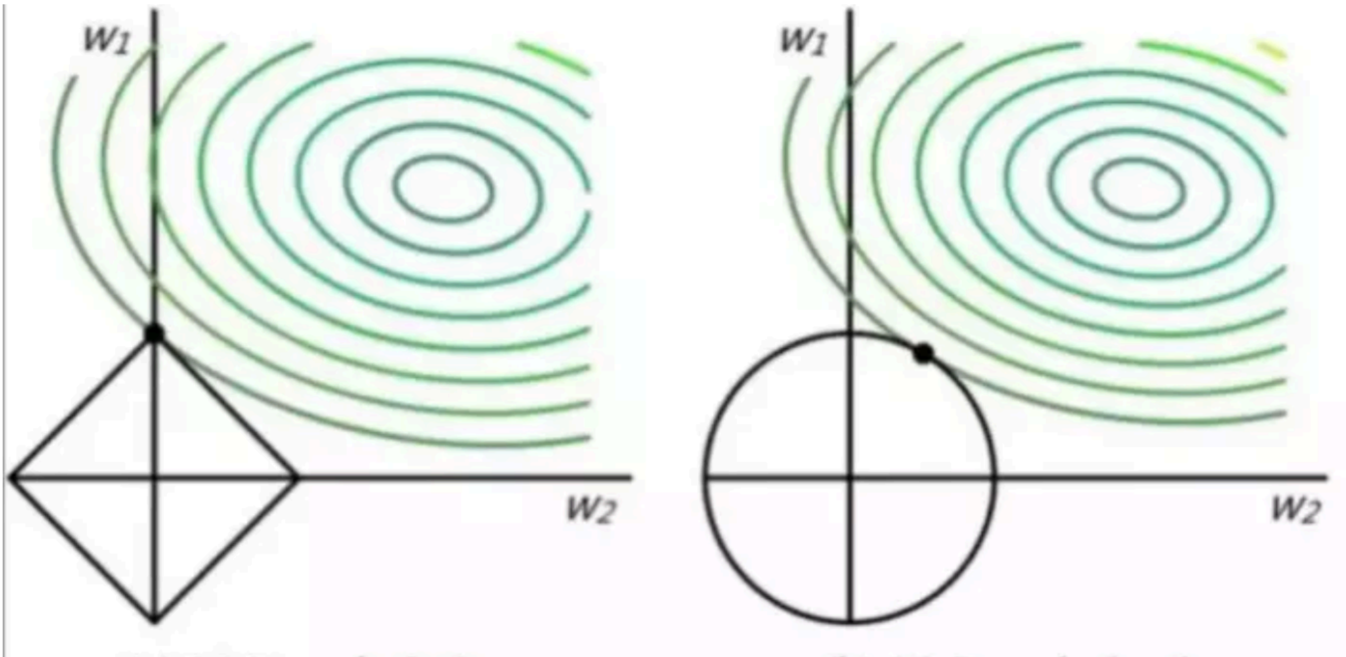
1. 正则化如何防止过拟合?

防止过拟合原理：模型越复杂，正则化项的值越大。要使正则化项也很小，那么模型复杂程度受到限制，因此就能有效地防止过拟合。

2. L1正则与L2正则



(a) L1 Regulariztion (b) L2 Regularization

L1正则化

公式：$\Omega(w)=||w||$，则，$\widetilde{J}(w)=J(w)+\lambda||w||$

求导：

- $\frac{\partial \widetilde{J}}{\partial w}=\frac{\partial J(w)}{\partial w}+\lambda sign(w)$
- 梯度下降：$w=w-\eta[\frac{\partial J(w)}{\partial w}+\lambda sign(w)]$ 注：η是学习率

特点：

- 能使得参数稀疏，具有特征选择的功能。
    - 原因：结果点在参数坐标轴上，所以w中会产生很多0值，使得参数矩阵稀疏
- 模型不是处处可微。
    - 原因：图形拐点处不可微

L2正则化

公式：$\Omega(w)=||w||^2$, 则，$\widetilde{J}(w)=J(w)+\frac{\lambda}{2}||w||^2$

求导：

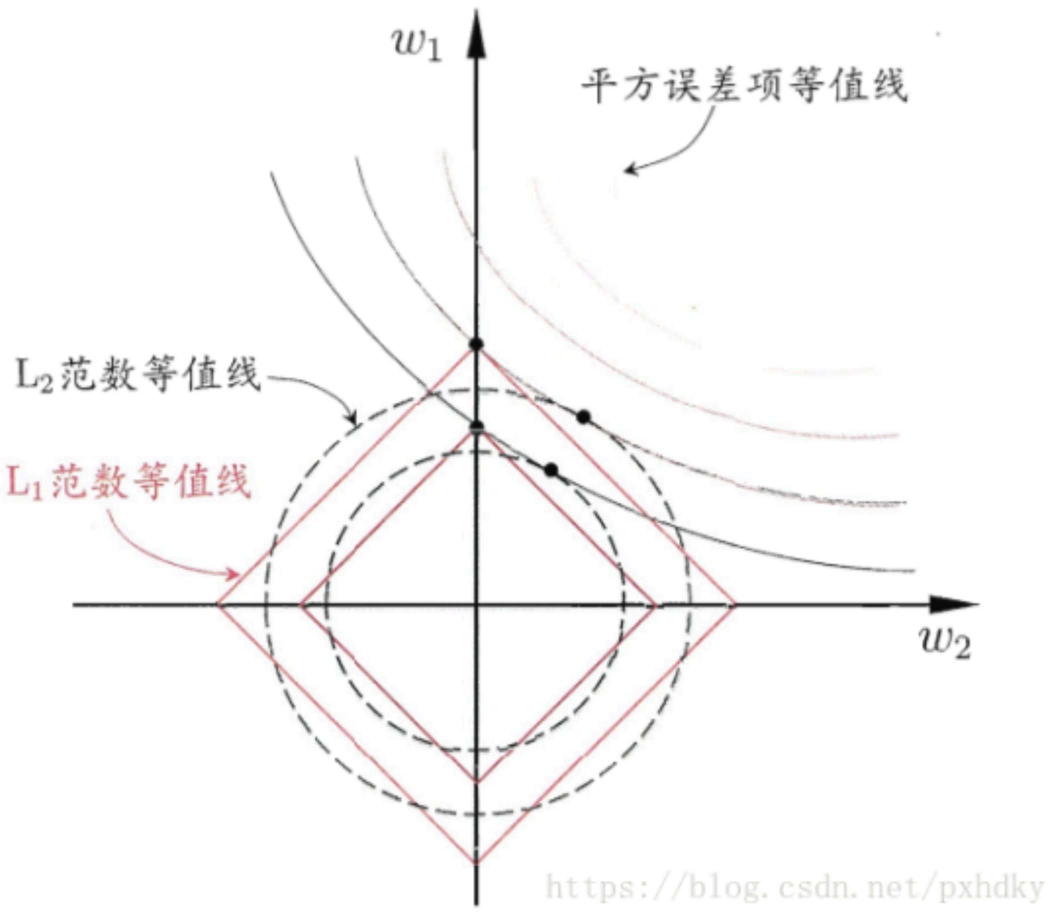
- $\frac{\partial\widetilde{J}}{\partial w}=\frac{\partial J(w)}{\partial w}+\lambda w$
- 反向传播：$w=w-\eta[\frac{\partial J(w)}{\partial w}+\lambda w]=(1-\eta\lambda)w-\eta\frac{\partial J(w)}{\partial w}$

特点：

- 能迅速使得参数变小，但不稀疏。
    - 原因：参数在每次更新的时候，都会先乘一个小于1的数$(1-\eta\lambda)$，从而使得w迅速的变小
    - 原因：由上图，在坐标轴上相交的概率大大降低了，从而避免了稀疏矩阵。
- 模型处处可微。
    - 原因：由图可知

3. 为什么L1正则化更容易获得稀疏矩阵?

假设仅有两个属性，w只有两个参数w_1,w_2，绘制不带正则项的目标函数-平方误差项等值线，再绘制$L1,L2$范数等值线，如图正则化后优化目标的解要在平方误差项和正则化项之间折中，即出现在图中等值线相交处采用。$L1$范数时，交点常出现在坐标轴上，即w_1或$w2$为0;而采用$L2$范数时，交点常出现在某个象限中，即w_1,w_2均非0。也就是说，范数比范数更易获得“稀疏”解。



4. L1正则和L2正则的先验分别服从什么分布?

最大似然估计 MLE $= argmax \sum log(P(x_i|\theta))$

最大后验概率 MAP $= argmax \sum log(P(x_i|\theta)) + log(P(\theta))$

所以： $MLE + log(P(\theta)) = MAP$

L1正则

拉普拉斯分布： $f(x|\mu,b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$

最大似然估计 + L1正则 = 最大后验概率（先验分布为**拉普拉斯**）

L2正则

高斯分布：$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

最大似然估计 + L2正则 = 最大后验概率（先验分布为**高斯分布**）