ℓ_1、ℓ_2 正则化

刘盛

Update: 2020 年 11 月 14 日

1 预备知识

1.1 深度学习基础知识

- 泛化、欠拟合、过拟合等基础概念见花书第五章。

1.2 条件极值

- 详细内容见陈纪修版 < 数学分析 > 下册第十一章第七小节：条件极值问题与 Lagrange 乘数法

1.3 似然函数

- 离散型：假定一个关于参数 θ、具有离散型概率分布 P 的随机变量 X，则在给定 X 的输出 x 时，参数 θ 的似然函数可表示为:

$$L(\theta|x) = p_\theta(x) = P_\theta(X = x) \tag{1}$$

其中,$p(x)$ 表示 X 取 x 时的概率。上式常常写为 $P(X = x|\theta)$ 或者 $P(X = x;\theta)$ 。需要注意的是，此处并非条件概率，因为 θ 不（总）是随机变量。

- 连续型：假定一个关于参数 θ、具有连续概率密度函数 f 的随机变量 X，则在给定 X 的输出 x 时，参数 θ 的似然函数可表示为：

$$L(\theta|x) = f_\theta(x) \tag{2}$$

- 最大似然函数：给定一个概率分布 D，假定其概率密度函数（连续分布）或概率聚集函数（离散分布）为 f_D，以及一个分布参数 θ，我们可以从这个分布中抽出一个具有 n 个值的采样 $X_1, X_2, \cdots, X_n$，通过利用 f_D，我们就能计算出其概率：

$$P(x_1, x_2, \cdots, x_n) = f_D(x_1, x_2, \cdots, x_n|\theta) \tag{3}$$

要在数学上实现最大似然估计法，我们首先要定义可能性：

$$like(\theta) = f_D(x_1, x_2, \cdots, x_n|\theta) \tag{4}$$

并且在 θ 的所有取值上，使这个函数最大化。这个使可能性最大的值即被称为 θ 的最大似然估计。

1.4 协方差

- 协方差：$cov(X,Y)=E[(X-E(X))(Y-E(Y))]$
- 方差 $var(X)=cov(X,X)=E[(X-E(X))(X-E(X))]$

设 $X=(X_1,X_2,\cdots,X_N)^T$ 为 n 维随机变量，称矩阵：

$$C=(c_{ij})_{n\times n}=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} \tag{5}$$

为 n 维随机变量 X 的协方差矩阵（covariance matrix），也记为 $D(X)$，其中 $c_{ij}=Cov(X_i,X_j)$，$i,j=1,2,\cdots,n$

二维随机变量 (X_1,X_2) 的协方差矩阵为

$$C=\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \tag{6}$$

其中 $c_{11}=E[X_1-E(X_1)]^2$，$c_{12}=E[X_1-E(X_1)][X_2-E(X_2)]$，$c_{21}=E[X_2-E(X_2)][X_1-E(X_1)]$，$c_{22}=E[X_2-E(X_2)]^2$

由于 $c_{ij} = Cov\left(X_i, X_j\right)$ ，$i, j = 1, 2, \cdots, n$，所以所以协方差矩阵为对称非负定矩阵。

对多维随机变量 $X = [X_1, X_2, X_3, \cdots, X_n]^T$, 我们往往需要计算各维度两两之间的协方差, 这样各协方差组成了一个 $n \times n$ 的矩阵，称为协方差矩阵。协方差矩阵是个对称矩阵，对角线上的元素是各维度上随机变量的方差。我们定义协方差矩阵为 $\sum$，这个符号与求和符号 $\sum$ 相同，需要根据上下文区分。矩阵内的元素 $\sum_{ij}$ 为

$$\sum_{ij} = cov\left(X_i, X_j\right) = E\left[\left(X_i - E\left(X_i\right)\right)\left(X_j - E\left(X_j\right)\right)\right] \tag{7}$$

$$\begin{aligned}\Sigma &= \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^T\right] \\ &= \begin{bmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) & \cdots & \operatorname{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \cdots & \operatorname{cov}(X_n, X_n) \end{bmatrix}\end{aligned} \tag{8}$$

$$\begin{bmatrix} \mathrm{E}\left[\left(X_1 - \mathrm{E}\left[X_1\right]\right)\left(X_1 - \mathrm{E}\left[X_1\right]\right)\right] & \mathrm{E}\left[\left(X_1 - \mathrm{E}\left[X_1\right]\right)\left(X_2 - \mathrm{E}\left[X_2\right]\right)\right] & \cdots & \mathrm{E}\left[\left(X_1 - \mathrm{E}\left[X_1\right]\right)\left(X_n - \mathrm{E}\left[X_n\right]\right)\right] \\ \mathrm{E}\left[\left(X_2 - \mathrm{E}\left[X_2\right]\right)\left(X_1 - \mathrm{E}\left[X_1\right]\right)\right] & \mathrm{E}\left[\left(X_2 - \mathrm{E}\left[X_2\right]\right)\left(X_2 - \mathrm{E}\left[X_2\right]\right)\right] & \cdots & \mathrm{E}\left[\left(X_2 - \mathrm{E}\left[X_2\right]\right)\left(X_n - \mathrm{E}\left[X_n\right]\right)\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}\left[\left(X_n - \mathrm{E}\left[X_n\right]\right)\left(X_1 - \mathrm{E}\left[X_1\right]\right)\right] & \mathrm{E}\left[\left(X_n - \mathrm{E}\left[X_n\right]\right)\left(X_2 - \mathrm{E}\left[X_2\right]\right)\right] & \cdots & \mathrm{E}\left[\left(X_n - \mathrm{E}\left[X_n\right]\right)\left(X_n - \mathrm{E}\left[X_n\right]\right)\right] \end{bmatrix} \tag{9}$$

2 ℓ_1、ℓ_2 正则化来源推导

2.1 ℓ_1、ℓ_2 范数

范数，是具有“长度”概念的函数。在高等代数、泛函分析及相关的数学领域，范数是一个函数，是矢量空间内的所有矢量赋予非零的正长度或大小。半范数可以为非零的矢量赋予零长度。

K 若 N 是数域上的线性空间，泛函 $\|\cdot\| : X \to \mathbb{R}$ 满足：

1. 正定性：$\|x\| \geq 0$, 且$\|x\| = 0 \Leftrightarrow x = 0$
2. 正齐次性：$\|cx\| = |c|\|x\|$
3. 次可加性 (三角不等式)：$\|x + y\| \leq \|x\| + \|y\|$

则，$\|\cdot\|$ 成为 X 上的一个范数。

如果线性空间上定义了范数，则称之为赋范线性空间。

简单来说也就是范数其实在 $[0, \infty)$ 范围内的值，是向量的投影大小，在机器学习中一般会勇于衡量向量的距离。范数有很多种，我们常见的有 ℓ_1-norm 和 ℓ_2-norm，其实还有 ℓ_3-norm、ℓ_4-norm 等等，所以抽象来表示，我们会写作 ℓ_p-norm，一般表示为:

$$\|x\|_p = \left(\sum_i |x_i|^p \right)^{1/p} \tag{10}$$

对于上面这个抽象的公式，如果我们代入 p 值，若 p 为 1，则就是我们常说的 ℓ_1-norm：

$$\|x\|_1 = \sum_i |x_i| = |x_1| + |x_2| + \ldots + |x_i| \tag{11}$$

若 p 为 2，则是我们常说的 ℓ_2-norm：

$$\|x\|_2 = \sqrt{\left(\sum_i |x_i|^2\right)} = \sqrt{x1^2 + x2^2 + \ldots + x_i^2} \tag{12}$$

2.2 基于优化角度的正则化

对于模型权重系数 w 求解是通过最小化目标函数实现的，即求解：

$$\min_w J(w; X, y) \tag{13}$$

模型的复杂度可用 VC 维[1]来衡量。通常情况下，模型 VC 维与系数 w 的个数成线性关系：即 w 数量越多，VC 维越大，模型越复杂[2]。因此，为了限制模型的复杂度，很自然的思路是减少系数 w 的个

1VC 维的定义https://www.jianshu.com/p/9e02fd1bdaa4

2https://blog.csdn.net/cjianwyr/article/details/54907917

- 如果 VC 维很小，那么发生预测偏差很大的坏事情的可能性也就很小，那这有利于 Ein(g) 接近 Eout(g)；但是，这是我们的假设空间的表达能力受到了限制，这样 Ein(g) 可能就没有办法做到很小。
- 如果 VC 维很大，那么假设空间的表达能力很强，我们很有可能选到一个 Ein(g) 很小的假设，但是 Ein(g) 和 Eout(g) 之差很大的坏事情发生的情况发生的可能性就变得很大，这样 Ein(g) 和 Eout(g) 根本不接近，我们就无法确定选择的假设在测试数据的时候表现的很好。

数，即让 w 向量中一些元素为 0 或者说限制 w 中非零元素的个数。为此，可在原优化问题中加入一个约束条件：

$$\min_{w} J(w; X, y) \quad \text{s.t. } \|w\|_0 \leq C \tag{14}$$

$\| \|_0$ 范数表示向量中非零元素的个数。但由于该问题是一个 NP 问题，不易求解，为此我们需要稍微松弛一下约束条件。为了达到近似效果，我们不严格要求某些权重 w 为 0，而是要求权重 w 应接近于 0，即尽量小。从而可用 ℓ_1、ℓ_2 范数来近似 ℓ_0 范数，即：

$$\min_{w} J(w; X, y) \quad \text{s.t. } \|w\|_1 \leq C \tag{15}$$

$$\min_{w} J(w; X, y) \quad \text{s.t. } \|w\|_2 \leq C \tag{16}$$

其中 $\alpha > 0$，我们假设 α 的最优解为 α^* ，则对拉格朗日函数求最小化等价于：

$$\min_{w} J(w; X, y) + \alpha^* \|w\|_1 \tag{17}$$

$$\min_{w} J(w; X, y) + \alpha^* \|w\|_2 \tag{18}$$

可以看出，上式与 $\min_w \tilde{J}(w; X, y)$ 等价。

故此，从优化的角度，ℓ_1、ℓ_2 正则化的可以表述为：

- ℓ_1 正则化等价于在原优化目标函数中增加约束条件 $\|w\|_1 \leq C$
- ℓ_2 正则化等价于在原优化目标函数中增加约束条件 $\|w\|_2 \leq C$

2.3 基于最大后验概率估计的正则化

我们将模型预测结果和真实标签的差值定义为残差 (residual)：$\epsilon = y - f(x)$。如果每一次的观测都属于独立事件，所有观测误差的期望和方差应该都一致：这符合中心极限定理，应该构成正态分布，并且误差的期望值应该是 0。所以大多数情况下，可以认为这个误差服从高斯分布，如下：

$$\epsilon \sim N\left(0, \sigma^2\right) \tag{19}$$

于是可以得到我们的观测到的标签服从如下高斯分布：$y \sim N\left(f(x), \sigma^2\right)$。此时，我们定义了产出观测数据的模型，处于“模型已定，参数未知”的情况，找到一组参数使我们观测到一系列 y 的概率最大 (最大似然估计的思路)。假设权重 w 是未知的参数，从而求得对数似然函数：

$$l(w) = \log[P(y \mid X; w)] = \log\left[\prod_i P\left(y^i \mid x^i; w\right)\right] \tag{20}$$

通过假设 y^i 的不同概率分布，即可得到不同的模型。例如若假设 $y^i \sim N\left(w^T x^i, \quad \sigma^2\right)$ 的高斯分布，则有：

$$l(w) = \log\left[\prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y^i - w^T x^i\right)^2}{2\sigma^2}}\right] = -\frac{1}{2\sigma^2}\sum_i \left(y^i - w^T x^i\right)^2 + C \tag{21}$$

式中 C 为常数项，在求解最大或者最小问题时，常数项和系数项并不影响 $l(w)$ 的解，因而可令 $J(w; X, y) = -l(w)$ 即可得到线性回归的代价函数。

在最大后验概率估计中，则将权重 w 看作随机变量，也具有某种分布，从而有：

$$P(w \mid X, y) = \frac{P(w, X, y)}{P(X, y)} = \frac{P(y \mid X, w)P(w)}{P(X, y)} \propto P(y \mid X, w)P(w) \tag{22}$$

同样取对数有：

$$\text{MAP} = \log P(y \mid X, w)P(w) = \log P(y \mid X, w) + \log P(w) \tag{23}$$

由式(23)可知看出后验概率函数是在似然函数的基础上增加了一项 $\log P(w)$ 。$P(w)$ 的意义是对权重系数 w 的概率分布的先验假设，在收集到训练样本 $\{X, y\}$ 后，则可根据 w 在 $\{X, y\}$ 下的后验概率对 w 进行修正，从而做出对 w 更好地估计。

若假设 w_j 的先验分布为 0 均值的高斯分布，即 $w_j \sim N\left(0, \sigma^2\right)$，则有

$$\log P(w) = \log \prod_j P\left(w_j\right) = \log \prod_j \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(w_j\right)^2}{2\sigma^2}} \right] = -\frac{1}{2\sigma^2} \sum_j w_j^2 + C' \tag{24}$$

可以看到，在高斯分布下 $\log P(x)$ 的效果等价于在代价函数中增加 ℓ_2 正则项。

若假设 w_j 服从均值为 0、参数为 a 的拉普拉斯分布，即：

$$P\left(w_j\right) = \frac{1}{\sqrt{2a}} e^{\frac{-\left|w_j\right|}{a}} \tag{25}$$

则有：

$$\log P(w) = \log \prod_j \frac{1}{\sqrt{2a}} e^{\frac{-\left|w_j\right|}{a}} = -\frac{1}{a} \sum_j \left|w_j\right| + C' \tag{26}$$

可以看到，在高斯分布下 $\log P(x)$ 的效果等价于在代价函数中增加 ℓ_1 正则项。

故此，从最大后验概率估计的角度，ℓ_1、ℓ_2 正则化的可以表述为：

- ℓ_1 正则化可通过假设权重 w 的先验分布为拉普拉斯分布，由最大后验概率估计导出；
- ℓ_2 正则化可通过假设权重 w 的先验分布为高斯分布，由最大后验概率估计导出。

3 ℓ_1 、ℓ_1 正则化效果分析

3.1 基于图像化的正则化

考虑带约束条件的优化解释，对 ℓ_1 正则化为：

$$\min_{w} J(w; X, y) \text{s.t. } \|w\|_2 \leq C \tag{27}$$

该问题的求解示意图如下所示：

图1中椭圆为原目标函数 $J(w)$ 的一条等高线，圆为半径 $\sqrt{C}$ 的 ℓ_2 范数球。由于约束条件的限制，w 必须位于 ℓ_2 范数球内。考虑边界上的一点 w ，图1中蓝色箭头为 $J(w)$ 在该处的梯度方向 $\nabla J(w)$ ，红色箭头为 ℓ_2 范数球在该处的法线方向。由于 w 不能离开边界（否则违反约束条件），因而在使用梯度下降法更新 w 时，只能朝 $\nabla J(w)$ 在 ℓ_2 范数球上 w 处的切线方向更新，即图1中绿色箭头的方向。如此 w 将沿着边界移动，当 $\nabla J(w)$ 与 ℓ_2 范数球上 w 处的法线平行时，此时 $\nabla J(w)$ 在切线方向的分量为 0，w 将无法继续移动，从而达到最优解 w^*（图 w 中红色点所示）。

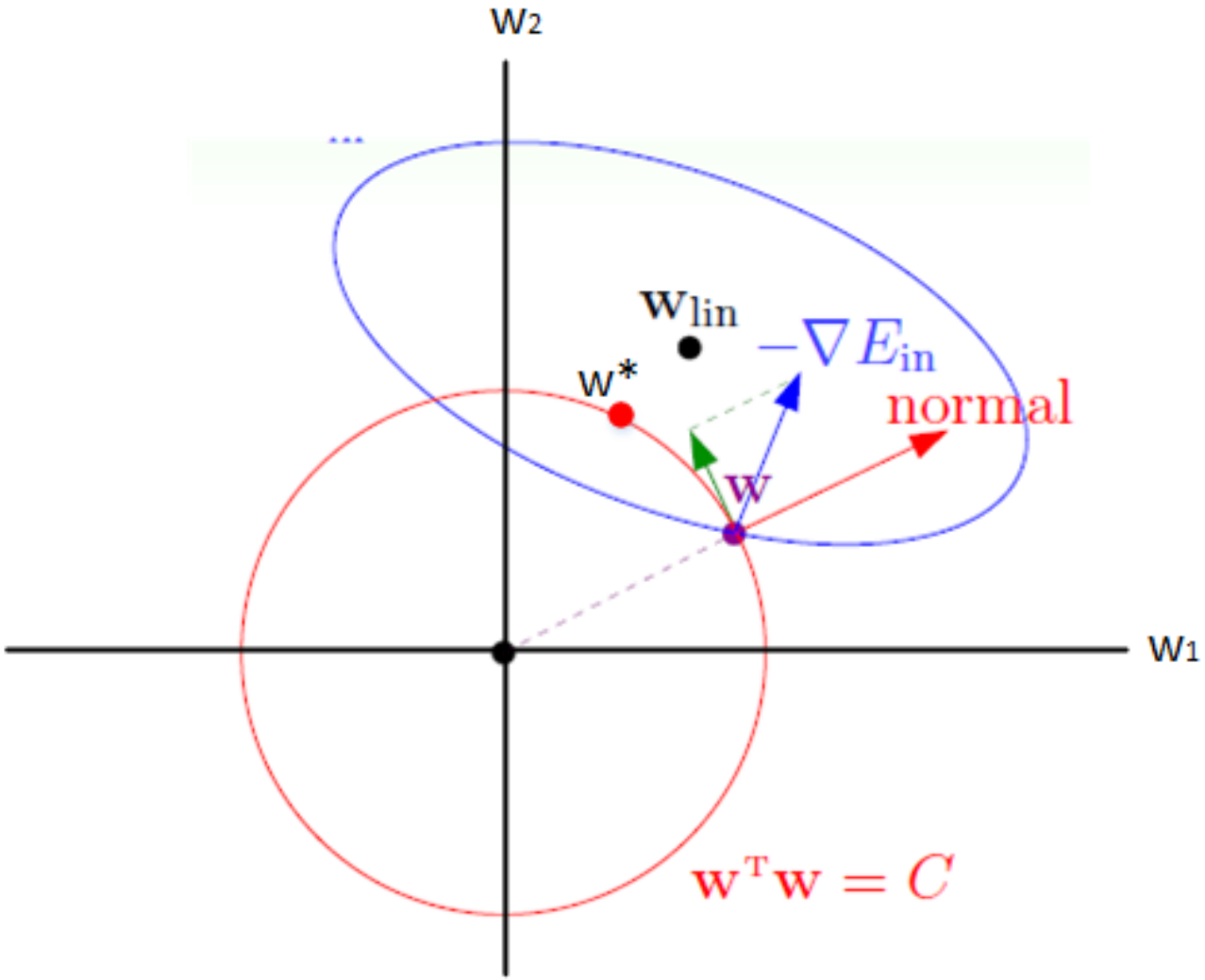


图 1: ℓ_2 正则化图解

从图1可知，可以得到一个很自然的问题：存在限定条件下，w 最终会在什么位置取得最优解呢？也就是说在满足限定条件的基础上，如何让 E_{in} 最小。

由图1可知，w 是沿着圆的切线方向运动，如图1绿色箭头所示。运动方向与 w 的方向（红色箭头方向）垂直。运动过程中，根据向量知识，只要 $-\nabla E_{in}$ 与运行方向有夹角，不垂直，则表明 $-\nabla E_{in}$ 仍会在 w 切线方向上产生分量，那么 w 就会继续运动，寻找下一步最优解。只有当 $-\nabla E_{in}$ 与 w 的切线方向垂直时，$-\nabla E_{in}$ 在 w 的切线方向才没有分量，这时候 w 才会停止更新，到达最接近 w_{lin} 的位置，且同时满足限定条件。$-\nabla E_{in}$ 与 w 的切线方向垂直，即 $-\nabla E_{in}$ 与 w 的方向平行。如上图所示，蓝色箭

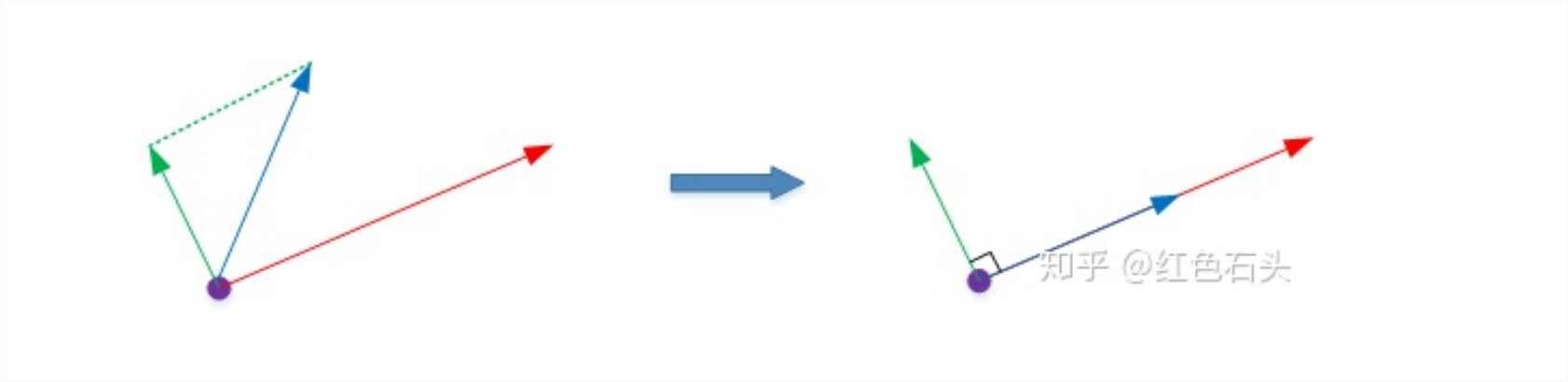


图 2: 梯度下降目标

头和红色箭头互相平行。这样，根据平行关系得到：

$$-\nabla E_{in} + \lambda w = 0 \tag{28}$$

由式(28)可知, 优化目标和限定条件整合在一个式子中了。换句话说只要在优化 E_{in} 的过程中满足

式(28)，就能实现正则化目标。

根据最优化算法的思想：梯度为 0 的时候，函数取得最优值。已知 ∇E_{in} 是 E_{in} 的梯度，若 λw 也能看成是某个表达式的梯度，则新的优化目标函数就可以表述为原目标优化函数 + 推导出的表达式。

显然 λw 可以看成是 $\frac{1}{2}\lambda w^2$ 的梯度：

$$\frac{\partial}{\partial w}\left(\frac{1}{2}\lambda w^2\right) = \lambda w \tag{29}$$

根据平行关系求得的等式，构造一个新的损失函数：

$$E_{aug} = E_{in} + \frac{\lambda}{2}w^2 \tag{30}$$

之所以这样定义，是因为对 $Eaug$ 求导，恰好得到上面所求的平行关系式。可以知道，式(30)右边第二项就是 ℓ_2 正则化项。

ℓ_1 正则化项的推导只需要将圆形约束条件改为正方形的约束条件就可以推导出 ℓ_1 正则化. 从图像化的角度，分析了 ℓ_1，ℓ_2 正则化的物理意义，推导 ℓ_1, ℓ_2 正则化项的损失函数。

ℓ_1、ℓ_2 范数是损失函数里的一个正则化项，作用就是降低模型复杂度，减小过拟合的风险。这里的正则化项，存在的目的就是作为一个“惩罚项”，对损失函数中的某一些参数做一些限制，是结构风险最小化策略的体现，就是选择经验风险（平均损失函数）和模型复杂度同时较小的模型。

针对线性回归模型，假设对其代价函数里加入正则化项，其中 ℓ_1、ℓ_2 正则化项的表示分别如式(31)所示，其中 $\lambda >= 0$，是用来平衡正则化项和经验风险的系数。

1. 使用 ℓ_1 范数正则化，其模型也被叫作 Lasso 回归（Least Absolute Shrinkage and Selection Operator，最小绝对收缩选择算子）
2. 使用 ℓ_2 范数正则化，其模型被叫做 Ridge 回归，中文为岭回归。

$$\begin{aligned}\mathbf{w}^* &= \arg\min_{\mathbf{w}} \sum_j \left(t\left(\mathbf{x}_j\right) - \sum_i w_i h_i\left(\mathbf{x}_j\right) \right)^2 + \lambda \sum_{i=1}^{k} |w_i| \\ \mathbf{w}^* &= \arg\min_{\mathbf{w}} \sum_j \left(t\left(\mathbf{x}_j\right) - \sum_i w_i h_i\left(\mathbf{x}_j\right) \right)^2 + \lambda \sum_{i=1}^{k} w_i^2\end{aligned} \tag{31}$$

此时，加入 ℓ_2 正则化项的损失函数为:

$$\mathcal{L} = \mathcal{L}_0 + \frac{\lambda}{2} \sum_{i=1} w_i^2 \tag{32}$$

式(32)中，第一项 $\mathcal{L}_0$ 是原始损失函数; 第二项是 ℓ_2 正则化项，它是所有权重的平方和，通过一个因子 $\frac{\lambda}{2}$ 进行量化调整可以看到正则化项不包含偏置，对通过系数 λ 权衡正则化项和原始损失函数的比重。从定义来看，正则化的效果会使得网络倾向于学习小一点的权重，否则第一项 $\mathcal{L}_0$ 将明显变化。换言之，正则化其实是一种对追求小权重和最小化原始损失函数这两个目标进行权衡的过程。两个目标之间的相对重要性由正则化系数入控制:λ 越小，越倾向于以最小化原始损失函数为主要目标;λ 越大，越倾向于以追求小权重为主要目标。

对式(32)求偏导可得：

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w} &= \frac{\partial \mathcal{L}_0}{\partial w} + \lambda w \\ \frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial \mathcal{L}_0}{\partial b}\end{aligned} \tag{33}$$

式中，w 为权重;b 为偏置。可知，偏置的更新规则:

$$b \to b - \eta \frac{\partial \mathcal{L}_0}{\partial b} \tag{34}$$

权重的更新规则:

$$w \to w - \eta \frac{\partial \mathcal{L}_0}{\partial w} - \eta \lambda w = (1 - \eta \lambda)\, w - \eta \frac{\partial \mathcal{L}_0}{\partial w} \tag{35}$$

可以发现，偏置的更新与 ℓ_2 正则化项无关，不受正则化影响，而权重的更新与 L，正则化项有关。在此，仅考虑引入 ℓ_2 正则化项这一个因素: 在未引入 ℓ_2 正则化项时，w 的原系数为 l; 引入 ℓ_2 正则化项后，w 的系数为 $1 - \eta\lambda$，其中，η、λ 均为正数，因此现有系数小于 1，n 即小于原系数。因此，引入 ℓ_2，正则化项将使得参数 w 减小，这也是权重衰减概念的由来。当考虑后面的偏导项时，w 的更新可能变大也可能变小。

此时，加入 ℓ_1 正则化项的损失函数为:

$$\mathcal{L} = \mathcal{L}_0 + \lambda \sum_w |w| \tag{36}$$

从定义来看，ℓ_1 正则化的定义和 ℓ_2 正则化类似，即通过一个有关权重的参数范数惩罚项, 使得网络规避较大的权重并优先选择较小的权重。当然，ℓ_1 正则化和 ℓ_2 正则化并不完全相同，因此也不会得到与 ℓ_2 正则化完全相同的效果。

对 ℓ_1 正则化的损失函数进行求导:

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}_0}{\partial w} + \lambda \operatorname{sgn}(w) \tag{37}$$

式中，sgn(w) 表示 w 的符号。此时，关于权重 w 的更新规则为:

$$w \to w' = w - \eta\lambda \operatorname{sgn}(w) - \eta\frac{\partial \mathcal{L}_0}{\partial w} \tag{38}$$

与原始的更新规则相比，加入 ℓ_1 正则化的更新规则多了第二项 $-\eta\lambda sgn\,(w)$，其中，η、λ 均为正数。此时，若 w 为正，则更新后 w 趋向于更小的值; 若 w 为负，则更新后 w 趋向于更大的值。因此，加入 L 正则化的效果就是使 w 的更新趋近于 0。当 w 等于 时，绝对值不可导，此时只能应用原始的无正则化方法进行更新。这就相当于去掉了第二项，因此可以约定 sgn(0)=0，这样就把 w=0 的情况也统一进来了。考虑整个神经网络，ℓ_1 正则化使得网络权重都趋近于 0，也就相当于降低了网络复杂度，防止过拟合。

这两种情形下，正则化的效果都是使权重尽可能地缩小。这符合我们的预期，但缩小权重的方式有所不同。在 ℓ_2 正则化中，权重通过一个与 w 成比例的量进行缩小; 而在 ℓ_1 正则化中，权重通过一个常量向 靠近从而达到缩小的目的。因此，对于一个绝对值很大的特定权重，经过 ℓ_1 正则化更新权重的缩小程度远比经过 ℓ_2 正则化要大得多。相反，对于一个绝对值很小的特定权重，经过 ℓ_1 正则化更新权重的缩小程度远比经过 ℓ_2 正则化要小得多。这就导致了最终的结果: 与 ℓ_2 正则化相比，ℓ_1 正则化更倾向于将网络的权重聚集在相对少量的重要连接上，而权重则会趋近于 0。

4 ℓ_1 正则

4.1 次梯度

次梯度的定义如下：

对于凸函数 $f:\mathbb{D}\to\mathbb{R}$, 其中 $\mathbb{D}$ 为定义在 n 维欧式空间上的凸集。若对 $\forall_y\in\mathbb{D}$ 有如下表达式成立:

$$f(y)\geq f(x)+<\nabla f(x),y-x> \tag{39}$$

则称向量 $\geq f(x)$ 为 f 在 x 处的次梯度。

4.2 Soft Thresholding Function

下面两种软阈值函数是等价的。

$$soft(x,T)=sign(x)\max\{|x|-T,0\}\Leftrightarrow soft(x,T)=\begin{cases}x+T & x\leq -T\\ 0 & |x|\leq T\\ x-T & x\geq T\end{cases} \tag{40}$$

其他等价形式3：

$$\eta_S(w,\lambda) = \mathrm{sgn}(w)(|w| - \lambda)_+ \tag{41}$$

软阈值 (Soft-Thresholding) 可以求解如下优化问题：

$$\arg\min \|X - B\|_2^2 + \lambda\|X\|_1 \tag{42}$$

其中，$X = [x_1, x_2, \cdots, x_n]^r$,$B = [k_1, b_2, \cdots, b_n]^r$

根据范数的定义，可以将上面优化问题的目标函数拆开：

$$\begin{aligned} F(X) &= \|X - B\|_2^2 + \lambda\|X\|_1 \\ &= \left[(x_1 - b_4)^2 + \lambda \mid x_1|\right]\right] + \left[(x_2 - b_2)^2 + \lambda\,|x_2|\right] + \cdots \left[(x_N - b_N)^2 + \lambda\,|x_N|\right] \end{aligned} \tag{43}$$

也就是说，我们可以通过求解 N 个独立的形如函数:

$$f(x) = (x - b)^2 + \lambda|x| \tag{44}$$

的优化问题，来求解式(42)。由数分中求极值方法可得，可以求函数 f(x) 导数：

$$\frac{df(x)}{dx} = 2(x - b) + \lambda\,\mathrm{sgn}(x) \tag{45}$$

这里要解释一下变量 x、绝对值 |x|、符号函数 sgn(x) 三者之间的关系 (数分上有)。

$$x = \mathrm{sgn}\,(x)\,|x| \Leftrightarrow |x| = \frac{x}{\mathrm{sgn}\,(x)} \Leftrightarrow |x| = \mathrm{sgn}\,(x)\,x \tag{46}$$

3sgn(x) 与 sign(x) 都是表示符号函数，下面在推导时使用的是 sgn(x) 记号。

当 x>0 时，|x|=x，因此其导数等于 sgn(x)=1；当 x<0 时，|x|=-x，因此其导数等于 sgn(x)=-1；综合起来，x 绝对值的导数等于 sgn(x)。令函数 f(x) 导数等于 0，得：

$$x = b - \frac{\lambda}{2}\operatorname{sgn}(x) \tag{47}$$

这个结果等号两端都有变量 x，需要再化简一下。下面分三种情况讨论[4]：

1. 当 b>$\lambda/2$ 时
   - 假设 x<0，则 sgn(x)=-1，所以 x=b+$\lambda/2$>0，与假设 x<0 矛盾；
   - 假设 x>0，则 sgn(x)=1，所以 x=b-$\lambda/2$>0，成立；

   此时在 x=b-$\lambda/2$>0 处取得极小值：

$$\begin{aligned} f(x)|_{\mathrm{x=b-\lambda/2}} &= \left(b - \frac{\lambda}{2} - b\right)^2 + \lambda\left(b - \frac{\lambda}{2}\right) = b\lambda - \frac{\lambda^2}{4} \\ &= -\left(\frac{\lambda}{2}\right)^2 + 2b\frac{\lambda}{2} \\ &= -\left(\frac{\lambda}{2}\right)^2 + 2b\frac{\lambda}{2} - b^2 + b^2 \\ &= -\left(\frac{\lambda}{2} - b\right)^2 + b^2 < b^2 = f(0) \end{aligned} \tag{48}$$

4一般情况下正则化参数都是大于零的，这里 λ 也是大于零的

即此时极小值小于 f(0)，而当 x<0 时

$$\begin{aligned}\frac{df(x)}{dx} &= 2(x-b)+\lambda\,\mathrm{sgn}(x)\\ &= 2x-2b-\lambda\\ &< 2x-2\left(\frac{\lambda}{2}\right)-\lambda\\ &= 2x-2\lambda<0\end{aligned} \tag{49}$$

即当 x<0 时函数 f(x) 为单调减函数（对任意 $\Delta x<0$，$f(0)<f(\Delta x)$）。因此，函数在 $x=b-\lambda/2>0$ 处取得最小值。

2. 当 $b<-\lambda/2$ 时
   - 假设 x<0，则 sgn(x)=-1，所以 $x=b+\lambda/2<0$，成立；
   - 假设 x>0，则 sgn(x)=1，所以 $x=b-\lambda/2<0$，与假设 x<0 矛盾；

   所以此时在 $x=b+\lambda/2<0$ 处取得极小值：

$$\begin{aligned}f(x)|_{x\to\Delta A} &= \left(b+\frac{\lambda}{2}-b\right)^2-\lambda\left(b+\frac{\lambda}{2}\right)=-b\lambda-\frac{\lambda^2}{4}\\ &= -\left(\frac{\lambda}{2}\right)^2-2b\frac{\lambda}{2}\\ &= -\left(\frac{\lambda}{2}\right)^2-2b\frac{\lambda}{2}-b^2+b^2\\ &= -\left(\frac{\lambda}{2}+b\right)^2+b^2<b^2=f(0)\end{aligned} \tag{50}$$

即此时极小值小于 f(0)，而当 x>0 时

$$\begin{aligned}\frac{df(x)}{dx} &= 2(x-b)+\lambda\,\mathrm{sgn}(x)\\ &= 2x-2b+\lambda\\ &> 2x-2\left(-\frac{\lambda}{2}\right)+\lambda\\ &= 2x+2\lambda>0\end{aligned} \tag{51}$$

即当 x>0 时函数 f(x) 为单调增函数（对任意 $\Delta x>0$，f(Δx>f(0)）。因此，函数在 x=b+$\lambda/2<0$ 处取得最小值。

3. 当-$\lambda/2$<b<$\lambda/2$ 时 (即 |b|<$\lambda/2$ 时)
   - 假设 x<0，则 sgn(x)=-1，所以 x=b+$\lambda/2>0$，与假设 x<0 矛盾；
   - 假设 x>0，则 sgn(x)=1，所以 x=b-$\lambda/2<0$，与假设 x<0 矛盾；

   即无论 x 为大于 0 还是小于 0 均没有极值点。

   下面讨论 x=0 是否为函数 f(x) 的极值点

   对于 $\Delta x\neq 0$

$$\begin{aligned}f(\Delta x) &= (\Delta x-b)^2+\lambda|\Delta x|\\ &= (\Delta x)^2-2\Delta x+b^2+\lambda|\Delta x|\\ &= (\Delta x)^2-2\Delta x b+\lambda|\Delta x|+f(0)\end{aligned} \tag{52}$$

当 $\Delta x > 0$ 时，利用条件 b<$\lambda/2$ 可得

$$\begin{aligned} f(\Delta x) &= (\Delta x)^2 - 2\Delta x b + \lambda \Delta x + f(0) \\ &> (\Delta x)^2 - 2\Delta x \frac{\lambda}{2} + \lambda \Delta x + f(0) \\ &= (\Delta x)^2 + f(0) > f(0) \end{aligned} \tag{53}$$

当 $\Delta x < 0$ 时，利用条件 b<$\lambda/2$ 可得 (注：此时 $|\Delta x|$=-Δx)

$$\begin{aligned} f(\Delta x) &= (\Delta x)^2 - 2\Delta x b + \lambda |\Delta x| + f(0) \\ &> (\Delta x)^2 - 2\Delta x \frac{\lambda}{2} + \lambda |\Delta x| + f(0) \\ &= (\Delta x)^2 + \lambda(-\Delta x) + \lambda |\Delta x| + f(0) \\ &= (\Delta x)^2 + 2\lambda |\Delta x| + f(0) > f(0) \end{aligned} \tag{54}$$

因此，函数在 x=0 处取得极小值，也是最小值。

综合以上三种情况，f(x) 的最小值在以下位置取得：

$$\arg\ \min f(x) = \begin{cases} b + \lambda/2 & , b < -\lambda/2 \\ 0 & |b| < \lambda/2 \\ b - \lambda/2 & , b > \lambda/2 \end{cases} \tag{55}$$

至此，我们可以得到优化问题

$$\arg\ \min_x \|X - B\|_2^2 + \lambda \|X\|_1 \tag{56}$$

的解为:

$$\text{soft}(B, \lambda/2) = \begin{cases} B + \lambda/2 & , B < -\lambda/2 \\ 0 & , |B| < \lambda/2 \\ B - \lambda/2 & , B > \lambda/2 \end{cases} \tag{57}$$

该式为软阈值函数 (Soft Thresholding Function) 的矩阵形式。

$$\hat{J}(w; \boldsymbol{X}, \boldsymbol{y}) = J\left(\boldsymbol{w}^*; \boldsymbol{X}, \boldsymbol{y}\right) + \sum_i \left[\frac{1}{2} H_{i,i} \left(w_i - w_i^*\right)^2 + \alpha \left|w_i\right|\right] \tag{58}$$

优化式(58)等价于优化这个目标函数：

$$arg\min \hat{J}\left(w; \boldsymbol{X}, \boldsymbol{y}\right) = arg\min \sum_i \left[\frac{1}{2} H_{i,i} \left(w_i - w_i^*\right)^2 + \alpha \left|w_i\right|\right] \tag{59}$$

5 proximal operator[5]

$$\min_{x \in \mathcal{X}} \; f_1(x) + f_1(x) \tag{60}$$

where $\mathcal{X}$ is a closed and convex set. Furthermore, f_1 is convex and lower semicontinuous; f_2 is convex and has Lipschitz continuous gradient.

5 A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization

Define the proximity operator $prox_{f_i}$:

$$prox_{f_i(x)} = argmin_{y \in \mathcal{X}} \ f_i(y) + \frac{1}{2}\|x - y\|^2 \tag{61}$$

The following forward-backward splitting iteration can be used to obtain a solution for problem:

$$x^{r+1} = prox_{\gamma f_1}\left(x^r - \gamma \nabla f_2\left(x^r\right)\right) \tag{62}$$

6 为什么 ℓ_1、ℓ_2 正则化可以防止过拟合

6.1 神经网络架构角度

当正则化 λ 设置得足够大，权重矩阵 被设置为接近于 0 的值 (ℓ_2 正则化是趋于零而不是等于 0,ℓ_1 正则化是等于 0)，直观理解就是把多隐藏单元的权重设为 0，于是基本上消除了这些隐藏单元的许多影响。

- 极端假设等于零的情况，这个被大大简化了的神经网络会变成一个很小的网络，小到如同一个逻辑回归单元，可是深度却很大，它会使这个网络从过度拟合的状态更接近左图的高偏差状态。

于是就有了两种情况，一种是 High Bias，另一种是 High Variance。当调节 λ 时，会存在一个中间值，使得结果会有一个接近“Just Right”的中间状态。

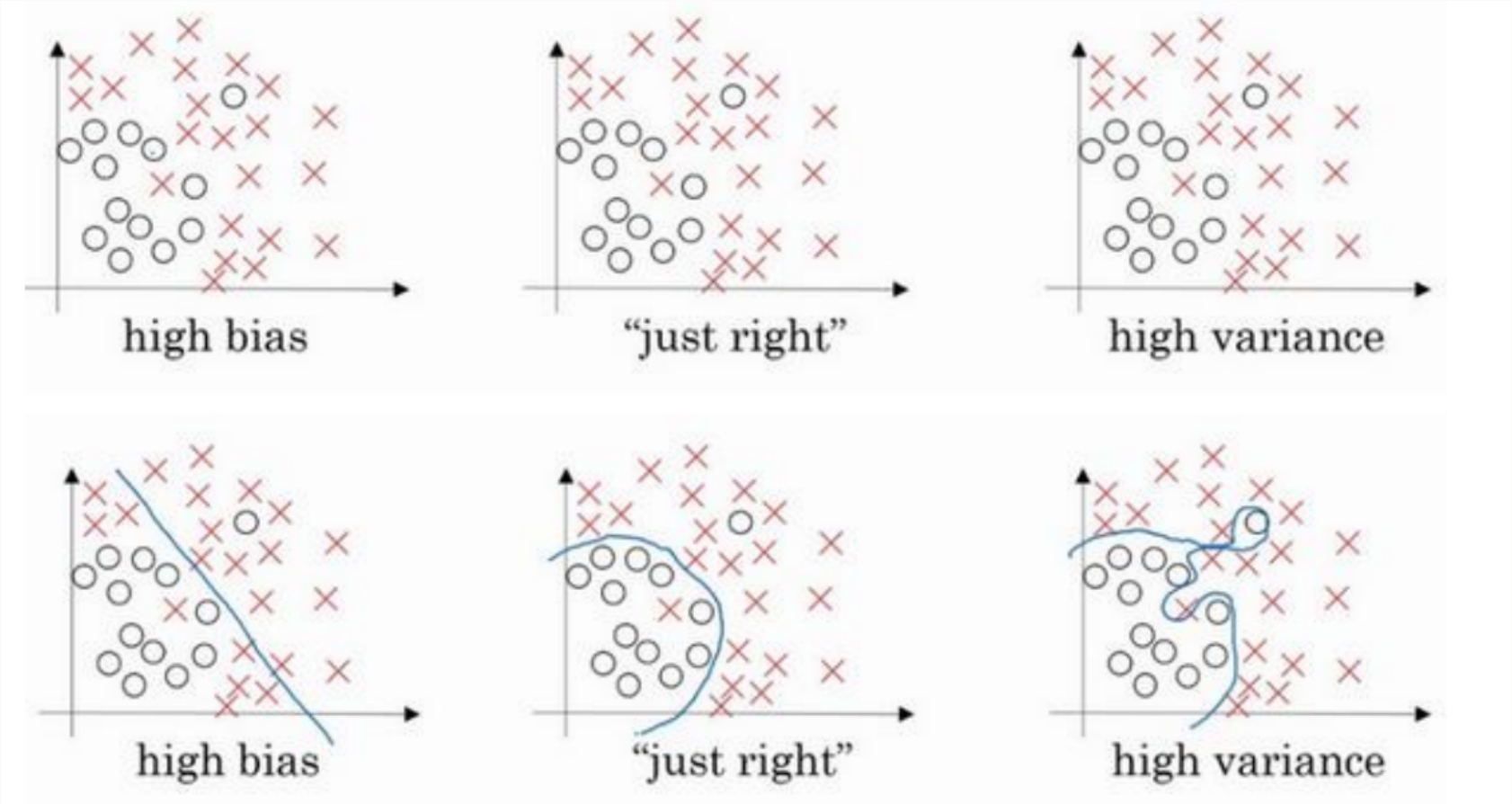


图 3: 左图是 High Bias，右图是 High Variance，中间是 Just Right

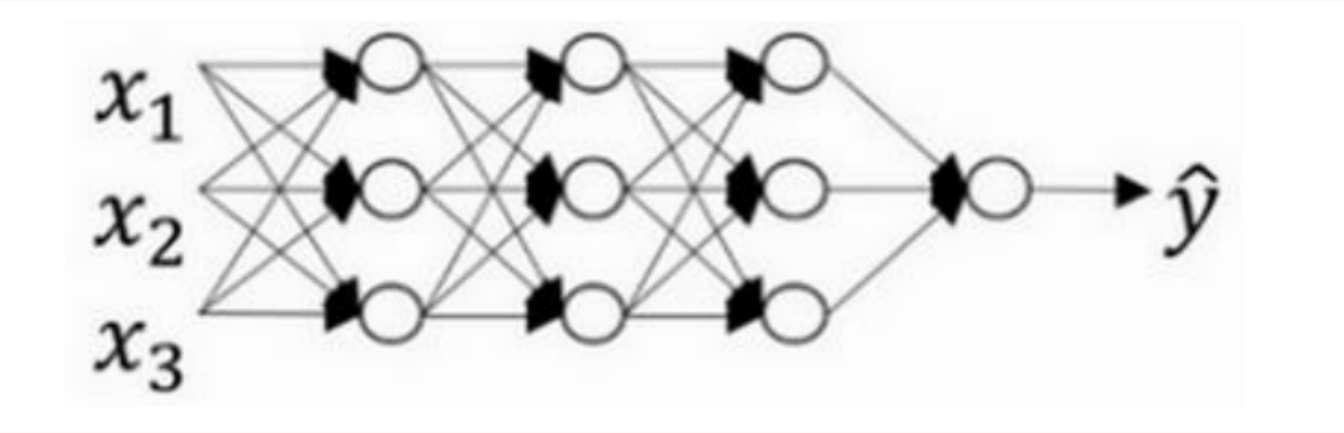


图 4: 简单的神经网络

6.2 激活函数角度

用 $tanh(z)$ 表示激活函数 $g(z)$, 可以发现，

- 当 $|z|$ 非常小，如果 z 只涉及少量参数，此时双曲正切函数近似的表现为一种线性函数
- 只有当 $|z|$ 很大时候，双曲正切函数才会体现它的非线性。

当正则化参数 λ 很大，激活函数的参数会相对较小。

$$Z_i = W_i X_i + b \tag{63}$$

- 如果 W 很小，相对来说，z 也会很小。特别是，如果 z 的值最终在这个范围内，都是相对较小的值，$g(z)$ 大致呈线性，每层几乎都是线性的，和线性回归函数一样。当每层都是线性的，那么整个网络就是一个线性网络，即使是一个非常深的深层网络，因具有线性激活函数的特征，最终我

图 5: tan 函数

们只能计算线性函数，因此，它不适用于非常复杂的决策，以及过度拟合数据集的非线性决策边界。这样就很容易出现欠拟合的情况。

- 如果 W 很大，相对来说，z 也会很大。这就会导致损失函数的值很大。这样就容易出现 High Variance。

7 广义拉格朗日函数

求解如下约束最优问题：

$$\min_{x} \ f(x) \ \ s.t. \ g(x) \leq 0 \tag{64}$$

引入广义拉格朗日函数：

$$L(x, \lambda) = f(x) + \lambda g(x) \tag{65}$$

需要证明：

$$\min_{x} \ f(x) \ \ s.t. \ g(x) \leq 0 \Leftrightarrow \min_{x} \ \max_{\lambda:\lambda\geq 0} \ L(x,\lambda) \tag{66}$$

1. 记 $P(x)=\max_{\lambda:\lambda\geq 0}$
   - 如果 g(x)>0, 由于 λ 可以取任意大，因此这时函数 P(x) 不可能取得最小值。
   - 如果 $g(x) \leq 0, \max_{\lambda:\lambda\geq 0} L(x,\lambda) = f(x)$，因此与式(66)等价。
2. 证明：

$$\max_{\lambda:\lambda\geq 0} \ \min_{x} \ L(x,\lambda) \ \leq \min_{x} \ \max_{\lambda:\lambda\geq 0} \ L(x,\lambda) \tag{67}$$

由于原始问题与对偶问题均由最优值，所以可以分别假设:

$$\begin{aligned} \max_{\lambda:\lambda\geq 0} \ \min_{x} \ L(x,\lambda) &= L(x_0,\lambda_0) \\ \min_{x} \ \max_{\lambda:\lambda\geq 0} \ L(x,\lambda) &= L(x_1,\lambda_1) \end{aligned} \tag{68}$$

那么，对于任意 x 都有：

$$L(x_0,\lambda_0) \leq L(x,\lambda_0) \tag{69}$$

对于任意的 λ 都有：

$$L(x_1,\lambda_1) \geq L(x_1,\lambda) \tag{70}$$

可得：

$$L(x_1,\lambda_1) \geq L(x_1,\lambda_0) \geq L(x_0,\lambda_0) \tag{71}$$

7.1 带约束的目标函数

7.1.1 无约束的优化问题

$$\min \ f(x) \tag{72}$$

其中，$x = (x_1, x_2)$。此时 $f(x)$ 在局部极小值点 $x^* = (x_1^*, x_2^*)$ 处的梯度必然为 0. 这个梯度为零的条件是局部极小值点的必要条件。这样，优化问题的求解变成了对该必要条件解方程组。

7.1.2 带等式约束的优化问题

$$\min_x \ f(x) \ \ s.t. h(x) = 0 \tag{73}$$

与无约束的问题不同。我们所要求的极小值点被限制在曲线 $h(x) = 0$ 上，我们将 $x|h(x) = 0$ 称为可行域, 解只能在这个可行域里取。如下图所示，曲线 $h(x) = 0$(黑色实曲线) 经过无约束极小值点（黑点）附近。那么满足约束的极小值点应该与黑点尽可能近。我们将 $f(x)$ 的等高线不断放大，直到与曲线 $h(x) = 0$ 相切，切点即为所求。相切是关键，是极小值点的必要条件。把 $h(x) = 0$ 沿着曲线方向参数化为 $x(t)$, $x^* = x(t^*)$ 。必有 $f(x)$ 在红点 x^* 的梯度方向与 $x(t)$ 的切线方向垂直，即

$$\nabla f(x^*) \cdot \dot{x}(t^*) = 0 \tag{74}$$

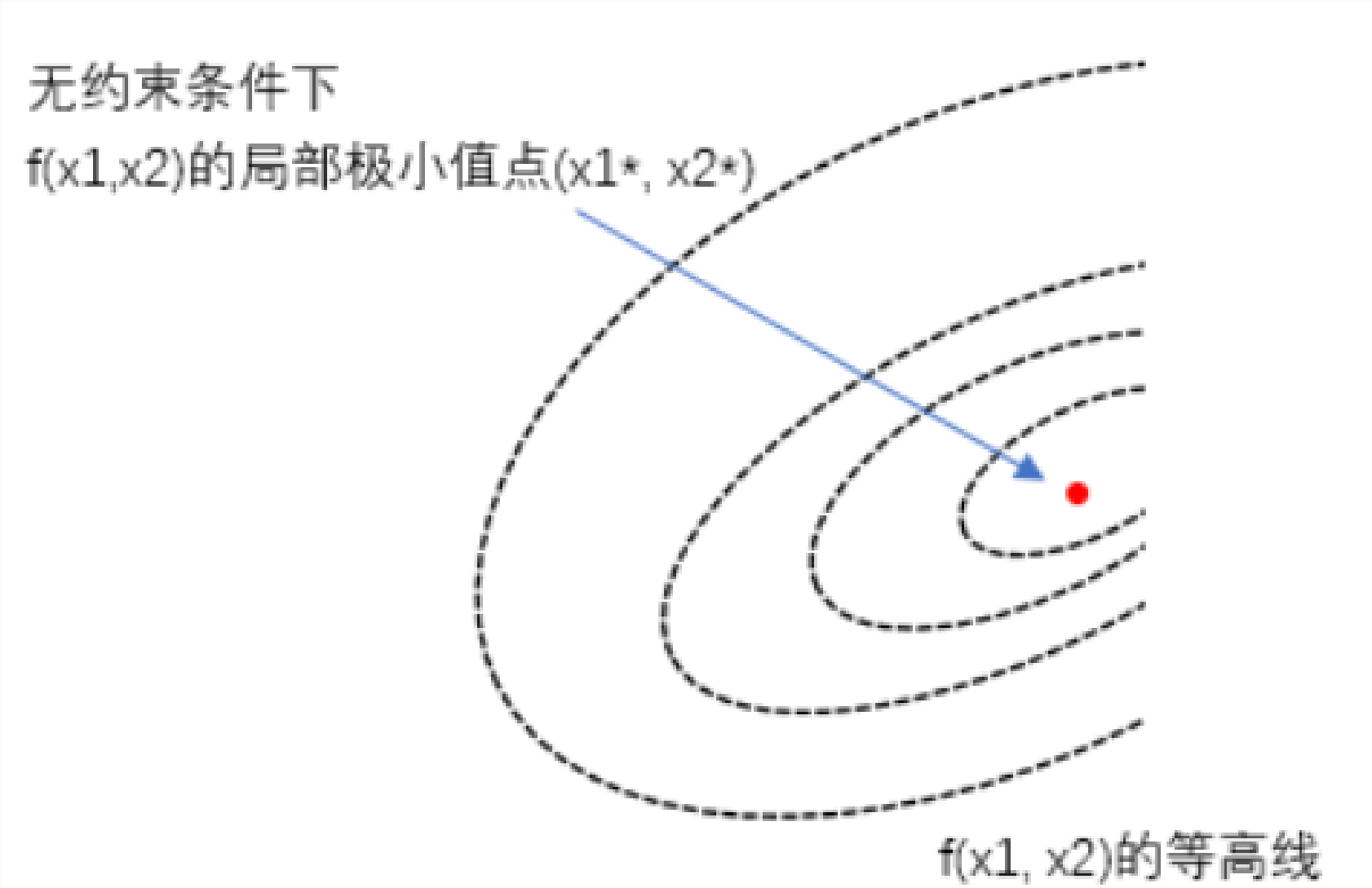


图 6: 无约束目标函数的最优值

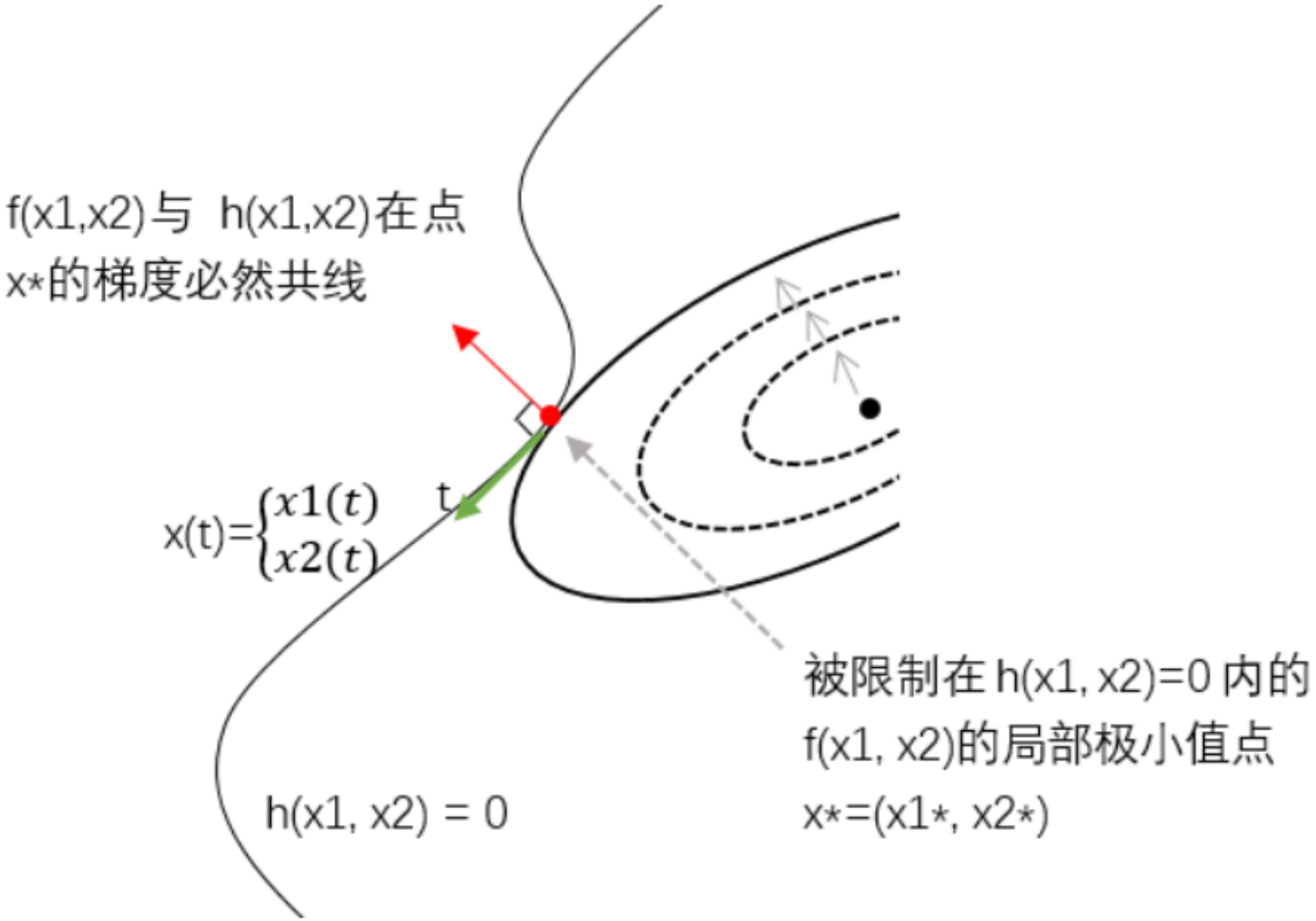


图 7: 带等式约束的优化问题

另外，由于 $h(x)=0$ 为常数，那么也有复合函数 $h(x(t))=0$, 因此 $h(x(t))$ 在 t 的导数必为 0，根据链式法则有

$$\nabla h(x)\cdot \dot{x}(t)=0 \quad \left(\text{ 内积为 } 0\text{ ，说明}\nabla h\left(x^{*}\right) \text{ 与}\dot{x}\left(t^{*}\right) \text{ 垂直}\right) \tag{75}$$

因为$\nabla f\left(x^{*}\right)$ 垂直于$\dot{x}\left(t^{*}\right)$，$\nabla h\left(x^{*}\right)$ 垂直于$\dot{x}\left(t^{*}\right)$，所以$\nabla f\left(x^{*}\right)$ 与$\nabla h\left(x^{*}\right)$ 共线, 有$\nabla f\left(x^{*}\right)+\lambda \nabla h\left(x^{*}\right)=0$

x^{*} 若为最小值点就必须满足上式和问题中的约束 $h(x^{*})$, 这个必要条件就叫作拉格朗日条件，为了好记，定义一个拉格朗日函数

$$L(x,\lambda)=f(x)+\lambda h(x) \tag{76}$$

令其偏导为 0，正好就得到拉格朗日条件。如此，带等式约束的优化问题转化为了无约束的优化问题，只需要对拉格朗日条件解方程组即可。这里 λ 就是拉格朗日乘子，有多少个等式约束就有多少个拉格朗日乘子。

7.1.3 带不等式约束的优化问题

$$\min_{x} \ f\left(x\right) \quad s.t.\ h\left(x\right)\leq 0 \tag{77}$$

当只有一个不等式起作用时, 如我们把问题 2 里的等式约束 $h\left(x\right)=0$ 改为 $h\left(x\right)\leq 0$ ，如下图所示，可行域变成了阴影部分，最小值点还是切点，情况和问题 2 完全一样，只需要把不等号当做等号去求解即可。

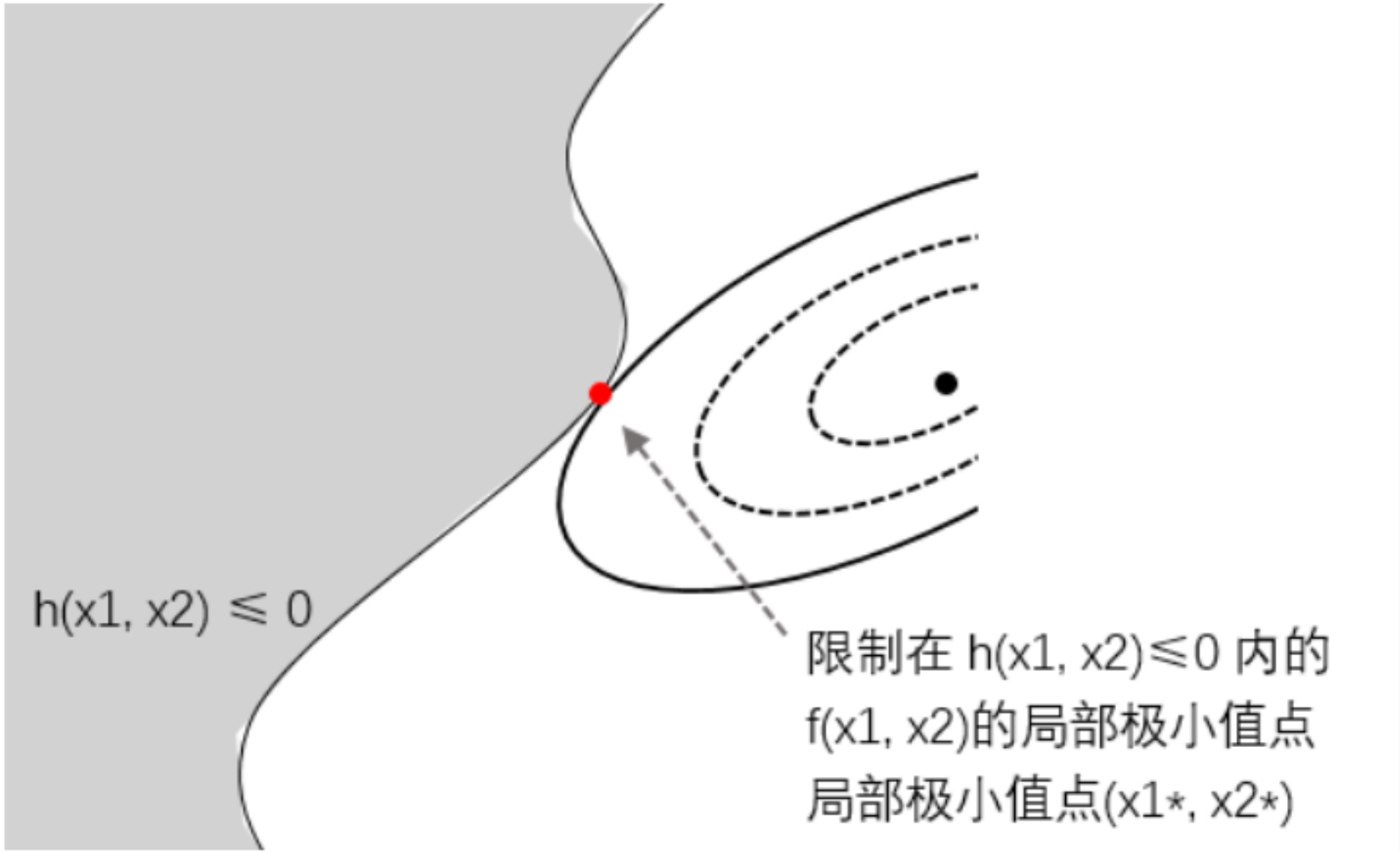


图 8: 带一个不等式约束的优化问题

当两个不等式起作用时，那么问题就来了

$$\min_{x} \ f(x) \quad s.t. \ g_1(x) \leq 0 \ g_1(x) \leq 0 \tag{78}$$

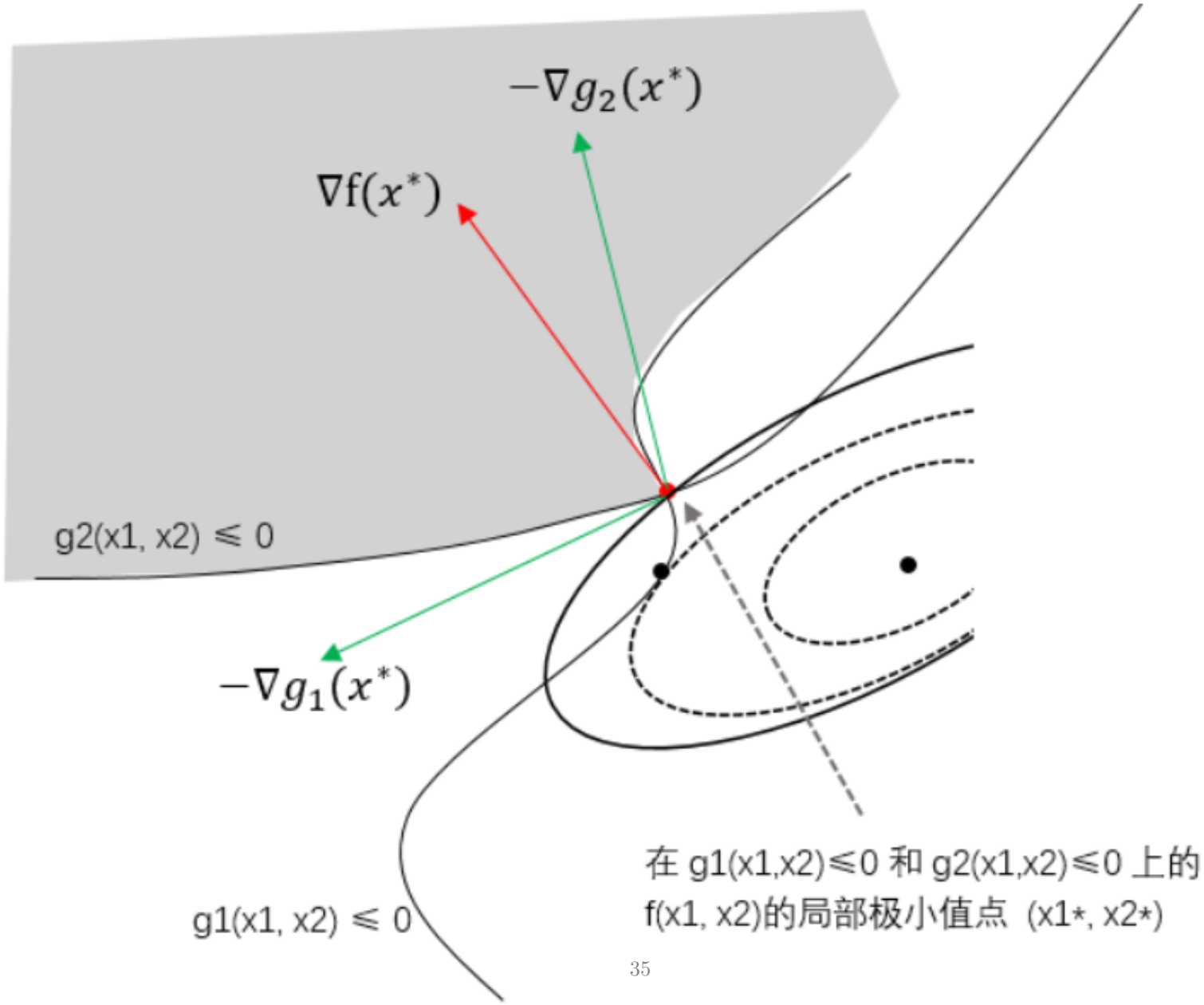
如图9，当 $f(x)$ 的等高线慢慢扩大时，等高线与可行域 (阴影部分) 第一次相遇的点是个顶点，2 个不等式同时起作用了。满足约束的最小值点从原来的黑点位置 (切点) 移动到了红点位置，现在跟哪条约束函数曲线都不相切。这时候就需要用到 kkt 条件了。这里的“条件”是指：某一个点它如果是最小值点的话，就必须满足这个条件（在含不等式约束的优化问题里）。这是个必要条件，前面说的也全部是必要条件。

这个问题的解 x^* 应满足的 KKT（卡罗需-库恩-塔克）条件为：

$$\begin{cases} \mu_1 \geq 0, \quad \mu_2 \geq 0 \\ \nabla f(x^*) + \mu_1 \nabla g_1(x^*) + \mu_2 \nabla g_2(x^*) = 0 \\ \mu_1 g_1(x^*) + \mu_2 g_2(x^*) = 0 \end{cases} \tag{79}$$

其中，μ 叫 KKT 乘子，有多少个不等式约束就有多少个 KKT 乘子。加上问题 3 中的约束部分，就是完整版的 KKT 条件。对于有等式的情况，你把其中一个不等式约束换成等式，可行域变成了半条曲线，最小值点还是那个红点，和下面这种情况是一样的。

下面看看 KKT 条件是怎么来的。在问题 2 中我们知道了约束曲线的梯度方向与曲线垂直, 我在上图画出了两条约束曲线的负梯度方向 (绿色箭头) 和等高线的梯度方向 (红色箭头) 。如果这个顶点是满足约束的最小值点, 那么该点处 (红点)，红色箭头一定在两个绿色箭头之间 ($-\nabla g(x)$ 方向


$-\nabla g_2(x^*)$
$\nabla f(x^*)$
g2(x1, x2) ≤ 0
$-\nabla g_1(x^*)$
g1(x1, x2) ≤ 0
在 g1(x1,x2)≤0 和 g2(x1,x2)≤0 上的
f(x1, x2)的局部极小值点 (x1*, x2*)

一定指向 g(x) 减小的方向, 即 g(x)<0 的那一边)。即 $\nabla f\left(x^{*}\right)$ 能被 $-\nabla g_{1}\left(x^{*}\right)$ 和 $-\nabla g_{2}\left(x^{*}\right)$ 线性表出 $\left(\nabla f\left(x^{*}\right)=-\mu_{1} \nabla g_{1}\left(x^{*}\right)-\mu_{2} \nabla g_{2}\left(x^{*}\right)\right)$, 且系数必非负 $\left(\mu_{1} \geq 0, \quad \mu_{2} \geq 0\right)$ 。也就是 kkt 条件中的 1 和 2:

$$\begin{cases} \mu_{1} \geq 0, \quad \mu_{2} \geq 0 \\ \nabla f\left(x^{*}\right)+\mu_{1} \nabla g_{1}\left(x^{*}\right)+\mu_{2} \nabla g_{2}\left(x^{*}\right)=0 \end{cases} \tag{80}$$

有时候，有的不等式约束实际上不起作用，如下面这个优化问题

$$\min f(x) \quad \text{s.t.} \quad g_{1}(x) \leq 0, \; g_{2}(x) \leq 0, \; g_{3}(x) \leq 0 \tag{81}$$

如下图的 $g_{3}\left(x_{1}, x_{2}\right) \leq 0$ 是不起作用的

对于最小值点 x^{*} , 三个不等式约束的不同在于

$$\begin{aligned} & g_{1}\left(x^{*}\right)=0 \quad (\text{ 起作用}) \\ & g_{2}\left(x^{*}\right)=0 \quad (\text{ 起作用}) \\ & g_{3}\left(x^{*}\right)<0 \quad (\text{不起作用, 最小值点不在} g_{3}(x)=0 \text{ 上}) \end{aligned} \tag{82}$$

这时，这个问题的 KKT 条件 1，2 成了：

$$\mu_{1} \geq 0, \quad \mu_{2} \geq 0, \quad \mu_{3} \geq 0 \nabla f\left(x^{*}\right)+\mu_{1} \nabla g_{1}\left(x^{*}\right)+\mu_{2} \nabla g_{2}\left(x^{*}\right)+\mu_{3} \nabla g_{3}\left(x^{*}\right)=0 \tag{83}$$

条件 2 中的 $\mu_{3} \nabla g_{3}\left(x^{*}\right)$ 这一项让我们很苦校啊, $g_{3}\left(x^{*}\right)$ 的绿色箭头跟我们的红色箭头没关系。要是能令 $\mu_{3}=0$ 就好了。加上条件 3:

$$\mu_{1} g_{1}\left(x^{*}\right)+\mu_{2} g_{2}\left(x^{*}\right)+\mu_{3} g_{3}\left(x^{*}\right)=0 \tag{84}$$

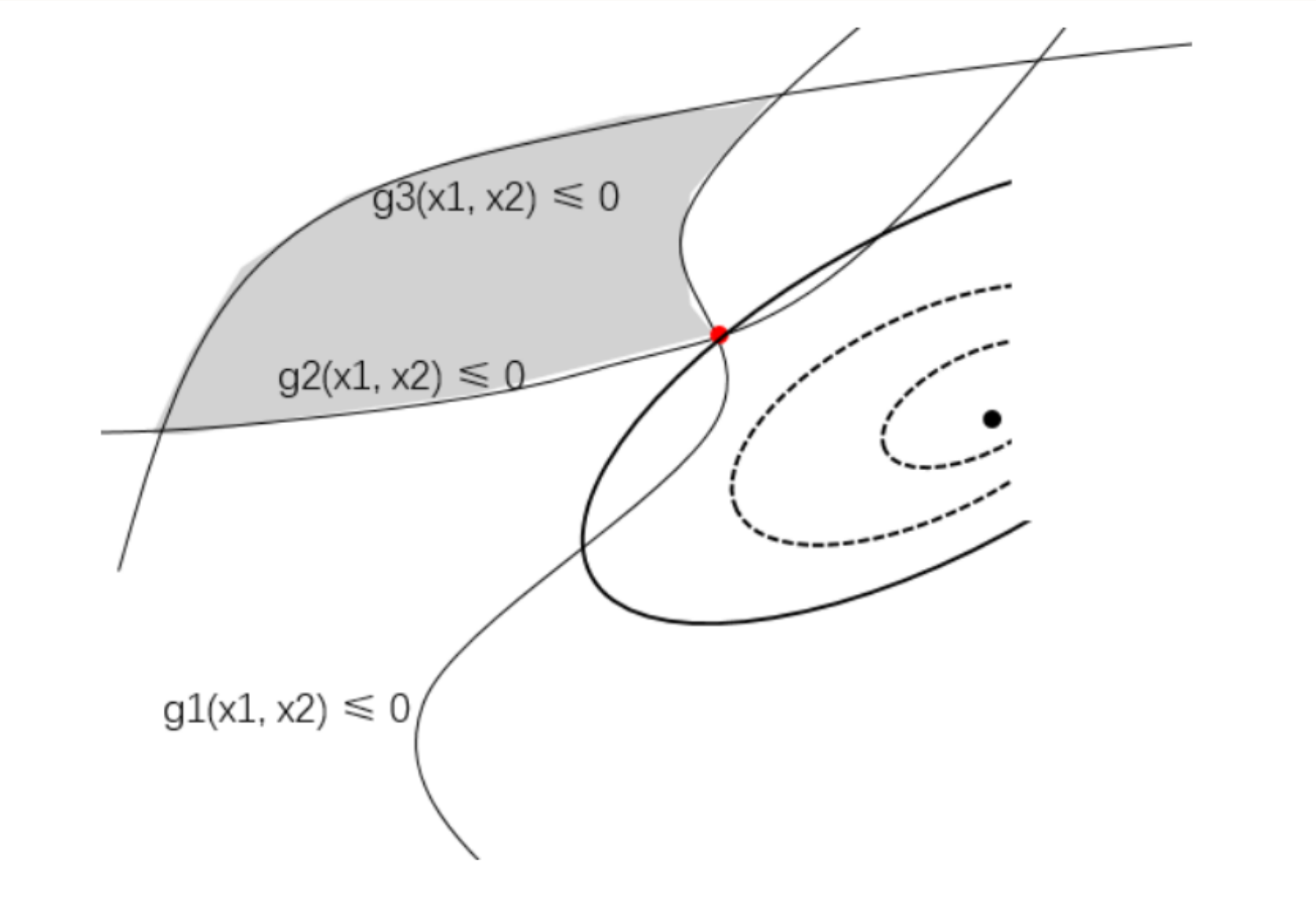


图 10: 带三个个不等式约束的优化问题

恰好能使 $\mu_3 = 0$ 。由于 $g_1(x^*) = 0, \quad g_2(x^*) = 0$, 所以前两项等于 0，第三项 $g_3(x^*) < 0$, 在条件 3 的作用下使得 $\mu_3 = 0$. 正好满足需求。如果再多几项不起作用的不等式约束, 比如 $g_4(x) \leq 0$。要使

$$\mu_1 g_1(x^*) + \mu_2 g_2(x^*) + \mu_3 g_3(x^*) + \mu_4 g_4(x^*) = 0 \tag{85}$$

就只能有 $\mu_3 g_3(x^*) + \mu_4 g_4(x^*) = 0$

同样地, $g_3(x^*) < 0, g_4(x^*) < 0$, 只能出现 $\mu_3 = \mu_4 = 0$ 或者 $\mu_3 \mu_4$ 异号的情况。但注意条件 1 限制了 $\mu_3 \geq 0, \quad \mu_4 \geq 0$, 所以只能有 $\mu_3 = \mu_4 = 0$ 。因此不管加了几个不起作用的不等式约束, 条件 2 都能完美实现：目标函数 $f(x)$ 的梯度 $\nabla f(x)$ 被起作用的不等式约束函数 $g(x)$ 的负梯度 $(-\nabla g(x))$ 线性表出且系数 μ 全部非负 (红色箭头被绿色箭头夹在中间) 。这样, 优化问题的求解就变成对所有 KKT 条件解方程组。

如果再定义一个拉格朗日函数

$$L(x, \mu) = f(x) + \mu_1 g_1(x) + \mu_2 g_2(x) + \ldots \tag{86}$$

令它对 x 的偏导为 0，就是 KKT 条件中的条件 2 了。最后说明一下，以上所有都是局部极小值点的必要条件。据此求得的解不一定是局部极小值点（更别提全局了），原因是上图中我所画的等高线也许根本就不闭合，也就是说我们一直想要靠近的等高线中间的黑点压根就是个鞍点或者近似鞍点.

7.2 KKT 条件

对于具有等式和不等式约束的一般优化问题:

$$\begin{aligned} &\min \ f(x) \\ s.t.\ &g_i(x) \le 0 \ \ (j = 1, 2, \cdots, m) \\ &h_k(x) = 0 \ \ (k = 1, 2, \cdots, l) \end{aligned} \tag{87}$$

KKT 条件给出了判断 x^* 是否为最优解的必要条件，即：

$$\begin{cases} \frac{\partial f}{\partial x_i} + \sum_{j=1}^{m} \mu_j \frac{\partial g_j}{\partial x_i} + \sum_{k=1}^{l} \lambda_k \frac{\partial h_k}{\partial x_i} = 0, (i = 1, 2, \ldots, n) \\ h_k(\mathbf{x}) = 0, (k = 1, 2, \cdots, l) \\ \mu_j g_j(\mathbf{x}) = 0, (j = 1, 2, \cdots, m) \\ \mu_j \ge 0 \end{cases} \tag{88}$$

8 噪声鲁棒性

对于具有输入向量 x 和单输出 y 的网络函数 y(x)，可以写入形式的平方和误差：

$$E = \left\langle \{y(\mathbf{x}) - t\}^2 \right\rangle \mathbf{x}, t \tag{89}$$

t 是目标变量，$\langle\cdot\rangle$ 为期望。现在假设每次向网络呈现输入向量 x 时，都会添加随机扰动 ϵ。然后，误差函数由数据分布和形式噪声分布的平均偏差给出:

$$\tilde{E} = \left\langle \{y(\mathbf{x}+\boldsymbol{\epsilon}) - t\}^2 \right\rangle_{\mathbf{x},t,\boldsymbol{\epsilon}} \tag{90}$$

对 $y(\mathbf{x}+\boldsymbol{\epsilon})$ 进行 Taylor 扩展:

$$y(\mathbf{x}+\boldsymbol{\epsilon}) = y(\mathbf{x}) + \boldsymbol{\epsilon}^{\mathrm{T}}\nabla y(\mathbf{x}) + \frac{1}{2}\boldsymbol{\epsilon}^{\mathrm{T}}\nabla\nabla y(\mathbf{x})\boldsymbol{\epsilon} + O\left(\boldsymbol{\epsilon}^3\right) \tag{91}$$

假定噪声分布是零均值，并且协方差矩阵与单位矩阵 I 成比例，比例为:ν，可得：

$$\begin{aligned} \langle\boldsymbol{\epsilon}\rangle &= 0 \\ \left\langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right\rangle &= \nu\mathbf{I} \end{aligned} \tag{92}$$

将式(91)代入式(90)中可得：

$$\tilde{E} = E + \nu\left\langle\|\nabla y\|^2\right\rangle_{\mathbf{x}} + \nu\left\langle(y-t)\nabla^2 y\right\rangle_{\mathbf{x},t} \tag{93}$$

通过比较式(90)和式(93)，定义：

$$y(\mathbf{x}) = \langle t \mid \mathbf{x}\rangle_t + O(\nu) \tag{94}$$

其中 $\langle t \mid \mathbf{x}\rangle$ 为目标数据的条件偏差，也称为可变目标回归。

将式(94)重置式(93)，对于输入向量 x^n 和匹配目标值 t^n 的数据集（其中 $n = 1,\ldots,N$），然后可以得到实际误差函数：

$$\tilde{E} = \sum_{n=1}^{N}\left\{y\left(\mathbf{x}^n\right) - t^n\right\}^2 + \nu\sum_{n=1}^{N}\left\|\nabla y\left(\mathbf{x}^n\right)\right\|^2 \tag{95}$$