mathbb{E}[|X|^p]} + \sqrt[p]{(\mathbb{E}[|Y|^p]} \ .$$

【Bhatia - Davis 不等式】 对 $X \in [a,b]$ 有

$$\mathbb{V}[X] \leqslant (b - \mathbb{E}[X])(\mathbb{E}[X] - a) \leqslant \frac{(b-a)^2}{4} \ .$$

重要不等式

【联合界 (Union Bound) 不等式】

$$P(X \cup Y) \leqslant P(X) + P(Y) \ .$$

【Markov 不等式】 对 $X \geqslant 0, \forall \epsilon > 0$, 有

$$P(X \geqslant \epsilon) \leqslant \frac{\mathbb{E}[X]}{\epsilon} \ .$$

【Chebyshev 不等式】 $\forall \epsilon > 0$ 有

$$P\left(|X - \mathbb{E}[X]| \geqslant \epsilon\right) \leqslant \frac{\mathbb{V}[X]}{\epsilon^2} \ .$$

重要不等式

【Cantelli 不等式】 $\forall \epsilon > 0$ 有　　亦称单边 Chebyshev 不等式.

$$P(X - \mathbb{E}[X] \geqslant \epsilon) \leqslant \frac{\mathbb{V}[X]}{\mathbb{V}[X] + \epsilon^2}$$

$$P(X - \mathbb{E}[X] \leqslant -\epsilon) \leqslant \frac{\mathbb{V}[X]}{\mathbb{V}[X] + \epsilon^2} .$$

【Chernoff 不等式】 对 m 个独立随机变量 $X_i \in [a, b]$ $(i = 1, \ldots, m)$, 令 $\bar{X} = \sum_{i=1}^{m} X_i / m$, 有

$$P(\bar{X} - \mathbb{E}[\bar{X}] \geqslant \epsilon) \leqslant e^{-2m\epsilon^2/(b-a)^2} ,$$

$$P(\bar{X} - \mathbb{E}[\bar{X}] \leqslant -\epsilon) \leqslant e^{-2m\epsilon^2/(b-a)^2} .$$

重要不等式

【Cantelli 不等式】 $\forall \epsilon > 0$ 有　　亦称单边 Chebyshev 不等式.

$$P(X - \mathbb{E}[X] \geqslant \epsilon) \leqslant \frac{\mathbb{V}[X]}{\mathbb{V}[X] + \epsilon^2}$$

$$P(X - \mathbb{E}[X] \leqslant -\epsilon) \leqslant \frac{\mathbb{V}[X]}{\mathbb{V}[X] + \epsilon^2} .$$

【Chernoff 不等式】 对 m 个独立随机变量 $X_i \in [a, b]$ $(i = 1, \ldots, m)$, 令 $\bar{X} = \sum_{i=1}^{m} X_i / m$, 有

$$P(\bar{X} - \mathbb{E}[\bar{X}] \geqslant \epsilon) \leqslant e^{-2m\epsilon^2/(b-a)^2} ,$$

$$P(\bar{X} - \mathbb{E}[\bar{X}] \leqslant -\epsilon) \leqslant e^{-2m\epsilon^2/(b-a)^2} .$$

重要不等式

在机器学习研究中常用到Chernoff不等式的另一种表达形式，若

$$P(\bar{X} \geqslant \mathbb{E}[\bar{X}] + \epsilon) \leqslant e^{-2m\epsilon^2} = \delta \ ,$$

则下式至少以$1-\delta$的概率成立.

$$\bar{X} \leqslant \mathbb{E}[\bar{X}] + \sqrt{\frac{1}{2m} \ln \frac{1}{\delta}} \ .$$

【Hoeffding 不等式】 对 m 个独立随机变量 $X_i \in [0,1] \ \ (i = 1, \ldots, m)$, 令 $\bar{X} = \frac{1}{m} \sum_{i=1}^{m} X_i$, 有

$$P\left(\bar{X} - \mathbb{E}[\bar{X}] \geqslant \epsilon\right) \leqslant e^{-2m\epsilon^2} \ .$$

重要不等式

【McDiarmid 不等式】 对 m 个独立随机变量 $X_i \in \mathcal{X}$ $(i = 1, \ldots, m)$, 若 $f: \mathcal{X}^m \to \mathbb{R}$ 是关于 X_i 的实值函数且 $\forall x_1, \ldots, x_m, x_i' \in \mathcal{X}$ 都有

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leqslant c_i \ ,$$

则 $\forall \epsilon > 0$ 有

$$P\left(f\left(X_1, \ldots, X_m\right) - \mathbb{E}\left[f\left(X_1, \ldots, X_m\right)\right] \geqslant \epsilon\right) \leqslant e^{-2\epsilon^2 / \sum_{i=1}^m c_i^2} \ . \qquad (1.25)$$

【Bennett 不等式】 对 m 个独立同分布的随机变量 X_i $(i = 1, \ldots, m)$, 令 $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$, 若 $X_i - \mathbb{E}[X_i] \leqslant 1$, 则有

$$P(\bar{X} \geqslant \mathbb{E}[\bar{X}] + \epsilon) \leqslant \exp\left(\frac{-m\epsilon^2}{2\mathbb{V}[X_1] + 2\epsilon/3}\right) \ .$$

重要不等式

在机器学习研究中常用到Bennett不等式的另一种表达形式，若

$$P(\bar{X} \geqslant \mathbb{E}[\bar{X}] + \epsilon) \leqslant \exp\left(\frac{-m\epsilon^2}{2\mathbb{V}[X_1] + 2\epsilon/3}\right) = \delta \ ,$$

则下式至少以$1-\delta$的概率成立.

$$\bar{X} \leqslant \mathbb{E}[\bar{X}] + \epsilon \leqslant \mathbb{E}[\bar{X}] + \frac{2\ln 1/\delta}{3m} + \sqrt{\frac{2\mathbb{V}[X_1]}{m}\ln\frac{1}{\delta}} \ .$$

【Bernstein 不等式】 对 m 个独立同分布的随机变量 $X_i \ \ (i = 1, \ldots, m)$, 令 $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$, 若存在 $b > 0$ 使得 $\forall k \geqslant 2$ 有 $\mathbb{E}[|X_i|^k] \leqslant k!\, b^{k-2}\mathbb{V}[X_1]/2$ 成立, 则有

$$P(\bar{X} \geqslant \mathbb{E}[\bar{X}] + \epsilon) \leqslant \exp\left(\frac{-m\epsilon^2}{2\mathbb{V}[X_1] + 2b\epsilon}\right) \ .$$

重要不等式

【Azuma 不等式】 对于均值为 μ 的 **鞅** (martingale) $\{Z_m,\ m \geqslant 1\}$, 令 $Z_0 = \mu$, 若 $-c_i \leqslant Z_i - Z_{i-1} \leqslant c_i$, 则 $\forall \epsilon > 0$ 有

$$P\left(Z_m - \mu \geqslant \epsilon\right) \leqslant e^{-\epsilon^2/2\sum_{i=1}^{m} c_i^2} ,$$

$$P\left(Z_m - \mu \leqslant -\epsilon\right) \leqslant e^{-\epsilon^2/2\sum_{i=1}^{m} c_i^2} .$$

令 $X_i = Z_i - Z_{i-1}$ 可以得到 **鞅差序列** (martingale difference sequence) $X_1, X_2, \ldots, X_m$, 于是有

$$P\left(\sum_{i=1}^{m} X_i \geqslant \epsilon\right) \leqslant e^{-\epsilon^2/2\sum_{i=1}^{m} c_i^2} ,$$

$$P\left(\sum_{i=1}^{m} X_i \leqslant -\epsilon\right) \leqslant e^{\epsilon^2/2\sum_{i=1}^{m} c_i^2} .$$

最小二乘问题和线性规划问题

一种最简单的优化问题是最小二乘问题

$$\min_{\boldsymbol{x}} \ f(\boldsymbol{x}) = \sum_{i=1}^{m} \left(\mathbf{a}_i^{\mathrm{T}}\boldsymbol{x} - b_i\right)^2 = \|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 \ , \qquad (1.28)$$

其中 $\boldsymbol{x} = (x_1; \ldots; x_d) \in \mathbb{R}^d$ 为 d 维优化变量, $\mathbf{A} = (\boldsymbol{a}_1; \ldots; \boldsymbol{a}_m) \in R^{m\times d}$, $\boldsymbol{b} = (b_1; \ldots; b_m) \in R^m$.

该问题存在闭式最优解 $x^* = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{b}$. 计算复杂度为 $O(md^2)$.

线性规划问题形如

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} \leqslant b_i \quad (i = 1, \ldots, m) \ , \end{aligned} \qquad (1.29)$$

其中 $\boldsymbol{c}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{\boldsymbol{m}} \in R^d$, $b_1, \ldots, b_m \in R$.

该问题虽无闭式解，但已有许多成熟的求解算法，当 $m \geq d$ 时计算复杂度仅为 $O(md^2)$.

一般优化问题形式

一般的，一个优化问题可以表示为

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) \leqslant 0 \quad (i=1,\ldots,m)\ , \end{aligned} \tag{1.30}$$

其中 $f:\mathbb{R}^d \mapsto \mathbb{R}$ 称为优化目标函数，$h_i:\mathbb{R}^d \mapsto \mathbb{R}$ $(i=1,\ldots,m)$ 称为约束函数.

该问题的最优解可以表达为 $\{\boldsymbol{x}^* \mid f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \ \ (\forall \boldsymbol{x} \in \Omega)\}$，其中 $\Omega = \{\boldsymbol{x} \mid h_i(\boldsymbol{x}) \leq 0 \ \ (i=1,\ldots,m)\}$ 称为可行域.

凸优化及其最优解

若式(1.30)中的目标函数和约束函数都是凸的，则该优化问题就是凸优化问题.

当目标函数 $f(\cdot)$ 可微时，$\boldsymbol{x}^*$ 是凸优化问题的最优解当且仅当 $\boldsymbol{x}^* \in \Omega$ 且 $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{z}-\boldsymbol{x}^*) \geq 0 \quad (\forall \boldsymbol{z} \in \Omega)$. 直观来看，$-\nabla f(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 处定义了可行域 Ω 的一个支撑面. 如下图1.5所示.

对于无约束的凸优化问题，$\boldsymbol{x}^*$ 是最优解当且仅当 $\boldsymbol{x}^* \in \Omega$ 且 $\nabla f(\boldsymbol{x}^*) = 0$.

值得注意的是凸优化问题的一些性质：

- 任意一个局部最优解都是全局最优解
- 通常可以在多项式时间内求解

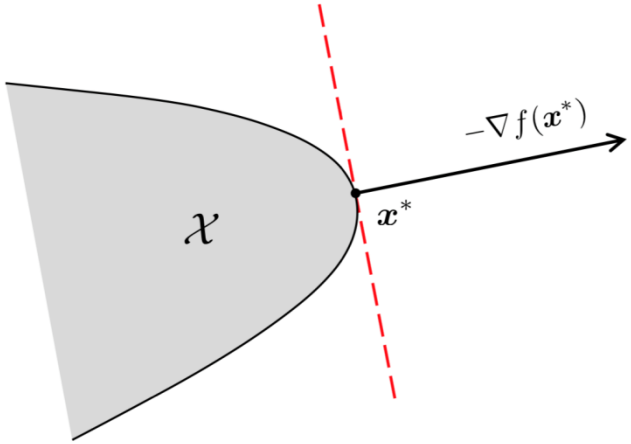


图 1.5 $-\nabla f(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 处定义了可行域 Ω 的一个支撑面

主问题

一个优化问题可以从两个角度来考察，即主问题和对偶问题.

主问题是(1.30)这样的原问题，或显示列出m 个不等式约束和 n 个等式约束写为

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) \leqslant 0 \quad (i=1,\ldots,m)\ , \\ & g_j(\boldsymbol{x}) = 0 \quad (j=1,\ldots,n)\ . \end{aligned} \tag{1.31}$$

我们假设其可行域 $\Omega \subset \mathbb{R}^d$ 非空，并将目标函数最优值记为 p^*.

对偶问题

对优化问题(1.31)，引入拉格朗日乘子 $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^{\mathrm{T}}$ 和 $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}}$，相应的拉格朗日函数 $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}$ 为

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x}) \ , \qquad (1.32)$$

其中 λ_i 和 μ_j 是分别针对不等式约束$h_i(\boldsymbol{x}) \leq 0$和等式约束 $g_j(\boldsymbol{x}) = 0$ 引入的拉格朗日乘子.

相应的拉格朗日对偶函数 $\Gamma : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}$ 为

$$\begin{aligned} \Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \inf_{\boldsymbol{x} \in \Omega} \ L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &= \inf_{\boldsymbol{x} \in \Omega} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x}) \right) \ . \qquad (1.33) \end{aligned}$$

对偶问题

由(1.31)可知，对于任意 $\lambda \succeq 0$ 都有

$$\sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x}) \leqslant 0 \ , \tag{1.34}$$

对于任意 $\tilde{\boldsymbol{x}} \in \Omega$ 有

$$\Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \Omega} \ L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \ \leqslant \ L(\tilde{\boldsymbol{x}}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \ \leqslant \ f(\tilde{\boldsymbol{x}}) \ , \tag{1.35}$$

于是，对任意 $\lambda \succeq 0$ 都有

$$\Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leqslant p^* \ , \tag{1.36}$$

即对偶函数(1.33)给出了主问题(1.31)的目标函数最优值 p^* 的下界

对偶问题

基于对偶函数(1.33)可以定义

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \quad \Gamma(\boldsymbol{\lambda},\boldsymbol{\mu}) \quad \text{s.t.} \ \boldsymbol{\lambda} \succeq 0 \ , \tag{1.37}$$

这就是主问题(1.31)的对偶问题，其中 λ 和 μ 称为对偶变量.

由于$\Gamma(\lambda,\mu)$ 是一个凹函数，而对偶问题(1.37)试图最大化一个凹函数，因此它是凸优化问题, 且该问题的目标函数最优值 d^* 是主问题的目标函数最优值 p^* 的下界，即

$$d^* \leqslant p^*, \tag{1.38}$$

这称为弱对偶性. 若

$$d^* = p^* \tag{1.39}$$

则称为强对偶性.

强对偶性的成立条件

对于一般的优化问题，强对偶性通常不成立.

但是，若主问题为凸优化问题，例如:

- (1.31)中的 $f(\boldsymbol{x})$ 和 $h_i(\boldsymbol{x})$ 均为凸函数，$g_j(\boldsymbol{x})$ 为仿射函数
- 可行域 Ω 中至少有一处使不等式约束严格成立

则强对偶性成立.

此时，将拉格朗日函数(1.32)分别对原变量和对偶变量求导，再令导数等于零即可求解。

KKT条件

KKT条件可以刻画主问题与对偶问题的最优解之间的关系.

令 $\boldsymbol{x}^*$ 为主问题的(1.31)的最优解, $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ 为对偶问题(1.37)的最优解. 当强对偶性成立时,

$$
\begin{aligned}
f(\boldsymbol{x}^*) &= \Gamma(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \\
&= \inf_{\boldsymbol{x}\in\Omega} \left\{ f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i^* h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j^* g_j(\boldsymbol{x}) \right\} \\
&\leqslant f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* h_i(\boldsymbol{x}^*) + \sum_{j=1}^{n} \mu_j^* h_j(\boldsymbol{x}^*) \\
&\leqslant f(\boldsymbol{x}^*)
\end{aligned}
\tag{1.40}
$$

KKT条件

显然，(1.40)中的不等式应该取等号. 于是有以下两个条件必定成立：

- 互补松弛条件：

$$\lambda_i^* h_i(\boldsymbol{x}^*) = 0 \quad (i = 1, \ldots, m) \ , \tag{1.41}$$

即 $\lambda_i^* > 0 \ \Rightarrow \ h_i(\boldsymbol{x}^*) = 0$ 以及 $h_i(\boldsymbol{x}^*) < 0 \ \Rightarrow \ \lambda_i^* = 0$

- $\boldsymbol{x}^*$ 是下面问题的最优解：

$$\begin{aligned} \boldsymbol{x}^* &= \arg\min_{\boldsymbol{x} \in \Omega} L(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \\ &= \arg\min_{\boldsymbol{x} \in \Omega} \left\{ f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i^* h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j^* g_j(\boldsymbol{x}) \right\} \ . \end{aligned} \tag{1.42}$$

通常 Ω 为全集或 $\boldsymbol{x}^*$ 位于 Ω 内部，因此拉格朗日函数 $L(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ 在 $\boldsymbol{x}^*$ 处的梯度为0.

KKT条件

相应的，KKT条件由以下几部分组成：

（1）主问题约束：

$$\begin{cases} h_i(\boldsymbol{x}^*) \leqslant 0 \ \ (i = 1, \ldots, m) \\ g_j(\boldsymbol{x}^*) = 0 \ \ (j = 1, \ldots, n) \end{cases}$$

（2）对偶问题约束：$\boldsymbol{\lambda}^* \succeq 0$

（3）互补松弛条件：$\lambda_i^* h_i(\boldsymbol{x}^*) = 0 \ \ (i = 1, \ldots, m)$

（4）拉格朗日函数在 $\boldsymbol{x}^*$ 处的梯度为0：

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\boldsymbol{x}^*) + \sum_{j=1}^{n} \mu_j^* \nabla g_j(\boldsymbol{x}^*) = 0$$

KKT条件

KKT条件具有如下重要性质：

- 强对偶性成立时，对于任意优化问题，KKT条件是最优解的必要条件，即主问题最优解和对偶问题最优解一定满足KKT条件；

- 对于凸优化问题，KKT条件是充分条件，即满足KKT条件的解一定是最优解；

- 对于强对偶性成立的凸优化问题，KKT条件是充分必要条件，即 x^* 是主问题最优解当且仅当存在 $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ 满足KKT条件.

支持向量机

支持向量机是一类经典的机器学习方法.

给定训练样本集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$, $y_i \in \{-1, +1\}$，支持向量机试图找到恰好位于两类训练样本“正中间”的划分超平面.

如下图1.6所示，距离超平面最近的这几个训练样本点被称为支持向量，两个异类支持向量到超平面的距离之和 $\gamma = \frac{2}{\|\boldsymbol{w}\|}$ 称为间隔.

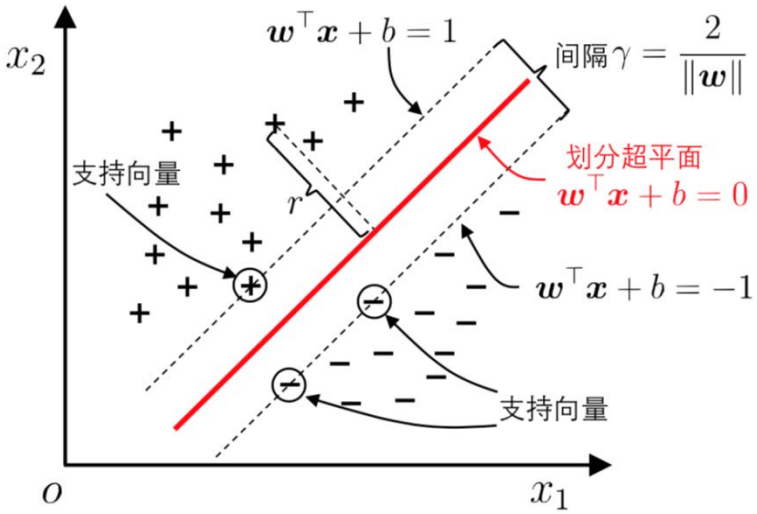


图 1.6 支持向量与间隔

支持向量机

假设超平面$(\boldsymbol{w}, b)$能将训练样本正确分类，即对于$(\boldsymbol{x}_i, y_i) \in D$，若$y_i = +1$，则有$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b > 0$；若$y_i = -1$，则有$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b < 0$.令

$$\begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \geqslant +1, & y_i = +1\ ; \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \leqslant -1, & y_i = -1\ , \end{cases} \tag{1.44}$$

则求解最大间隔划分超平面对应于优化问题

$$\begin{aligned} &\min_{\boldsymbol{w},b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t.} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, m. \end{aligned} \tag{1.45}$$

支持向量机

由(1.32)可知，(1.45)的拉格朗日函数为：

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b)\right) , \tag{1.46}$$

其中拉格朗日乘子 $\alpha_i \geq 0$, $\boldsymbol{\alpha} = (\alpha_1; \alpha_2; \ldots; \alpha_m)$.

令$L(\boldsymbol{w}, b, \boldsymbol{\alpha})$对 $\boldsymbol{w}$ 和 b 的偏导为零可得

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i , \tag{1.47}$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i . \tag{1.48}$$

支持向量机

将(1.47)代入(1.46)，即可将 $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ 中的 $\boldsymbol{w}$ 和 b 消去，再考虑(1.48)的约束，就得到主问题(1.45)的对偶问题

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j - \sum_{i=1}^{m} \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \ , \\ & \alpha_i \geqslant 0 \ , \quad i = 1, 2, \ldots, m \ . \end{aligned} \tag{1.49}$$

上述过程需满足KKT条件

$$\begin{cases} \alpha_i \geqslant 0 \ ; \\ y_i f(\boldsymbol{x}_i) - 1 \geqslant 0 \ ; \\ \alpha_i \left(y_i f(\boldsymbol{x}_i) - 1 \right) = 0 \ . \end{cases} \tag{1.50}$$

线性不可分问题

对原始空间中线性不可分的问题，可将样本从原始空间映射到一个高维特征空间.

例如图1.7中所示的“异或问题”在原始二维空间中线性不可分，但若将原始二维空间映射到合适的三维空间则变得线性可分.

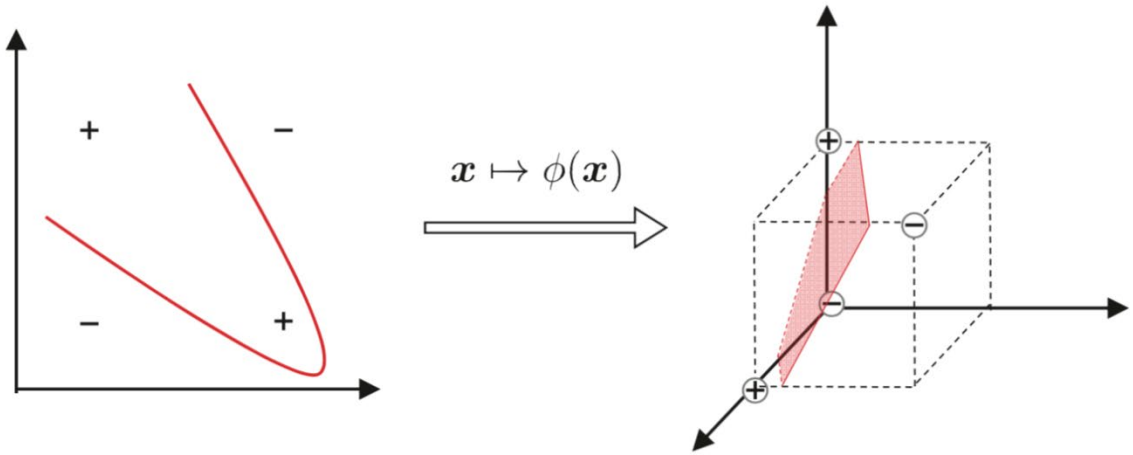


图 1.7 异或问题与非线性映射

线性不可分问题

引入非线性映射$\phi(\boldsymbol{x})$后，支持向量机求解的主问题(1.45)变成

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b) \geqslant 1, \quad i=1,2,\ldots,m, \end{aligned} \tag{1.51}$$

相应的，对偶问题(1.49)变成

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j) - \sum_{i=1}^{m}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0 \ , \\ & \alpha_i \geqslant 0 \ , \quad i=1,2,\ldots,m \ . \end{aligned} \tag{1.52}$$

核函数

注意到，(1.52)涉及到计算 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 映射到特征空间之后的内积 $\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$，当特征空间维数很高时，直接计算 $\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$ 通常很困难. 为此，考虑核函数

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j) \ , \tag{1.53}$$

即 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 在特征空间的内积等于它们在原始样本空间中通过 $\kappa(\cdot,\cdot)$ 计算的结果.

此时，(1.52)可重写为

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{m}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0 \ , \\ & \alpha_i \geqslant 0 \ , \quad i = 1, 2, \ldots, m \ . \end{aligned} \tag{1.54}$$

核函数

关于核函数有下面的定理：

定理 1.1 **核函数** 令 $\mathcal{X}$ 为输入空间, $\kappa(\cdot,\cdot)$ 是定义在 $\mathcal{X}\times\mathcal{X}$ 上的对称函数, 则 κ 是核函数当且仅当对于任意数据 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$, **核矩阵**(kernel matrix) $\mathbf{K}$ 总是半正定的. 这里 $\mathbf{K}$ 是一个 m 阶方阵, $\mathbf{K}_{ij}=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$.

表1.2列出了几种常用的核函数

表 1.2 常用核函数

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)^d$	$d\geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$	$\sigma>0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|}{\sigma}\right)$	$\sigma>0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\tanh(\beta\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j+\theta)$	tanh 为双曲正切函数, $\beta>0,\theta<0$

每个核函数都隐式地定义了一个特征空间，称为再生希尔伯特空间，其好坏对支持向量机的性能至关重要.

软间隔

目前为止，我们一直假设训练样本在样本空间或特征空间中线性可分. 然而，在现实任务中往往很难确定合适的满足这一条件的核函数，有时貌似可分的结果甚至可能是由于过拟合而造成的. 因此有必要允许支持向量机在少量样本上出错.

为此，引入软间隔的概念，允许某些样本不满足约束：

$$y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\geqslant 1\ . \tag{1.55}$$

在最大化间隔的同时，不满足约束的样本应尽可能少. 于是得到如下优化目标

$$\min_{\boldsymbol{w},b}\ \frac{1}{2}\|\boldsymbol{w}\|^2+\beta\sum_{i=1}^{m}\ell_{0/1}\left(y_i\left(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b\right)-1\right)\ , \tag{1.56}$$

其中$\beta>0$ 是一个常数，

$$\ell_{0/1}(x)=\begin{cases}1, & \text{if } x<0;\\ 0, & \text{otherwise.}\end{cases} \tag{1.57}$$

软间隔

由于 $\ell_{0/1}$ 非凸、不连续，(1.56)不易求解，因此支持向量机用hinge损失函数 $\ell_{hinge}(x) = \max(0, 1-x)$ 作为替代损失得到(1.56)的等价形式(1.59)

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \beta \sum_{i=1}^{m} \max\left(0, 1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b\right)\right) \ . \tag{1.59}$$

同时引入松弛变量 $\xi_i \geq 0$，可将(1.59)重写为

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \beta \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b) \geqslant 1 - \xi_i \\ & \xi_i \geqslant 0 \ , \ i = 1, 2, \ldots, m, \end{aligned} \tag{1.60}$$

即软间隔支持向量机优化的主问题.

软间隔

相应的，软间隔支持向量机优化的对偶问题为

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_{i=1}^{m}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0 \ , \\ & 0 \leqslant \alpha_i \leqslant \beta \ , \quad i = 1,2,\ldots,m \ . \end{aligned} \tag{1.61}$$

对比(1.61)和硬间隔下的对偶问题(1.54)可看出，两者的唯一差别就在于对偶变量的约束不同：前者是 $0 \leq \alpha_i \leq \beta$，后者是 $0 \leq \alpha_i$.