# Convex Optimization and Support Vector Machines

王璐

对于没有限制条件的凸优化，如果目标函数可导，则使得目标函数梯度为 $\mathbf{0}$ 的点是全局最小值点。本章以 Support Vector Machine (SVM) 为例，介绍带限制条件的凸优化问题的一般解法。SVM 是最好的监督学习 (supervised learning) 算法之一。监督学习就是从一些事先标记过的训练数据中建立一个模型或学习一个函数，这个模型或函数可以对输入的特征做出预测（输出）。

## 1 SVM: Margins

本节将介绍 SVM 的一个重要概念 – margin. Margin 代表一种预测的 "信心"。在 logistic 回归中，我们用 logistic 函数预测概率

$$P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x})}.$$

决策时可以采用如下规则：如果 $P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) \geq 0.5$, 预测 $Y = 1$, 反之 $Y = 0$. 或者等价地，如果 $\boldsymbol{\theta}^\top \boldsymbol{x} \geq 0$, 预测 $Y = 1$, 反之 $Y = 0$. $\boldsymbol{\theta}^\top \boldsymbol{x}$ 的值越大，$P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta})$ 越接近 1，我们对预测 $Y = 1$ 越有信心; 同理如果 $P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) \approx 0$, 我们对预测 $Y = 0$ 就越有信心。图1展示了一个数据集，其中 $\times$ 代表标记为 1 的点，$\circ$ 代表标记为 0 的点。图1中的实线是用这些训练数据得到的一条决策边界 (decision boundary): $\boldsymbol{\theta}^\top \boldsymbol{x} = 0$. 图1中的 A, B, C 是要预测的点，其中 A 点远离决策边界且 $\boldsymbol{\theta}^\top \boldsymbol{x}_A \gg 0$, 因此我们对预测 A 点值为 1 很有信心; 相反，C 点很靠近边界，尽管依据决策规则 ($\boldsymbol{\theta}^\top \boldsymbol{x}_C > 0$) 预测 C 点值为 1, 但只要稍微变动一下决策边界，C 点的预测可能就变为 0. 因此我们对 C 的预测没有对 A 的预测那么有信心，对 B 预测的信心介于两者之间。图1表明：当要预测的点越远离决策边界, 我们对它的预测越有信心。

SVM 既可以预测分类，也可以预测连续值，以下我们先从一个最简单的 0-1 线性 SVM 分类器 (classifier) 入手引入 margin 的概念。在训练集 $S = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ 中，每个点 $i$ 由一个特征 (feature) 向量 $\boldsymbol{x}_i$ 和一个标签 $y_i$ 组成，为了后续计算方便，令 $y_i \in \{-1, 1\}$ (注意不是 $\{0, 1\}$). 我们希望决策边界具有如下形式：

$$\boldsymbol{\omega}^\top \boldsymbol{x} + b = 0. \tag{1}$$

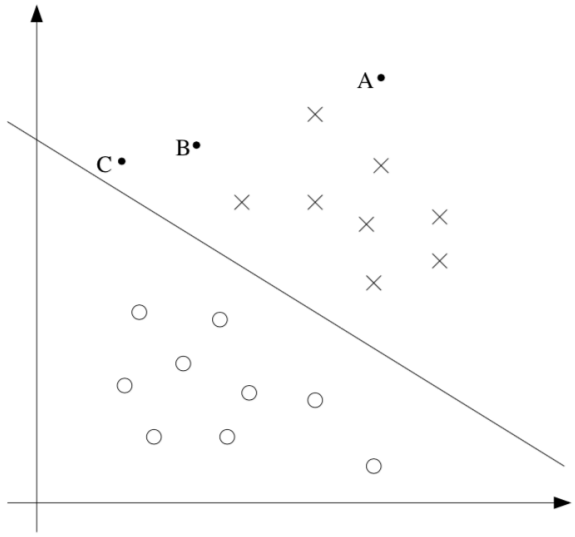


Figure 1: 对 A, B, C 三点的预测信心。Picture source: Andrew Ng

由于 SVM 对截距项的计算与其它系数不同，所以在(1)中将截距项 $b$ 单独写出来。有了决策边界，决策规则为：如果 $\boldsymbol{\omega}^\top \boldsymbol{x} + b \geq 0$, 预测 $y = 1$; 反之，预测 $y = -1$.

如果将点 $i$ 的 **margin** $\gamma_i$ 定义为

$$\gamma_i = y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b) \tag{2}$$

可以看到 $\gamma_i > 0$ 表明对点 $i$ 的预测是正确的，同时较大的 $\gamma_i$ 代表对预测值较大的信心。但是(2)有一个问题：如果将 $\boldsymbol{\omega}$ 和 $b$ 同时扩大 2 倍，决策边界不变，但是对预测的信心，即 margin $\gamma_i$ 却扩大了 2 倍。为了保证 margin 可识别，需要对(2)中的系数加一些规范化条件 (normalization condition)，比如令 $\|\boldsymbol{\omega}\|_2 = 1$ 或者令

$$\gamma_i = y_i \left( \frac{\boldsymbol{\omega}^\top \boldsymbol{x}_i + b}{\|\boldsymbol{\omega}\|_2} \right). \tag{3}$$

以下将 $\|\cdot\|_2$ 简记为 $\|\cdot\|$.

现在我们来考察一下定义(3)的**几何意义**。在图2中，A 点的坐标为 $\boldsymbol{x}_i$, 同时 $y_i = 1$; A 点在决策边界 $\boldsymbol{\omega}^\top \boldsymbol{x} + b = 0$ 上的投影为 $B$. 设 AB 的距离为 $d_{AB}$, 则由 A 点出发，沿着单位向量 $-\boldsymbol{\omega}/\|\boldsymbol{\omega}\|$ 走 $d_{AB}$ 个单位即到达 $B$ 点，所以 B 的坐标为 $\boldsymbol{x}_i - d_{AB}\boldsymbol{\omega}/\|\boldsymbol{\omega}\|$. 注意到 B 点在决策边界上，因此满足

$$\boldsymbol{\omega}^\top \left( \boldsymbol{x}_i - d_{AB} \frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|} \right) + b = 0.$$

解得

$$d_{AB} = \frac{\boldsymbol{\omega}^\top \boldsymbol{x}_i + b}{\|\boldsymbol{\omega}\|} \tag{4}$$

 

比较(3)和(4), 我们证明了 $y_i = 1$ 时，$\gamma_i$ 等于点 $i$ 到决策边界的距离。类似可证该结论对 $y_i = -1$ 的点也成立。

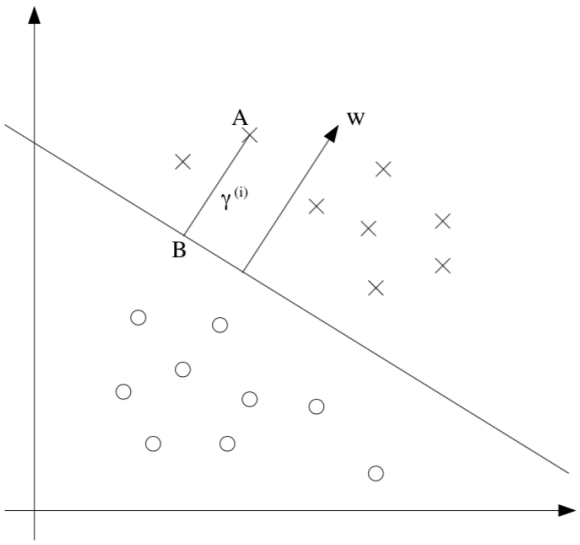


Figure 2: Margin 的几何意义。Picture source: Andrew Ng

假设训练集是线性可分的，即存在超平面 $\boldsymbol{\omega}^\top \boldsymbol{x} + b = 0$ 可以将正负点区分开。我们会在 Section **??** 中讨论线性不可分的情形。SVM 希望训练集中的所有点都远离决策边界，令

$$\gamma = \min_i \ \gamma_i$$

SVM 的目标是寻找一条决策边界使得 $\gamma$ 最大:

$$\max_{\boldsymbol{\omega},b} \ \gamma \ \ \text{s.t.} \ \ y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b)/\left\|\boldsymbol{\omega}\right\| \geq \gamma, \ i = 1, \ldots, n$$

等价于

$$\max_{\boldsymbol{\omega},b} \ \gamma \ \ \text{s.t.} \ \ y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b) \geq \gamma \left\|\boldsymbol{\omega}\right\|, \ i = 1, \ldots, n. \tag{5}$$

在(5)中令 $\|\boldsymbol{\omega}\| = 1/\gamma$, 则最大化 $\gamma$ 等价于最小化 $\|\boldsymbol{\omega}\|$:

$$\min_{\boldsymbol{\omega},b} \ \left\|\boldsymbol{\omega}\right\| \ \ \text{s.t.} \ \ y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b) \geq 1, \ i = 1, \ldots, n. \tag{6}$$

为了计算方便，将(6)写为以下形式:

$$\min_{\boldsymbol{\omega},b} \ \frac{1}{2}\left\|\boldsymbol{\omega}\right\|^2 \ \ \text{s.t.} \ \ -y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b) + 1 \leq 0, \ i = 1, \ldots, n. \tag{7}$$

此时我们将最大化最小 margin 的问题(5)转化为非常容易求解的优化问题(7). 当优化问题的目标函数和限制条件都是线性函数时，有通用的 linear programming 算法求解；当目标函数是二次函

 

数、限制条件是线性时，也有通用的 quadratic programming (QP) 算法。但是如果将(7)写为它的 Lagrange dual form, 可以设计出比通用的 QP 更高效的算法，这种算法还可以在很高维的空间高效地寻找最优的非线性 margin 决策曲面。为此我们需要先了解一些凸优化的理论。

## 2 Convex Optimization

考虑一般的凸优化问题：

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \\ &\text{s.t. } g_i(\boldsymbol{x}) \le 0,\ i=1,\dots,m \\ &\quad h_j(\boldsymbol{x}) = 0,\ j=1,\dots,p. \end{aligned} \tag{8}$$

其中函数 $f:\mathbb{R}^d\to\mathbb{R}$ 和 $g_i:\mathbb{R}^d\to\mathbb{R}$, $i=1,\dots,m$ 都是可导的 convex 函数，函数 $h_j:\mathbb{R}^d\to\mathbb{R}$, $j=1,\dots,p$ 都是仿射函数 (affine functions).

回顾 convex 函数和仿射函数的定义。如果函数 $g:G\to\mathbb{R}$ 满足 $G$ 是一个凸集且对于任意两点 $\boldsymbol{x}_1,\boldsymbol{x}_2\in G$, $\forall\theta\in[0,1]$ 有下式成立：

$$g\left(\theta\boldsymbol{x}_1+(1-\theta)\boldsymbol{x}_2\right)\le\theta g(\boldsymbol{x}_1)+(1-\theta)g(\boldsymbol{x}_2)$$

称 $g$ 是一个 **convex 函数**。

**仿射函数** $h:\mathbb{R}^d\to\mathbb{R}$ 的形式为 $h(\boldsymbol{x})=\boldsymbol{a}^\top\boldsymbol{x}+b$, 其中 $\boldsymbol{a}\in\mathbb{R}^d$, $b\in\mathbb{R}$. 仿射函数既是 convex 函数又是 concave 函数。

可以将带限制的优化问题(8)写为以下等价的无限制优化问题：

$$\min_{\boldsymbol{x}}\Theta_P(\boldsymbol{x})\triangleq f(\boldsymbol{x})+\infty\sum_{i=1}^m\mathbf{1}\left(g_i(\boldsymbol{x})>0\right)+\infty\sum_{j=1}^p\mathbf{1}\left(h_j(\boldsymbol{x})\ne0\right) \tag{9}$$

称(9)为 **primal optimization**. 但是(9)很难求解因为 primal objective $\Theta_P(\boldsymbol{x})$ 不连续更不可导。考虑用某种可导函数替换惩罚函数 $\infty\cdot\mathbf{1}(u>0)$，比如线性函数 $\alpha u$. 由于 $\infty\cdot\mathbf{1}(u>0)$ 只惩罚 $u>0$ 的部分，当 $\alpha\ge0$ 时，函数 $\alpha u$ 是 $\infty\cdot\mathbf{1}(u>0)$ 的一个下界函数，如图3所示。类似地，函数 $\beta u$ 总是 $\infty\cdot\mathbf{1}(u\ne0)$ 的一个下界函数（$\beta$ 的取值没有限制）。

定义 **Lagrangian**：

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})=f(\boldsymbol{x})+\sum_{i=1}^m\alpha_i g_i(\boldsymbol{x})+\sum_{j=1}^p\beta_j h_j(\boldsymbol{x}). \tag{10}$$

称(10)中 $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_m)$ 和 $\boldsymbol{\beta}=(\beta_1,\dots,\beta_p)$ 的元素为 Lagrange multipliers. 可以证明

$$\Theta_P(\boldsymbol{x})=\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\ \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\ \text{ s.t. }\alpha_i\ge0,\forall i. \tag{11}$$

 

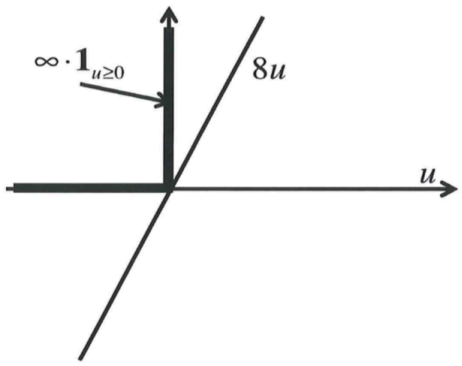


Figure 3: 惩罚函数 $\infty \cdot \mathbf{1}(u \geq 0)$ 和它的一个下界函数 $8u$. Picture source: Cynthia Rudin

*Proof.* 对于任意 $\boldsymbol{x}$,

- 如果某个限制条件成立：假设 $g_i(\boldsymbol{x}) \leq 0$, 为使 $\alpha_i g_i(\boldsymbol{x})$ 尽可能大，应该令 $\alpha_i = 0$; 如果 $h_j(\boldsymbol{x}) = 0$, 则 $\beta_j$ 取任何值都不会改变 $\mathcal{L}$ 的值。
- 如果某个限制条件不成立：假设 $g_i(\boldsymbol{x}) > 0$, 为使 $\alpha_i g_i(\boldsymbol{x})$ 尽可能大，应该令 $\alpha_i = \infty$ ; 如果 $h_j(\boldsymbol{x}) \neq 0$, 为使 $\beta_j h_j(\boldsymbol{x})$ 尽可能大 (达到 $\infty$)，应该令 $\beta_j = +\infty$ 或 $-\infty$.

因此通过调整 $\alpha_i$'s 和 $\beta_j$'s 的值总可以使(11)成立。 □

由(11)可知 $\Theta_P(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的 convex 函数。首先, $f(\boldsymbol{x})$ 是 convex 函数。其次，每个 $g_i(\boldsymbol{x})$ 是 convex 函数，$\alpha_i \geq 0$，因此每个 $\alpha_i g_i(\boldsymbol{x})$ 是 convex 函数。由于每个 $h_j(\boldsymbol{x})$ 是线性函数，不论 $\beta_j$ 的符号正或负, $\beta_j h_j(\boldsymbol{x})$ 总是 convex 函数。由于 convex 函数的和仍是 convex 函数，所以 $\mathcal{L}$ 是 $\boldsymbol{x}$ 的 convex 函数。最后，一列 convex 函数的上确界仍是 convex 函数。所以 $\Theta_P(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的 convex 函数。

根据(11)，可以将 primal optimization (9)转化为以下目标函数可导的优化问题：

$$\min_{\boldsymbol{x}} \Theta_P(\boldsymbol{x}) = \min_{\boldsymbol{x}} \left[ \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i \geq 0,\forall i} \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \right]. \tag{12}$$

如果点 $\boldsymbol{x}$ 满足所有限制条件，即 $g_i(\boldsymbol{x}) \leq 0, \forall i$ 且 $h_j(\boldsymbol{x}) = 0, \forall j$, 称点 $\boldsymbol{x}$ 为 **primal feasible**. 假设点 $\boldsymbol{x}^\star$ 使 $\Theta_P(\boldsymbol{x})$ 达到最小，最小值为 $p^\star = \Theta_P(\boldsymbol{x}^\star)$.

---

如果交换(12)中 min 和 max 的顺序，就得到了另一个不同的优化问题，称为(12)的 dual problem：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i \geq 0,\forall i} \left[ \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i \geq 0,\forall i} \Theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) \tag{13}$$

 

此处定义 dual objective $\Theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$. 如果 $\alpha_i \geq 0$, $i = 1,\ldots,m$, 称点 $(\boldsymbol{\alpha},\boldsymbol{\beta})$ 为 **dual feasible**. 假设 $(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)$ 使 $\Theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})$ 达到最大，最大值为 $d^\star = \Theta_D(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)$.

**Theorem 1.** *对任意一对 primal and dual problems (12)和(13), 总有 $d^\star \leq p^\star$.*

*Proof.* 如果点 $(\boldsymbol{\alpha},\boldsymbol{\beta})$ 是 dual feasible, 则以下下界关系成立：

$$\alpha_i g_i(\boldsymbol{x}) \leq \infty \cdot \mathbf{1}\left(g_i(\boldsymbol{x}) > 0\right),\ \forall i$$
$$\beta_j h_j(\boldsymbol{x}) \leq \infty \cdot \mathbf{1}\left(h_j(\boldsymbol{x}) \neq 0\right),\ \forall j.$$

因此

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \leq \Theta_p(\boldsymbol{x}),\ \forall \boldsymbol{x}.$$

两边关于 $\boldsymbol{x}$ 取最小值，不等式依然成立

$$\underbrace{\min_{\boldsymbol{x}}\ \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})}_{\Theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})} \leq \underbrace{\min_{\boldsymbol{x}}\ \Theta_p(\boldsymbol{x})}_{p^\star}. \tag{14}$$

(14)对所有 dual feasible 的点 $(\boldsymbol{\alpha},\boldsymbol{\beta})$ 都成立，因此

$$d^\star = \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i\geq 0,\forall i}\left[\min_{\boldsymbol{x}}\ \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right] \leq \min_{\boldsymbol{x}}\ \Theta_p(\boldsymbol{x}) = p^\star.$$

□

如果 primal and dual problems 满足 $d^\star = p^\star$，称这种情况为 **strong duality**. 很多条件可以保证 strong duality 成立，最常用的是 **Slater's condition**: 如果优化问题(8)的解 $\boldsymbol{x}^\star$ 使所有不等式限制条件都严格成立，即 $g_i(\boldsymbol{x}^\star) < 0$, $i = 1,\ldots,m$, 称 primal/dual problem pair 满足 Slater's condition.

## 2.1 KKT 条件

对于带限制的优化问题(8), 找到满足 KKT 条件的解等价于找到全局最优解 (global minimum).

之前我们用 $\boldsymbol{x}^\star$ 表示 primal optimization (12)的最优解，用 $(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)$ 表示 dual optimization (13)的最优解。当 strong duality 成立时，可得到以下结论。

**Lemma 1** (Complementary Slackness)**.** *如果 strong duality 成立，那么 $\alpha_i^\star g_i(\boldsymbol{x}^\star) = 0$, $i = 1,\ldots,m$.*

*Proof.* 由定义出发可得

$$d^\star = \Theta_D(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star) = \min_{\boldsymbol{x}}\ \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)$$

 

$$\leq \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) \tag{15}$$

$$\leq \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i\geq 0,\forall i} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \Theta_p(\boldsymbol{x}^\star) \tag{16}$$

$$= f(\boldsymbol{x}^\star) = p^\star \tag{17}$$

其中不等式(15)是因为 $\min_{\boldsymbol{x}}$ 小于任意 $\boldsymbol{x}$ 处的值, 当然包括 $\boldsymbol{x}^\star$; 同理可得不等式(16); 等式(17)成立是因为 primal optimization (12)的最优解一定是 primal feasible, 即满足所有限制条件。

当 strong duality 成立时，$d^\star = p^\star$，因此上式中的所有不等式都可以写为等式。此时有

$$\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = f(\boldsymbol{x}^\star) + \sum_{i=1}^m \alpha_i^\star g_i(\boldsymbol{x}^\star) + \sum_{j=1}^p \beta_j^\star h_j(\boldsymbol{x}^\star) = f(\boldsymbol{x}^\star) \tag{18}$$

所以

$$\sum_{i=1}^m \alpha_i^\star g_i(\boldsymbol{x}^\star) + \sum_{j=1}^p \beta_j^\star h_j(\boldsymbol{x}^\star) = 0$$

由于 $\boldsymbol{x}^\star$ 是 primal feasible，因此 $h_j(\boldsymbol{x}^\star) = 0$, $j = 1, \ldots, p$. 所以

$$\sum_{i=1}^m \alpha_i^\star g_i(\boldsymbol{x}^\star) = 0. \tag{19}$$

注意到

(i) 因为 $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 是 dual feasible, 所以 $\alpha_i^\star \geq 0$, $i = 1, \ldots, m$;

(ii) 因为 $\boldsymbol{x}^\star$ 是 primal feasible, 所以 $g_i(\boldsymbol{x}^\star) \leq 0$, $i = 1, \ldots, m$.

由 (i)(ii) 可得 $\alpha_i^\star g_i(\boldsymbol{x}^\star) \leq 0$, $i = 1, \ldots, m$，再由(19)得

$$\alpha_i^\star g_i(\boldsymbol{x}^\star) = 0,\ i = 1, \ldots, m.$$

□

**Remark**

- 由 Lemma 1可得，当 strong duality 成立时，在 primal/dual problem 的最优解 $(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 处有以下结论成立：
  - 如果某个 $\alpha_i^\star > 0$，则对应的 $g_i(\boldsymbol{x}^\star) = 0$，此时称该限制条件 $g_i$ 为 active constraint 或 binding constraint.
  - 如果某个 $g_i(\boldsymbol{x}^\star) < 0$, 则对应的 $\alpha_i^\star = 0$.

 

- 当 strong duality 成立时，根据 Lemma 1的证明，$\boldsymbol{x}^\star$ 是 convex 函数 $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 的最小值点，因此满足梯度为零：

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^\star) + \sum_{i=1}^{m} \alpha_i^\star \nabla_{\boldsymbol{x}} g_i(\boldsymbol{x}^\star) + \sum_{j=1}^{p} \beta_j^\star \nabla_{\boldsymbol{x}} h_j(\boldsymbol{x}^\star) = \mathbf{0}. \tag{20}$$

一般称等式(20)为 **Lagrangian stationarity**. (20)表明在最优解 $\boldsymbol{x}^\star$ 处，目标函数 $f$ 的梯度和限制函数的梯度方向相反，模长相同，如图4所示。图4中的曲线代表 $f$ 的等高线 (contours)，直线代表等式限制条件，在点 $\boldsymbol{x}^\star$ 处，目标函数 $f$ 的梯度和限制函数的梯度方向相反、模长相同，再沿直线移动 $\boldsymbol{x}$ 也不会进一步减小函数 $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 的值。由 Lemma 1的证明可知，此时 $\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = f(\boldsymbol{x}^\star)$，因此在限制条件满足的范围内再沿直线移动 $\boldsymbol{x}$ 也不会使 $f(\boldsymbol{x})$ 下降。

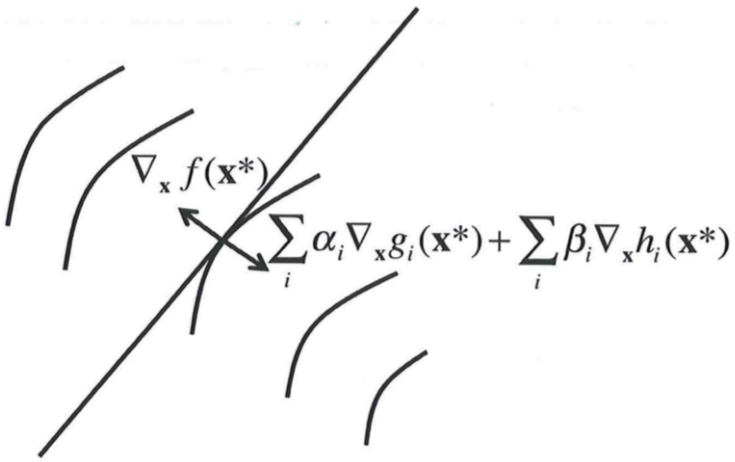


Figure 4: 最优解 $\boldsymbol{x}^\star$ 处目标函数的梯度与限制函数的梯度关系。Picture source: Cynthia Rudin

---

现在可以给出 primal dual optimization pair 的全局最优解满足的条件了，这些条件被称为 Karush-Kuhn-Tucker (KKT) 条件。

**Theorem 2** (KKT conditions)**.** 如果点 $\boldsymbol{x}^\star \in \mathbb{R}^d$, $\boldsymbol{\alpha}^\star \in \mathbb{R}^m$, $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ 满足以下条件：

- *(Primal feasibility)* $g_i(\boldsymbol{x}^\star) \leq 0$, $i = 1, \ldots, m$ 且 $h_j(\boldsymbol{x}^\star) = 0$, $j = 1, \ldots, p$.
- *(Dual feasibility)* $\alpha_i^\star \geq 0$, $i = 1, \ldots, m$.
- *(Complementary Slackness)* $\alpha_i^\star g_i(\boldsymbol{x}^\star) = 0$, $i = 1, \ldots, m$.

 

- *(Lagrangian stationary)* $\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = \mathbf{0}$.

则 $\boldsymbol{x}^\star$ 是 *primal optimal,* $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 是 *dual optimal. 如果 strong duality 成立*, 则任何 *primal optimal* $\boldsymbol{x}^\star$ *及任何 dual optimal* $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ *必须满足以上条件*。

**Remarks**

1. 如果 strong duality 不成立，KKT 条件是找到优化问题(8)全局最优解的充分条件。

2. 如果 strong duality 成立，KKT 条件是找到(8)全局最优解的充要条件。

历史上，KKT 条件最初是 Karush 在硕士论文 (1939) 中提出的，但没有引起任何注意，直到 1950 年两位数学家 Kuhn 和 Tucker 重新发现才获得关注。

 