规则化

范数

矩阵求导

Matlab 编程

一. 监督学习问题，就是“minimize your error while regularizing your parameters”,即在规则化参数的同时最小化误差。最小化误差是为了让我们的模型拟合我们的训练数据，而规则化参数是防止我们的模型过渡拟合我们的训练数据。训练误差最小并不是我们的最终目标，我们的目标是希望模型的测试误差最小。规则化的使用可以根据不同的先验知识，引入不同的范数和约束，强行让学习到的模型具有人们想要的特性，比如稀疏，低秩，平滑。

如果用奥卡姆剃刀原理来解释，即在所有可能的模型中，我们应该选择能够很好解释已知数据并且十分简单的模型。根据通俗的说法，即规则化是结构风险最小化策略的实现，是在经验风险上加一个正则化（regularizer）或惩罚项（penalty term）。

一般而言，监督学习可以看做最小化下面的目标函数：

$$w^* = argmin_w \sum_i L(y_i, f(x_i; w)) + \lambda \Omega(w)$$

机器学习的大部分带参数的模型都和这个不但神似，而且形似。大部分无非就是变换这两个项而已。对于第一项 Loss 函数，如果是 Square loss，那就是最小二乘；如果是 Hinge loss，那么就是著名的 SVM；如果是 exp-loss，那么就是牛逼的 Boosting；如果是 log-loss,那就是 logistic regression 了。一些 **loss function** 的例子：

Ex1. “gold standard”,”0-1” loss: L_{01}

$$L_{01}(m) = \begin{cases} 0 & if\ m \geq 0 \\ 1 & if\ m < 0 \end{cases}$$

Ex2. “hinge loss”: L_{hinge} for soft margin SVM

$$\begin{aligned} J(w) &= \frac{1}{2}\|w\|^2 + \sum_i max(0, 1 - y^i w^t x^i) \\ &= \frac{1}{2}\|w\|^2 + \sum_i max(0, 1 - m_i(w)) \\ &= R_2(w) + \sum_i L_{hinge}(m_i) \end{aligned}$$

Ex3. “log loss”: equivalent to the cross entropy loss function used to train a logistic regression model:

$$J(w) = \lambda\|w\|^2 + \sum_i y^i \log g_w(x^i) + (1 - y^i)(\log(1 - g(x^i))),\ y^i \in (0,1)$$

$$g(x^i) = \frac{1}{1+e^{-f_w(x^i)}}$$

$$f_w(x^i) = w^T x^i$$

because：

$$1 - g(x^i) = 1 - \frac{1}{1+e^{-f_w(x^i)}}$$

$$= \frac{e^{-f_w(x^i)}}{1+e^{-f_w(x^i)}}$$

$$= \frac{1}{1+e^{f_w(x^i)}}$$

So:

$$J(w) = \lambda\|w\|^2 + \sum_i \log(1 + e^{-\hat{y(i)} f_w(x^i)})$$

$$L_{log}(m) = \log(1 + e^{-m})$$

$$m^i = \hat{y(i)} f_w(x^i)$$

$$\hat{y(i)} = \begin{cases} -1 & if y^i = 0 \\ 1 & if\ y^i = 1 \end{cases}$$

Ex4. “Linear Regression”: use L_2 describe the squared loss term

$$L_2 = (f_w(x) - y)^2 = (m-1)^2$$

Ex5. “exponential loss term” for boosting

$$J(w) = \lambda R(\mathrm{w}) + \sum_i \exp(-y^i f_w(x^i))$$

$$L_{exp}(m_i) = \exp(-m_i(w))$$

Figure1: Loss function.

In fig(1)we show the curves of these five functions, the blue line is for 0-1 loss function,

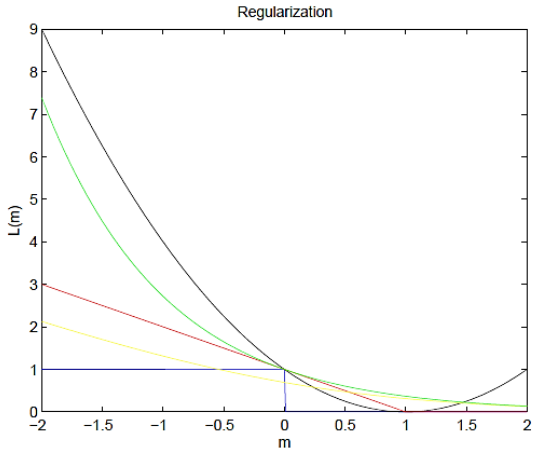


Figure 14.1. loss functions

the black line is for squared loss function,the red line is for Hinge loss, the yellow line is for logistic loss and the green line is for boosting loss. We can learn the following four points from analysis of these five loss terms.

1. L_{hinge}, L_{log}对训练样本中的噪声数据是不敏感的;
2. 所有损失都是L_{01}的凸代理;
3. 如果我们的目标不是去最小化新样本$(\mathrm{x}^{new}, y^{new})$的 zero-one L_{01} 损失，而是去最大化概率$\mathrm{P}(y^{new}|\mathrm{x}^{new})$，那么$L_{log}$将是一个正确的选择。但这意味着我们对结果的评判将伴随着无上限的损失。
4. 有限的边界和凸之间存在矛盾。从某种意义上说，loss 应该是有边界的，我们不希望对一个错误分类的样本赋予无限大的损失，但是有边界的又暗示了非凸，从而导致难以优化，比如，L_{01}。

二. 规则化函数$\Omega(w)$也有很多选择，一般他是模型复杂度的单调递增函数，模型越复杂，规则化值就越大。比如，规则化项可以是模型参数向量的范数。然而，不同的选择对参数 w 的约束不同，取得的效果也不同，通常有以下几类范数：零范数，一范数，二范数，迹范数，Frobenius 范数和核函数等。

一些规则化项：

$$\mathrm{R}_2 = \frac{1}{2}\|w\|^2$$

$$\mathrm{R}_1 = \sum_i |w_i|$$

$$\mathrm{R}_0 = |\{i: w_i \neq 0\}|$$

$$\mathrm{R}_p = (\sum_i |w_i|^p)^{\frac{1}{p}}$$

R_2是最常用的，容易用梯度下降算法来进行最优化，而且也可以根据损失相加法在梯度中引入λw"。 R_0是最容易理解的规则化项，即特征选择，这会引起向量/矩阵的稀疏，从而降低计算量。同时，R_2，R_1是凸的，而$\mathrm{R}_p(p<1)$，包括R_0是非凸的。因此，通常我们会用R_1来近似R_0来产生稀疏的结果。下面是一些简单的规则化效果：

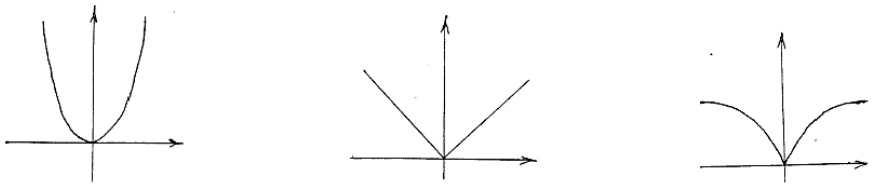


Figure 14.2. Plots of the function $R_p(w)$ corresponding to $p = \{2, 1, .3\}$ for a one-dimensional w.

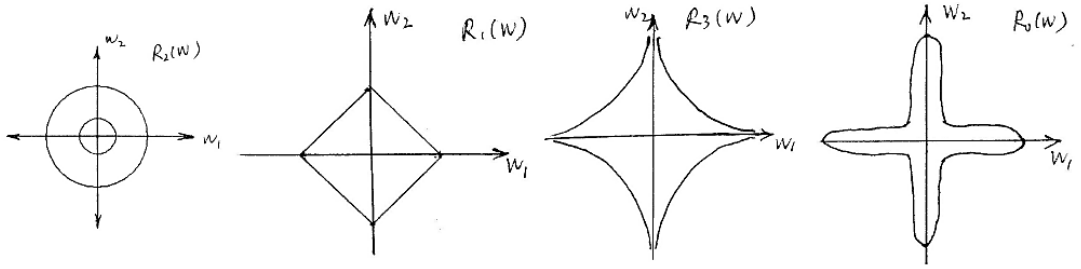


Figure 14.3. Iso-contour lines for the function $R_p(w)$ corresponding to $p = \{2, 1, .3, 0\}$ for a two-dimensional w.

下面，我们着重介绍 L0，L1，L2 范数。

1. L0 范数和 L1 范数

L0 范数是指向量中非 0 的元素的个数，如果用 L0 范数来规范化参数矩阵 W，就是希望 W 尽可能的稀疏。L1 范数是指向量中各个元素绝对值之和，也称为“稀疏规则算子”（Lasso regularization）。那么为什么通常说 L1 会使得权值矩阵稀疏？有一个说法是 L1 是 L0 范数的最优凸近似。其实，还有一个更美的解释：任何的规则化算子，如果他在$W_i = 0$的地方不可微，并且可以分解为一个求和的形式，那么这个规则化算子就可以实现稀疏。也就是说，W 的 L1 范数是绝对值|w|，且在 w=0 处是不可微的。

对了，这儿要强调下用 L1 代替 L0 进行稀疏的原因：L0 范数是难优化求解(NP 难问题)；L1 范数是 L0 范数的最优凸近似，而且更容易优化求解。

$$\text{Min } \|x\|_0 \quad \text{s.t. } Ax = b \quad \xLeftrightarrow{\text{在一定条件下，以概率 1 意义下等价}} \quad \text{Min } \|x\|_1 \quad \text{s.t. } Ax = b$$

OK，用一句话而言，就是 L1 范数和 L0 范数都可以实现稀疏，但 L1 因为比 L0 有更好的优化求解特性而被广泛采用。那么，参数稀疏有什么好处呢？

1) 特征选择(Feature Selection)

稀疏规则化的关键作用是他能实现特征的自动选择。一般而言，xi 的大部分元素（也就是特征）都和最终的输出 yi 没有关系或者不提供任何信息，在最小化目标函数的时候考虑 xi 这些额外的特征，虽然可以获得更小的训练误差，但在预测新的样本时，这些没用的信息反而会被考虑进去，从而干扰了对正确 yi 的预测。稀疏规则化算子的引入，可以进行自动选择，去掉那些没有信息的特征，或者说是把这些特征的权值重置为 0。

2) 可解释性(Interpretability)

稀疏的另一个优点是使得模型更容易解释。如果通过学习，原本 1000 个维度的 W 仅有 5 个非零的 wi，那么可以认为这些对应的特征在分析上面提供的信息是巨大的，决策性的。

2. L2 范数

L2 范数，是一个更加常用的规则化范数，他有两个美称：一是岭回归(Ridge Regression)，二是权值衰减(Weight decay)。他的强大功效就是改善机器学习的一个重要问题：过拟合。（过拟合：通俗地讲，就是应试能力很强，实际应用能力很差）。例如下面的图中所示（来自 Ng 的 machine learning course）：

从左到右分别为欠拟合（underfiiting, High-bias 偏倚），合适的拟合和过拟合（overfitting, high variance 偏差）。

关于 Bias 和 Variance 的介绍：

1. 关于误差：如果将数据分为：Train set, cross validation set, testing set 三类，那么有误差

   Train error:

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2$$

   Cross validation error:

$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x_{cv}{}^i) - y_{cv}{}^i)^2$$

   Test error:

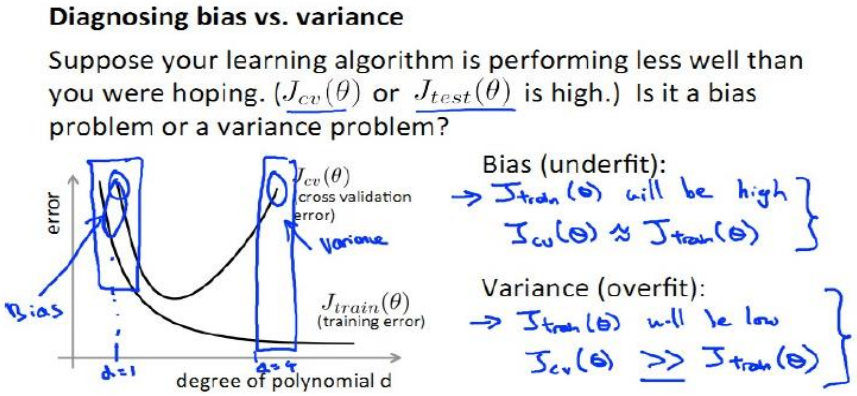
$$J_{test}(\theta) = \frac{1}{2m_{test}}\sum_{i=1}^{m_{test}}(h_\theta(x_{test}{}^i) - y_{test}{}^i)^2$$

进而，我们可以进行模型选择：首先，建立 d 个 model 假设，分别在 training set 上求得使其 training error 最小的θ向量，得到 d 个θ。然后对这 d 个 model 假设，在 cross validation set 上计算J_{cv}，取 cv set error 最小的一个 model 作为 hypothsis。

2. 由以上的J_{cv}和J_{train}，就产生了 bias 和 variance 的概念：

   Bias: J(train)大，J(cv)大，J(train)≈J(cv)，bias 产生于复杂度(degree of polynomial d)小，underfit 阶段

   Variance: J(train)小，J(cv)大，J(train)<<J(cv)，variance 产生于复杂度(d)大，overfit 阶段



下面给出 bias 和 variance 的由来和具体定义：

对于给定的数据集 D，对于这些数据集上的点我们可以计算每个 index 下的平均值（期望）$t(x) = E(y|x)$，则我们有 mean square error：

$$\text{MSE} = \frac{1}{n}\sum(f(x) - t(x)^2$$

Decomposing generalization error into bias and variance:

$$\begin{aligned}(f(x;D) - t(x))^2 &= [f(x;D) - E_D(f(x;D)) + E_D(f(x;D)) - t(x)]^2 \\ &= (f(x;D) - E_D(f(x;D))^2 + (E_D(f(x;D)) - t(x))^2 + 2(f(x;D) - E_D(f(x;D))(E_D(f(x;D)) - t(x))\end{aligned}$$

$$E_D\{(f(x;D) - t(x))^2\} = E_D\{(f(x;D) - E_D(f(x;D))^2\} + \{E_D(f(x;D)) - t(x)\}^2$$

极端理解：

1、记住训练集合上所有的点的 label，这样的系统低偏倚，高方差

2、无论输入是什么，总是预测一个相同的，这样的系统高偏倚，低方差

Variance

估计本身的方差

Bias

估计的期望和样本数据希望得到的回归函数之间的差别

一些参考的解释：

1.bias 表明一种老是学不到点子上的感觉，不准; variance 表明一种学习的结果容易飘，不稳

2.用仪器测量做类比，bias 好比系统误差，variance 好比随机误差

3. Bias-variance 分解是机器学习中一种重要的分析技术。给定学习目标和训练集规模，它可以把一种学习算法的期望误差分解为三个非负项的和，即本真噪音、bias 和 variance。

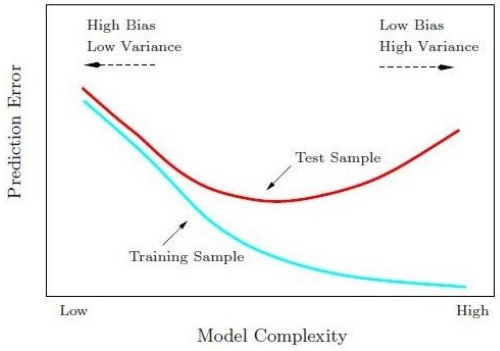
本真噪音是任何学习算法在该学习目标上的期望误差的下界；(任何方法都克服不了的误差)

bias 度量了**某种学习算法的平均估计结果所能逼近学习目标的程度**；（**独立于训练样本的误差**，刻画了匹配的准确性和质量：一个高的偏差意味着一个坏的匹配）

variance 则度量了**在面对同样规模的不同训练集时，学习算法的估计结果发生变动的程度**。（**相关于观测样本的误差**，刻画了一个学习算法的精确性和特定性：一个高的方差意味着一个弱的匹配）

Boosting 通过样本变量全部参与，故 Boosting 主要是降低 bias（同时也有降低 variance 的作用，但以降低 bias 为主）；而 Bagging 通过样本随机抽样部分参与（单个学习器训练），故 bagging 主要是降低 variance。

4.Everyone in machine learning knows about overfitting, but it comes in many forms that are not immediately obvious. One way to understand overfitting is by **decomposing generalization error into bias and variance. Bias is a learner's tendency to consistently learn the same wrong thing. Variance is the tendency to learn random things irrespective of the real signal.** Figure 1 illustrates this by an analogy with throwing darts at a board.



总结：Variance 是估计本身的方差，bias 是估计的期望和样本数据希望得到的回归函数之间的差别。

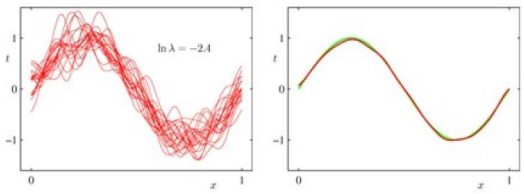
这样一来，对于正则化项中的λ:

a. λ小，d 大->overfit（flexible）->

对于不同的训练数据集的拟合结果抖动很大->variance 大；

Bias 是估计均值和实际期望的偏差->bias 小；

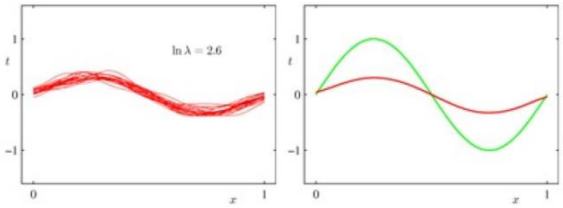
下图中，左图为拟合的 20 条曲线；右图红线为 20 条曲线的期望，绿色为实际数据期望所得的拟合曲线。



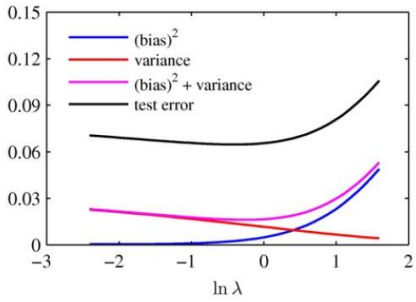
b. λ大，d 小->underfit（stable）->

对于不同的训练数据集的拟合结果抖动较小->variance 小；

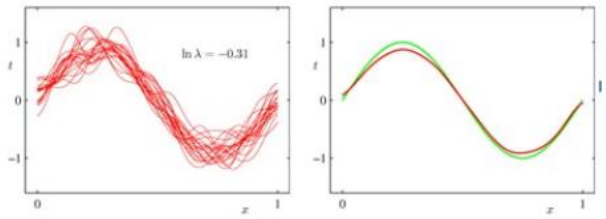
Bias 是估计均值和实际期望的偏差，不能很好得进行回归->bias 大；



c. 下图为λ，$bias, variance, error$之间的关系：



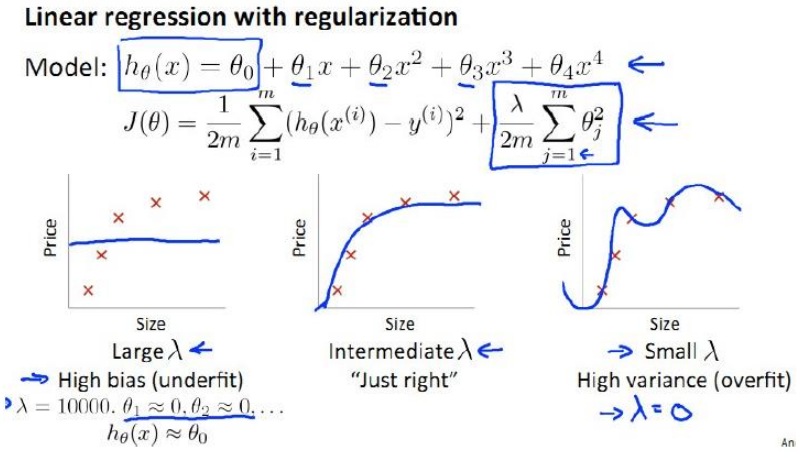
我们希望数据的 variance 和 bias 都不要大:



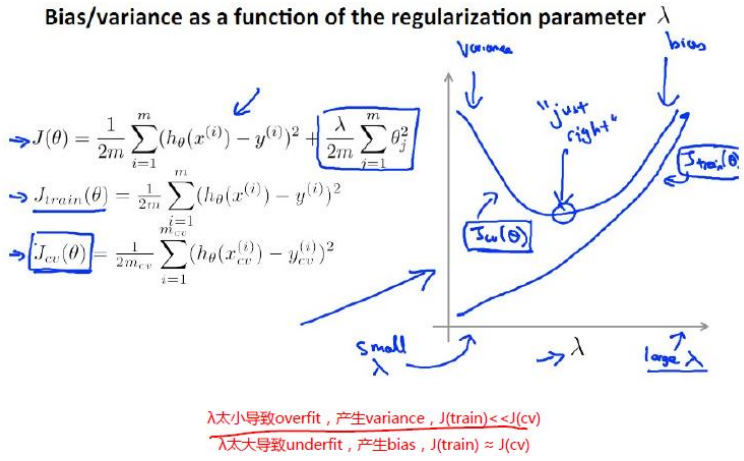
这时就是一个 variance 和 bias 之间的 tradeoff 问题了。

3. 规则化和 bias/variance

现在，我们将上一节的 bias 和 variance 应用于 regularization 中。这里的 regularization 就是为了防止 overfit 而在 cost function 中引入的一个分量，其中λ太大导致 underfit，λ太小导致 overfit。



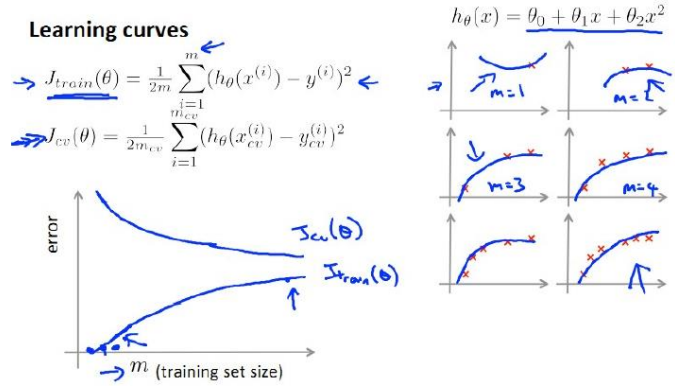
同样，模型的选择，首先选出每个 cost function 下令 J(θ)最小的θ，然后取出另 Jcv(θ)最小的一组定位最终的λ。



λ太小导致overfit，产生variance，J(train)<<J(cv)

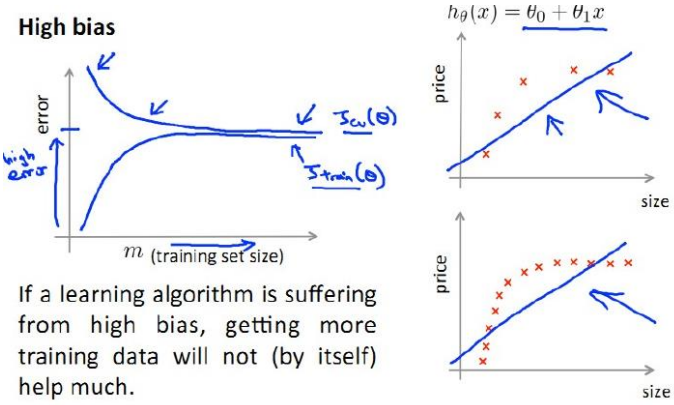
λ太大导致underfit，产生bias，J(train) ≈ J(cv)

4. 那么什么时候增加训练样本才是有效的呢？

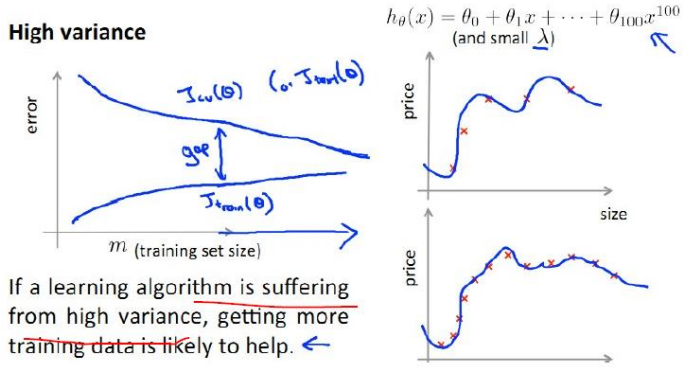


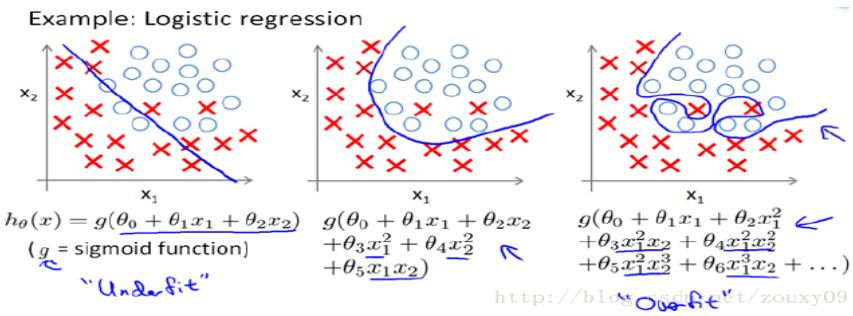
从上面这幅图中我们可知（不知的话用极限思维想想），训练数据越少（如果只有一个），J(train)越小，J(cv)越大；m 越大，J(train)越大（因为越难 perfectly 拟合），J(cv)越小（因为越精确）。分别就 High Bias 和 High Variance 来看看增加 training set 个数，即 m，是否有意义。

High bias：（Underfit 的 high variance:在一定数量后增加 m 无济于事）



High variance:（Overfit 的 high variance:增加 m 使得 J(train)和 J(cv)之间的 gap 减小，有助于性能提高）





又如：

那么，L2 为什么可以防止过拟合呢？L2 范数是指向量各元素的平方和再求平方根。我们如果让 L2 范数的规则项$\|W\|_2$最小，可以使得 W 的每个元素都很小，都接近于 0，但和 L1 范数不同。他不会让它等于 0，而是接近于 0.越小的参数说明模型越简单，越简单的模型则越不容易过拟合。通过 L2 范数，我们可以实现对模型空间的限制，从而在一定程度上避免过拟合。

L2 范数的好处：

1).学习理论的角度：L2 范数可以防止过拟合，提高模型的泛化能力；

2).从优化计算的角度，L2 范数有助于处理 condition number 不好的情况下矩阵求逆困难的问题。

关于 condition number：优化通常有两个难题：一是局部最小值，二是 ill-condition 病态问题。假设我们有个方程 Ax=b,如果 A 或 b 稍微的改变，X 的解就会发生很大的变化，那么这个方程就是 ill-conditioned 的，反之这是 well-conditioned。

如下，左边是 ill-conditioned 的系统,右边是 well-conditioned 的系统：

equations	solution	equations	solution
$\begin{bmatrix}1 & 2\\ 2 & 3.999\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}4\\ 7.999\end{bmatrix}$	$\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}2\\ 1\end{bmatrix}$	$\begin{bmatrix}1 & 2\\ 2 & 3\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}4\\ 7\end{bmatrix}$	$\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}2\\ 1\end{bmatrix}$
$\begin{bmatrix}1 & 2\\ 2 & 3.999\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}4.001\\ 7.998\end{bmatrix}$	$\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}-3.999\\ 4.000\end{bmatrix}$	$\begin{bmatrix}1 & 2\\ 2 & 3\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}4.001\\ 7.001\end{bmatrix}$	$\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}1.999\\ 1.001\end{bmatrix}$
$\begin{bmatrix}1.001 & 2.001\\ 2.001 & 3.998\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}4\\ 7.999\end{bmatrix}$	$\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}3.994\\ 0.001388\end{bmatrix}$	$\begin{bmatrix}1.001 & 2.001\\ 2.001 & 3.001\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}4\\ 7\end{bmatrix}$	$\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}2.003\\ 0.997\end{bmatrix}$

上面的 Condition number 就是用来衡量 ill-condition 系统的可信度的。Condition number 衡量的是输入发生微小变化的时候，输出会发生多大的变化。也就是系统对微小变化的敏感度。Condition number 小的就是 well-conditioned,大的就是 ill-conditioned。

如果方阵 A 是非奇异的，那么 A 的 condition number 就定义为：

$$\mathcal{K}(\mathrm{A}) = \|A\|\|A^{-1}\|$$

也就是矩阵 A 的 norm 乘以他逆的 norm,具体的值，取决于 norm 的选择。为什么要范数呢？范数就相当于衡量一个矩阵的大小或者向量的长度。对于 Ax=b,我们有以下的结论：

$$\frac{\|\Delta x\|}{\|x\|} \le \|A\|\cdot\|A^{-1}\|\cdot\frac{\|\Delta b\|}{\|b\|} \quad \frac{\|\Delta x\|}{\|x\|} \le K(A)\cdot\frac{\|\Delta b\|}{\|b\|} \quad \frac{\|\Delta x\|}{\|x+\Delta x\|} \le K(A)\frac{\|\Delta A\|}{\|A\|}$$

用一句话来总结，就是 condition number 是一个矩阵（或者他所描述的线性系统）的稳定性或敏感度的衡量，如果矩阵的 condition number 在 1 附近，那么他就是 well-conditioned,如果远大于 1，那么就是 ill-conditioned。

回到第一句话，从矩阵或者数值计算的角度来说，L2 范数有助于处理 condition number 不好的情况下矩阵求逆困难的问题。因为目标函数如果是二次的，对于线性回归来说，那实际上是有解析解的，求导并令导数等于零即可得到最优解：

$$\hat{w} = (X^T\mathrm{X})^{-1}X^Ty$$

然而，如果我们的样本 X 的数目比每个样本的维度还要小的时候，矩阵X^TX将不会是满秩的，也就是X^TX将变得不可逆，这样$\hat{w}$将没办法直接计算出来，或者说有无穷多个解（方程的个数小于未知数的个数）。总之，我们过拟合了。

但如果加上 L2 规则项的话，就变成了下面这种情况，可以直接求逆了。

$$\hat{w} = (X^T\mathrm{X} + \lambda I)^{-1}X^Ty$$

我们通常并不直接求矩阵的逆，而是通过解线性方程组的方式（例如高斯消元法）来计算。考虑没有规则项的时候，也就是 λ=0 的情况，如果矩阵X^TX的 condition number 很大的话，解线性方程组就会在数值上相当不稳定，而这个规则项的引入则可以改善 condition number。另外，如果使用迭代优化的算法，condition number 太大仍然会导致问题：它会拖慢迭代的收敛速度，而规则项从优化的角度来看，实际上是将目标函数变成 λ-stronglyconvex（λ 强凸）的了：

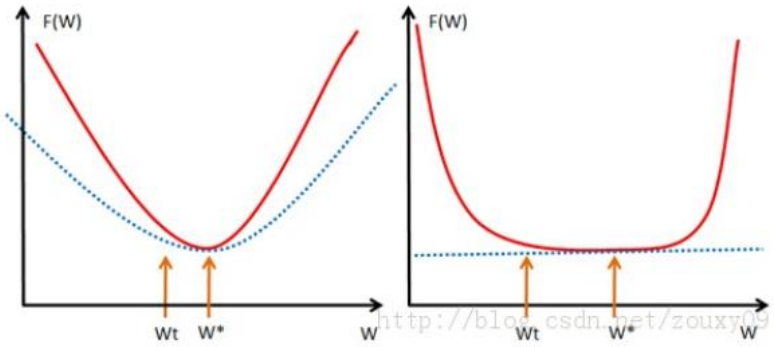
$$f(y) \ge f(x) + < \nabla f(x), y - x > + \frac{\lambda}{2}\|y - x\|^2$$

其中，当 λ = 0是退回到普通 convex 函数的定义（<>表示内积）。

对比普通的凸：

$$f(y) \geq f(x)+< \nabla f(x), y-x > +o(\|y-x\|)$$

直观来讲，convex 性质是指函数曲线位于该点处的切线，也就是线性近似之上，而 strongly convex 则进一步要求位于该处的一个二次函数上方，也就是说要求函数不要太平坦而是保证有一定的向上弯曲的趋势。专业点说，就是 convex 可以保证函数的任意一点的都位于他的一阶泰勒函数之上，而 strongly convex 可以保证函数在任意一点都存在一个非常漂亮的二次下界 quadratic lower bound。参考下图：



我们可以看到，左图在最优解附近的时候，还有较大的梯度值，这样我们可以在较少的迭代次数内达到w^*。而右图在最优解附近非常平缓，优化会变得很慢。

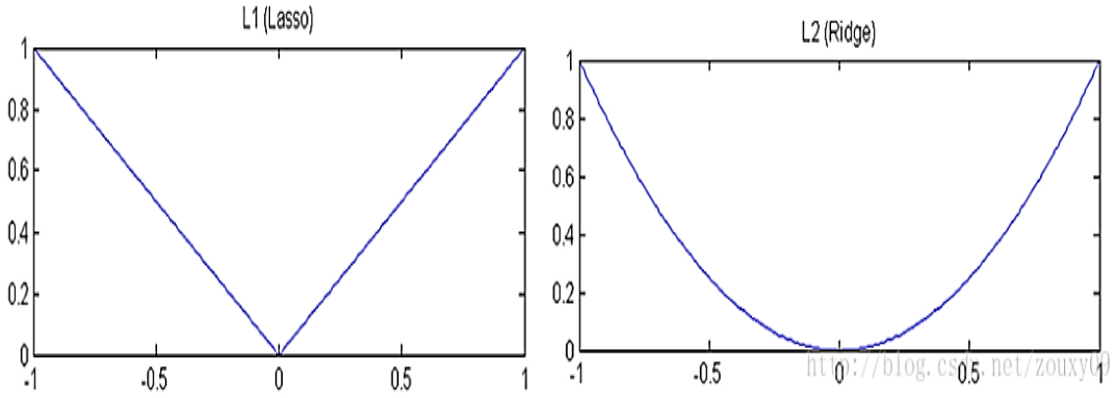
至于有多好呢，这里面有一个 bound，这个 bound 的好坏也要取决于 strongly convex 性质中的常数 α 的大小。看到这里，不知道大家学聪明了没有。如果要获得 strongly convex 怎么做？最简单的就是加入一项：$\frac{\alpha}{2}\|w\|^2$。

实际上，在梯度下降过程中，目标函数收敛的速率上街实际上和矩阵X^TX的 condition number 有关。X^TX 的 condition number 越小，上界就越小，收敛速度也就越快。

综上，L2 不但可以防止过拟合，而且可以当让我们的优化求解变得稳定和快速。

3. L1 和 L2 的差别

1) 下降速度：L1 和 L2 的都是规则化的方式我们将权值参数以 L1 或者 L2 的方式放到代价函数里面去，然后模型就会尝试去最小化这些权值参数。而这个最小化就像一个下坡的过程，L1 和 L2 的差别就在于这个“坡”的不同。如下图所示，L1 就是按绝对值的“坡”下降，L2 是按二次函数的“坡”下降。所以实际在 0 附近，L1 下降速度比 L2 的下降快速很多。



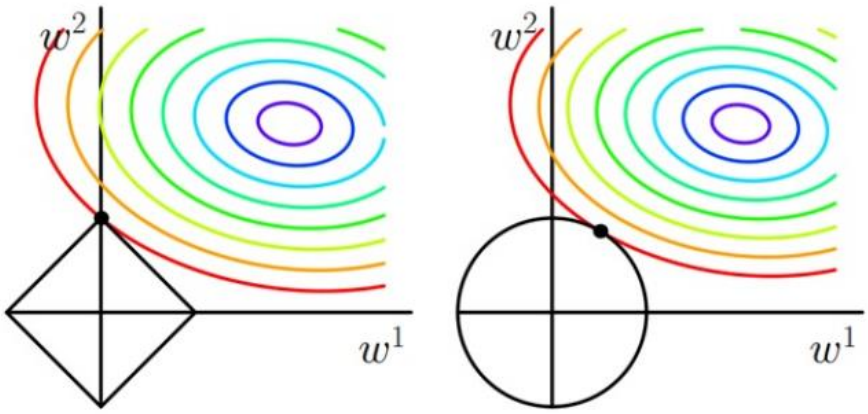
2) 模型空间的限制：

实际上，对于 L1 和 L2 规则化的代价函数来说，我们可以写成：

$$\text{Lasso: } min_w \frac{1}{n}\|y-Xw\|^2, \quad s.t. \|w\|_1 \leq C$$

$$\text{Ridge: } min_w \frac{1}{n}\|y-Xw\|^2, \quad s.t. \|w\|_2 \leq C$$

也就是我们将模型空间限制在 w 的一个 L1-ball 中。为便于可视化，我们考虑二维的情况，在(w1,w2)平面上目标函数的等高线，而约束条件成为平面上半径为 C 的一个 norm ball。等高线与 norm ball 首次相交的地方就是最优解：



(a) ℓ_1-ball meets quadratic function. ℓ_1-ball has corners. It's very likely that the meet-point is at one of the corners.

(b) ℓ_2-ball meets quadratic function. ℓ_2-ball has no corner. It is very unlikely that the meet-point is on any of axes.

可以看到，L1-ball 与 L2-ball 的不同就在于 L1 在和每个坐标轴相交的地方都有“角”出现，而目标函数的测地线除非位置摆得非常好，大部分时候都会在角的地方相交。注意到在角的位置就会产生稀疏性。除了交点以外，还有很多边的轮廓也是既有很大的概率成为第一次相交的地方，又会产生稀疏性。

相比之下，L2 没有这样的性质。因为没有角，所以第一次相交的地方出现在具有稀疏性的位置的概率就变得非常小。

因此一句话：L1 会趋向于产生少量的特征，而其他的特征都是 0；而 L2 会选择很多的特征，这些特征接近于 0.Lasso 在特征选择时候非常有用，而 Ridge 就只是一种规则化而已。

4. 核范数

核范数$\|w\|_*$是指矩阵奇异值的和，英文名叫 Nuclear Norm。相对于 L1 和 L2 而言，可能有些陌生。他的作用却十分霸气：约束 Low-Rank(低秩)。例如在线性代数中：

$$\begin{cases} x_1 - x_2 + x_3 = 5 \\ x_1 + x_2 + x_3 = 7 \\ 2x_1 + 2x_2 + 2x_3 = 14 \end{cases}$$

手工求秩的话，我们通过矩阵初等变换把 A 化为阶梯型矩阵，若该阶梯矩阵有 r 个非零行，那么 A 的秩 rank(A)就等于 r。从物理意义上讲，矩阵的秩度量的就是矩阵行列之间的相关性。如果矩阵的各行或各列之间是线性无关的，那么矩阵就是满秩的。OK，既然秩可以度量相关性，而矩阵的相关性实际上又带有了矩阵的结构信息。如果矩阵之间各行的相关性很强，那么就表示这个矩阵实际可以投影到更低纬度的线性空间，也就是用几个向量就可以完全表达了，他就是低秩的。总结地讲，就是如果矩阵表达的是结构性的信息，例如图像，用户-推荐表等等，那么这个矩阵各行之间存在着一定的相关性，那这个矩阵就是低秩的。

如果 X 是一个 m 行 n 列的数值矩阵，rank(X)是 X 的秩，假如 rank(X)远小于 m 和 n，则我们称 X 为低秩矩阵。低秩矩阵每行或每列都可以用其他的行或列表出，可见他包含大量的冗余信息。利用这种冗余信息，可以对缺失的数据进行恢复，也可以对数据进行特征提取。

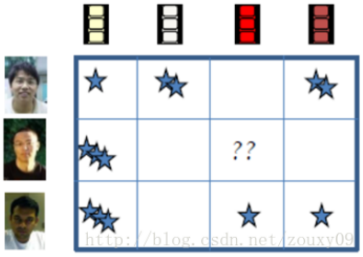
然而，rank()是非凸的，在优化问题里很难求解，那么就需要寻找它的凸近似来近似它。没错，rank(X)的凸近似就是核范数$\|w\|_*$。

一些有意思的应用：

1) 矩阵填充(Matrix Completion)

主流的应用是推荐系统。推荐系统有一种方法是通过分析用户的历史记录来给用户推荐的。例如我们在看一部电影的时候，如果喜欢看，就会给它打个分，例如 3 颗星。然后系统，例如 Netflix 等知名网站就会分析这些数据，看看到底

每部影片的题材到底是怎样的？针对每个人，喜欢怎样的电影，然后会给对应的用户推荐相似题材的电影。但有一个问题是：我们的网站上面有非常多的用户，也有非常多的影片，不是所有的用户都看过说有的电影，不是所有看过某电影的用户都会给它评分。假设我们用一个“用户-影片”的矩阵来描述这些记录，例如下图，可以看到，会有很多空白的地方。如果这些空白的地方存在，我们是很难对这个矩阵进行分析的，所以在分析之前，一般需要先对其进行补全。也叫矩阵填充。



那到底怎么无中生有呢？每个空白的地方的信息是否蕴含在其他已有的信息之上呢？如果有，怎么提取出来呢？Yeah,这就是低秩生效的地方。这叫低秩矩阵重构问题。他可以用如下的模型表示：已知数据是一个给定的 m*n 矩阵 A，如果其中一些元素因为某种原因丢失了，我们能否根据其他的行和列的元素，将这些元素恢复出来呢？当然，如果没有其他的条件，想要恢复很挺难得。但如果我们已知 A 的秩 rank(A)<<m 且 rank(A)<<n,那么我们可以通过矩阵各行（列）之间的线性关系将丢失的元素给出。而这种低秩假设在实际中是十分合理的，比如一个用户对某电影评分是其他用户对这部电影评分的线性组合。所以低秩重构可以预测用户对为评价过的视频的喜好程度，从而实现矩阵填充。

2) 鲁棒 PCA

主成分分析，这种方法可以有效的找出数据中最重要的元素和结构，去除噪声和冗余，将原有数据降维，同时揭示隐藏在复杂数据背后的简单结构。最简单的主成分分析方法就是 PCA。从线性代数的角度看，PCA 的目标就是使用另一组基去重新描述得到的数据空间。希望这组新的基下，能够尽量揭示原有的数据间的关系。

鲁棒主成分分析(robust pca)考虑的是这样一个问题：一般我们的数据矩阵 X 会包含结构信息，也会包含噪声。那么我们可以将这个矩阵分解为两个矩阵相加，一个是低秩的（由于内部有一定的结构信息，造成各行或列之间是线性相关的），另一个是稀疏的（由于含有噪声，而噪声是稀疏的），则鲁棒主成分分析可以写成以下的优化问题：

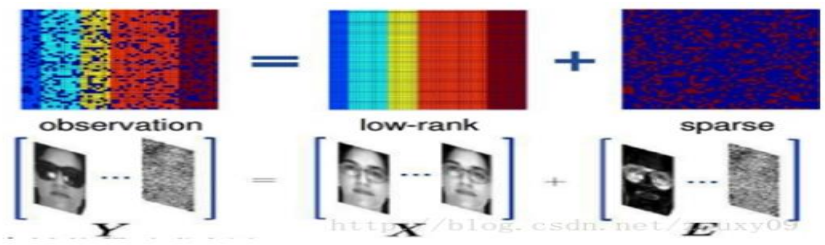
$$min_{A,E} rank(A) + \lambda \|E\|_0 \qquad s.t. X = A + E$$

与经典 PCA 问题一样，鲁棒 PCA 本质上也是寻找数据在低维空间上的最佳投影问题。对于低秩数据观测矩阵 X，假如 X 受到随机（稀疏）噪声的影响，则 X 的低秩性就会破坏，使 X 变成满秩的。所以我们就需要将 X 分解成包含其真实结构的低秩矩阵和稀疏噪声矩阵之和。找到了低秩矩阵，实际上就找到了数据的本质低维空间。那有了 PCA，为什么还有这个 Robust PCA 呢？Robust 在哪？因为 PCA 假设我们的数据的噪声是高斯的，对于大的噪声或者严重的离群点，PCA 会被它影响，导致无法正常工作。而 Robust PCA 则不存在这个假设。它只是假设噪声是稀疏的，而不管噪声的强弱如何。

由于 rank 和 L0 范数在优化上存在非凸和非光滑的特性，所以我们一般将它转化成为求解以下松弛的凸优化问题：

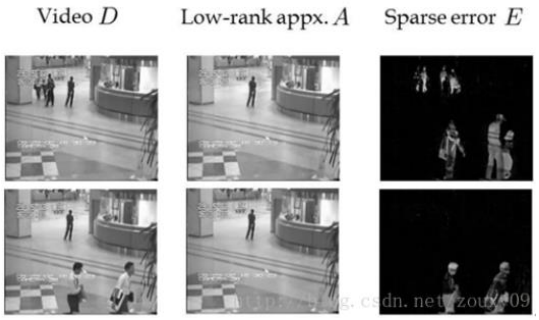
$$min_{A,E} \|A\|_* + \lambda \|E\|_1 \qquad s.t. X = A + E$$

说个应用吧。考虑同一副人脸的多幅图像，如果将每一副人脸图像看成是一个行向量，并将这些向量组成一个矩阵的话，那么可以肯定，理论上，这个矩阵应当是低秩的。但是，由于在实际操作中，每幅图像会受到一定程度的影响，例如遮挡，噪声，光照变化，平移等。这些干扰因素的作用可以看做是一个噪声矩阵的作用。所以我们可以把我们的同一个人脸的多个不同情况下的图片各自拉长一列，然后摆成一个矩阵，对这个矩阵进行低秩和稀疏的分解，就可以得到干净的人脸图像（低秩矩阵）和噪声的矩阵了（稀疏矩阵），例如光照，遮挡等等。



3) 背景建模

背景建模的最简单情形是从固定摄相机拍摄的视频中分离背景和前景。由于背景比较稳定，图像序列帧与帧之间具有极大的相似性，所以仅由背景像素组成的矩阵具有低秩特性；同时由于前景是移动的物体，占据像素比例较低，故前景像素组成的矩阵具有稀疏特性。视频观测矩阵就是这两种特性矩阵的叠加，因此，可以说视频背景建模实现的过程就是低秩矩阵恢复的过程。



4) 变换低秩纹理(TILT)

以上章节所介绍的针对图像的低秩逼近算法，仅仅考虑图像样本之间像素的相似性，却没有考虑到图像作为二维的像素集合，其本身所具有的规律性。事实上，对于未加旋转的图像，由于图像的对称性与自相似性，我们可以将其看做是一个带噪声的低秩矩阵。当图像由端正发生旋转时，图像的对称性和规律性就会被破坏，也就是说各行像素间的线性相关性被破坏，因此矩阵的秩就会增加。

低秩纹理映射算法(TransformInvariant Low-rank Textures，TILT)是一种用低秩性与噪声的稀疏性进行低秩纹理恢复的算法。它的思想是通过几何变换 τ 把 D 所代表的图像区域校正成正则的区域，如具有横平竖直、对称等特性，这些特性可以通过低秩性来进行刻画。

5. 规则化参数选择

现在我们重新看看我们的目标函数：

$$w^* = argmin_w \sum_i L(y_i, f(x_i; w)) + \lambda \Omega(w)$$

里面除了 loss 和规则项两块外，还有一个参数λ。它有个霸气的名字，叫 hyper-parameters(超参)。他是一个非常重要的部分。他的取值很大程度上决定我们模型的性能，事关模型生死。他主要平滑 loss 和规则项两项，λ越大，就表示模型的规则化项要比模型训练误差更重要，也就是相比于拟合我们模型的数据，我们更希望我们的模型能满足我们约束的$\Omega(w)$的特征。反之亦然。如果$\lambda = 0$，那么就是过拟合了。

我们真正希望的，是我们的模型既能拟合我们的数据，又能有我们约束他的特性。只有他们两者的完美结合，才能让我们的模型在我们的任务上发挥出强大的性能。在这点上，大家可能深有体会。还记得你复现了很多论文，然后复现出来的代码跑出来的准确率没有论文说的那么高，甚至还差之万里。这时候，你就会怀疑，到底是论文的问题，还是你实现的问题？实际上，除了这两个问题，我们还需要深入思考另一个问题：论文提出的模型是否具有 hyper-parameters？论文给出了它们的实验取值了吗？经验取值还是经过交叉验证的取值？这个问题是逃不掉的，因为几乎任何一个问题或者模型都会具有 hyper-parameters，只是有时候它是隐藏着的，你看不到而已，但一旦你发现了，证明你俩有缘，那请试着去修改下它吧，有可能有“奇迹”发生哦。

OK，那么我们选择参数λ的目标是什么？我们希望模型的训练误差和泛化能力都很强。这时候，你有可能还反映过来，这不是说我们的泛化性能是我们的参数 λ 的函数吗？那我们为什么按优化那一套，选择能最大化泛化性能的 λ 呢？Oh，sorry to tell you that，因为泛化性能并不是 λ 的简单的函数！它具有很多的局部最大值！而且它的搜索空间很大。所以大家确定参数的时候，一是尝试很多的经验值，这和那些在这个领域摸爬打滚的大师是没得比的。当然了，对于某些模型，大师们也整理了些调参经验给我们。例如 Hinton 大哥的那篇 **A Practical Guide to Training RestrictedBoltzmann Machines** 等等。

还有一种方法是通过分析我们的模型来选择。怎么做呢？就是在训练之前，我们大概计算下这时候的 loss 项的值是多少？Ω(w)的值是多少？然后针对他们的比例来确定我们的 λ，这种启发式的方法会缩小我们的搜索空间。

另外一种最常见的方法就是交叉验证 Cross validation 了。先把我们的训练数据库分成几份，然后取一部分做训练集，一部分做测试集，然后选择不同的 λ 用这个训练集来训练 N 个模型，然后用这个测试集来测试我们的模型，取 N 模型里面的测试误差最小对应的 λ 来作为我们最终的 λ。。

如果我们的模型一次训练时间就很长了，那么很明显在有限的时间内，我们只能测试非常少的 λ 。这就是为什么我们要选择优化也就是收敛速度快的算法，为什么要用 GPU、多核、集群等来进行模型训练、为什么具有强大计算机资源的工业界能做很多学术界也做不了的事情（当然了，大数据也是一个原因）的原因了。努力做个“调参”高手吧！祝愿大家都能“调得一手好参”！

三.矩阵求导

Two competing notational conventions split the field of matrix calculus into two separate groups. The two groups can be distinguished by whether they write the derivative of a scalar with respect to a vector as a column vector or a row vector. Serious mistakes can result when combining results from different authors without carefully verifying that compatible notations are used. Therefore great care should be taken to ensure notational consistency. And a number of authors mix and match their layout choices in various ways. Serious mistakes can result from carelessly combining formulas written in different layouts, and converting from one layout to another requires care to avoid errors.

Derivatives with vectors：

1. Vector-by-scalar： a vector $\mathbf{y} = [y_1, y_2, \dots, y_m]^T$ and a scalar x , then we have

$$\frac{\partial \boldsymbol{y}}{\partial x} = \left[\frac{\partial y_1}{\partial x}, \frac{\partial y_2}{\partial x}, \dots, \frac{\partial y_m}{\partial x}\right]^T$$

2. Scalar-by-vector: a scalar y and a vector $\mathbf{x} = [x_1, x_2, \dots, x_m]^T$, then we have

$$\frac{\partial y}{\partial \boldsymbol{x}} = \left[\frac{\partial \mathrm{y}}{\partial x_1}, \frac{\partial \mathrm{y}}{\partial x_2}, \dots, \frac{\partial \mathrm{y}}{\partial x_m}\right]^T$$

3. Vector-by-vector: a vector $\mathbf{y} = [y_1, y_2, \dots, y_m]^T$ a vector $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \mathrm{y}_1}{\partial x_1} & \cdots & \frac{\partial \mathrm{y}_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathrm{y}_m}{\partial x_1} & \cdots & \frac{\partial \mathrm{y}_m}{\partial x_n} \end{bmatrix}$$

Derivatives with matrices:

1. Matrix-by-scalar: a matrix function $\mathbf{Y}$ and a scalar x

$$\frac{\partial \boldsymbol{Y}}{\partial x} = \begin{bmatrix} \frac{\partial \mathrm{y}_{11}}{\partial x} & \cdots & \frac{\partial \mathrm{y}_{1n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathrm{y}_{m1}}{\partial x} & \cdots & \frac{\partial \mathrm{y}_{mn}}{\partial x} \end{bmatrix}$$

2. Matrix-by-scalar: a scalar y and a matrix $\mathbf{X}$

$$\frac{\partial y}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{p1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1q}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}$$

Other matrix derivatives:

1. matrix derivative of a matrix function F(X) that maps from n×m matrices to p×q matrices:

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial F}{\partial x_{11}} & \cdots & \frac{\partial F}{\partial x_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F}{\partial x_{1m}} & \cdots & \frac{\partial F}{\partial x_{nm}} \end{bmatrix}$$

2. Given $\emptyset$, a differentiable function of an n x m matrix $\mathbf{X} = (x_{i,j})$:

$$\frac{\partial \emptyset}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial \emptyset}{\partial x_{11}} & \cdots & \frac{\partial \emptyset}{\partial x_{1q}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F}{\partial x_{n1}} & \cdots & \frac{\partial F}{\partial x_{nq}} \end{bmatrix}$$

3. Given $\mathbf{F} = (f_{i,j})$, a differentiable m x n function of an n x m matrix $\mathbf{X}$:

$$\frac{\partial \boldsymbol{F}(\boldsymbol{X})}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial f_{1,1}}{\partial X} & \cdots & \frac{\partial f_{1,p}}{\partial X} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{m,1}}{\partial X} & \cdots & \frac{\partial f_{m,p}}{\partial X} \end{bmatrix}$$

Result of differentiating various kinds of aggregates with other kinds of aggregates

	Scalar y		Vector y (size m)		Matrix Y (size $m \times n$)	
	Notation	Type	Notation	Type	Notation	Type
Scalar x	$\frac{\partial y}{\partial x}$	scalar	$\frac{\partial \mathrm{y}}{\partial x}$	(numerator layout) size-m column vector (denominator layout) size-m row vector	$\frac{\partial \mathrm{Y}}{\partial x}$	(numerator layout) $m \times n$ matrix
Vector x (size n)	$\frac{\partial y}{\partial \mathrm{x}}$	(numerator layout) size-n row vector (denominator layout) size-n column vector	$\frac{\partial \mathrm{y}}{\partial \mathrm{x}}$	(numerator layout) $m \times n$ matrix (denominator layout) $n \times m$ matrix	$\frac{\partial \mathrm{Y}}{\partial \mathrm{x}}$	?
Matrix X (size $p \times q$)	$\frac{\partial y}{\partial \mathrm{X}}$	(numerator layout) $q \times p$ matrix (denominator layout) $p \times q$ matrix	$\frac{\partial \mathrm{y}}{\partial \mathrm{X}}$	?	$\frac{\partial \mathrm{Y}}{\partial \mathrm{X}}$	?

The results of operations will be transposed when switching between numerator-layout and denominator-layout notation.

Numerator-layout-notation

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}.$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}.$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}.$$

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{21}} & \cdots & \frac{\partial y}{\partial x_{p1}} \\ \frac{\partial y}{\partial x_{12}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{p2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1q}} & \frac{\partial y}{\partial x_{2q}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}.$$

Denominator-layout notation

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}.$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix}.$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}.$$

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1q}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{p1}} & \frac{\partial y}{\partial x_{p2}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}.$$

Identities:

Identities: vector-by-vector $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

Condition	Expression	Numerator layout, i.e. by y and $\mathbf{x}^{\mathrm{T}}$	Denominator layout, i.e. by $\mathbf{y}^{\mathrm{T}}$ and x
a is not a function of **x**	$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{0}$	
	$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{I}$	
A is not a function of **x**	$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{A}$	$\mathbf{A}^{\mathrm{T}}$
A is not a function of **x**	$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{A}}{\partial \mathbf{x}} =$	$\mathbf{A}^{\mathrm{T}}$	$\mathbf{A}$
a is not a function of **x**, **u** = **u**(**x**)	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	
a = a(**x**), **u** = **u**(**x**)	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial a}{\partial \mathbf{x}}$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial a}{\partial \mathbf{x}}\mathbf{u}^{\mathrm{T}}$
A is not a function of **x**, **u** = **u**(**x**)	$\frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} =$	$\mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}^{\mathrm{T}}$
u = **u**(**x**), **v** = **v**(**x**)	$\frac{\partial (\mathbf{u}+\mathbf{v})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$	
u = **u**(**x**)	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}$
u = **u**(**x**)	$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}$

Identities: scalar-by-vector $\frac{\partial y}{\partial \mathbf{x}} = \nabla_{\mathbf{x}} y$

Condition	Expression	Numerator layout, i.e. by $\mathbf{x}^{\mathrm{T}}$; result is row vector	Denominator layout, i.e. by x; result is column vector
a is not a function of **x**	$\frac{\partial a}{\partial \mathbf{x}} =$	$\mathbf{0}^{\mathrm{T}}$ [4]	$\mathbf{0}$ [4]
a is not a function of **x**, u = u(**x**)	$\frac{\partial au}{\partial \mathbf{x}} =$	$a\frac{\partial u}{\partial \mathbf{x}}$	
u = u(**x**), v = v(**x**)	$\frac{\partial (u+v)}{\partial \mathbf{x}} =$	$\frac{\partial u}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}$	
u = u(**x**), v = v(**x**)	$\frac{\partial uv}{\partial \mathbf{x}} =$	$u\frac{\partial v}{\partial \mathbf{x}} + v\frac{\partial u}{\partial \mathbf{x}}$	
u = u(**x**)	$\frac{\partial g(u)}{\partial \mathbf{x}} =$	$\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$	
u = u(**x**)	$\frac{\partial f(g(u))}{\partial \mathbf{x}} =$	$\frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$	
u = **u**(**x**), **v** = **v**(**x**)	$\frac{\partial (\mathbf{u}\cdot\mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}^{\mathrm{T}}\mathbf{v}}{\partial \mathbf{x}} =$	$\mathbf{u}^{\mathrm{T}}\frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^{\mathrm{T}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ • assumes numerator layout of $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}, \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}\mathbf{u}$ • assumes denominator layout of $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}, \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$
u = **u**(**x**), **v** = **v**(**x**), **A** is not a function of **x**	$\frac{\partial (\mathbf{u}\cdot\mathbf{A}\mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{v}}{\partial \mathbf{x}} =$	$\mathbf{u}^{\mathrm{T}}\mathbf{A}\frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ • assumes numerator layout of $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}, \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}\mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}\mathbf{A}^{\mathrm{T}}\mathbf{u}$ • assumes denominator layout of $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}, \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$

$\mathbf{a}$ is not a function of $\mathbf{x}$	$\frac{\partial (\mathbf{a} \cdot \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x} \cdot \mathbf{a})}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^{\mathrm{T}} \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^{\mathrm{T}} \mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{a}^{\mathrm{T}}$	$\mathbf{a}$
$\mathbf{A}$ is not a function of $\mathbf{x}$ $\mathbf{b}$ is not a function of $\mathbf{x}$	$\frac{\partial \mathbf{b}^{\mathrm{T}} \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{b}^{\mathrm{T}} \mathbf{A}$	$\mathbf{A}^{\mathrm{T}} \mathbf{b}$
$\mathbf{A}$ is not a function of $\mathbf{x}$	$\frac{\partial \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{x}^{\mathrm{T}} (\mathbf{A} + \mathbf{A}^{\mathrm{T}})$	$(\mathbf{A} + \mathbf{A}^{\mathrm{T}}) \mathbf{x}$
$\mathbf{A}$ is not a function of $\mathbf{x}$ $\mathbf{A}$ is symmetric	$\frac{\partial \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} =$	$2\mathbf{x}^{\mathrm{T}} \mathbf{A}$	$2\mathbf{A}\mathbf{x}$
$\mathbf{A}$ is not a function of $\mathbf{x}$	$\frac{\partial^2 \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}}{\partial \mathbf{x}^2} =$	$\mathbf{A} + \mathbf{A}^{\mathrm{T}}$	
$\mathbf{A}$ is not a function of $\mathbf{x}$ $\mathbf{A}$ is symmetric	$\frac{\partial^2 \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}}{\partial \mathbf{x}^2} =$	$2\mathbf{A}$	
	$\frac{\partial (\mathbf{x} \cdot \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^{\mathrm{T}} \mathbf{x}}{\partial \mathbf{x}} =$	$2\mathbf{x}^{\mathrm{T}}$	$2\mathbf{x}$
$\mathbf{a}$ is not a function of $\mathbf{x}$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial (\mathbf{a} \cdot \mathbf{u})}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^{\mathrm{T}} \mathbf{u}}{\partial \mathbf{x}} =$	$\mathbf{a}^{\mathrm{T}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ • assumes numerator layout of $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} \mathbf{a}$ • assumes denominator layout of $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$
$\mathbf{a}$, $\mathbf{b}$ are not functions of $\mathbf{x}$	$\frac{\partial\, \mathbf{a}^{\mathrm{T}} \mathbf{x} \mathbf{x}^{\mathrm{T}} \mathbf{b}}{\partial\, \mathbf{x}} =$	$\mathbf{x}^{\mathrm{T}} (\mathbf{a}\mathbf{b}^{\mathrm{T}} + \mathbf{b}\mathbf{a}^{\mathrm{T}})$	$(\mathbf{a}\mathbf{b}^{\mathrm{T}} + \mathbf{b}\mathbf{a}^{\mathrm{T}}) \mathbf{x}$
$\mathbf{A}$, $\mathbf{b}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{e}$ are not functions of $\mathbf{x}$	$\frac{\partial\, (\mathbf{A}\mathbf{x} + \mathbf{b})^{\mathrm{T}} \mathbf{C} (\mathbf{D}\mathbf{x} + \mathbf{e})}{\partial\, \mathbf{x}} =$	$(\mathbf{D}\mathbf{x} + \mathbf{e})^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} \mathbf{A} + (\mathbf{A}\mathbf{x} + \mathbf{b})^{\mathrm{T}} \mathbf{C} \mathbf{D}$	$\mathbf{D}^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} (\mathbf{A}\mathbf{x} + \mathbf{b}) + \mathbf{A}^{\mathrm{T}} \mathbf{C} (\mathbf{D}\mathbf{x} + \mathbf{e})$
$\mathbf{a}$ is not a function of $\mathbf{x}$	$\frac{\partial\, \|\mathbf{x} - \mathbf{a}\|}{\partial\, \mathbf{x}} =$	$\frac{(\mathbf{x} - \mathbf{a})^{\mathrm{T}}}{\|\mathbf{x} - \mathbf{a}\|}$	$\frac{\mathbf{x} - \mathbf{a}}{\|\mathbf{x} - \mathbf{a}\|}$

Identities: vector-by-scalar $\frac{\partial \mathbf{y}}{\partial x}$

Condition	Expression	Numerator layout, i.e. by $\mathbf{y}$, result is column vector	Denominator layout, i.e. by $\mathbf{y}^{\mathrm{T}}$, result is row vector
$\mathbf{a}$ is not a function of x	$\frac{\partial \mathbf{a}}{\partial x} =$	$\mathbf{0}$[4]	
a is not a function of x, $\mathbf{u} = \mathbf{u}(x)$	$\frac{\partial a\mathbf{u}}{\partial x} =$	$a\frac{\partial \mathbf{u}}{\partial x}$	
$\mathbf{A}$ is not a function of x, $\mathbf{u} = \mathbf{u}(x)$	$\frac{\partial \mathbf{A}\mathbf{u}}{\partial x} =$	$\mathbf{A}\frac{\partial \mathbf{u}}{\partial x}$	$\frac{\partial \mathbf{u}}{\partial x}\mathbf{A}^{\mathrm{T}}$
$\mathbf{u} = \mathbf{u}(x)$	$\frac{\partial \mathbf{u}^{\mathrm{T}}}{\partial x} =$	$\left(\frac{\partial \mathbf{u}}{\partial x}\right)^{\mathrm{T}}$	
$\mathbf{u} = \mathbf{u}(x)$, $\mathbf{v} = \mathbf{v}(x)$	$\frac{\partial (\mathbf{u} + \mathbf{v})}{\partial x} =$	$\frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial x}$	
$\mathbf{u} = \mathbf{u}(x)$, $\mathbf{v} = \mathbf{v}(x)$	$\frac{\partial (\mathbf{u} \times \mathbf{v})}{\partial x} =$	$\mathbf{u} \times \frac{\partial \mathbf{v}}{\partial x} + \frac{\partial \mathbf{u}}{\partial x} \times \mathbf{v}$	
$\mathbf{u} = \mathbf{u}(x)$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial x} =$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial x}$ Assumes consistent matrix layout; see below.	$\frac{\partial \mathbf{u}}{\partial x} \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}$
$\mathbf{u} = \mathbf{u}(x)$	$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial x} =$	$\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}} \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial x}$ Assumes consistent matrix layout; see below.	$\frac{\partial \mathbf{u}}{\partial x} \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}$

Identities: scalar-by-matrix $\frac{\partial y}{\partial \mathbf{X}}$

Condition	Expression	Numerator layout, i.e. by $\mathbf{X}^{\mathrm{T}}$	Denominator layout, i.e. by $\mathbf{X}$
a is not a function of $\mathbf{X}$	$\frac{\partial a}{\partial \mathbf{X}} =$	$\mathbf{0}^{\mathrm{T}}$ [5]	$\mathbf{0}$ [5]
a is not a function of $\mathbf{X}$, $u = u(\mathbf{X})$	$\frac{\partial au}{\partial \mathbf{X}} =$	$a\frac{\partial u}{\partial \mathbf{X}}$	
$u = u(\mathbf{X})$, $v = v(\mathbf{X})$	$\frac{\partial (u+v)}{\partial \mathbf{X}} =$	$\frac{\partial u}{\partial \mathbf{X}} + \frac{\partial v}{\partial \mathbf{X}}$	
$u = u(\mathbf{X})$, $v = v(\mathbf{X})$	$\frac{\partial uv}{\partial \mathbf{X}} =$	$u\frac{\partial v}{\partial \mathbf{X}} + v\frac{\partial u}{\partial \mathbf{X}}$	
$u = u(\mathbf{X})$	$\frac{\partial g(u)}{\partial \mathbf{X}} =$	$\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial \mathbf{X}}$	
$u = u(\mathbf{X})$	$\frac{\partial f(g(u))}{\partial \mathbf{X}} =$	$\frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial \mathbf{X}}$	
$\mathbf{U} = \mathbf{U}(\mathbf{X})$	[6] $\frac{\partial g(\mathbf{U})}{\partial X_{ij}} =$	$\operatorname{tr}\left(\frac{\partial g(\mathbf{U})}{\partial \mathbf{U}}\frac{\partial \mathbf{U}}{\partial X_{ij}}\right)$	$\operatorname{tr}\left(\left(\frac{\partial g(\mathbf{U})}{\partial \mathbf{U}}\right)^{\mathrm{T}}\frac{\partial \mathbf{U}}{\partial X_{ij}}\right)$
		Both forms assume *numerator* layout for $\frac{\partial \mathbf{U}}{\partial X_{ij}}$. i.e. mixed layout if denominator layout for $\mathbf{X}$ is being used.	
	$\frac{\partial \operatorname{tr}(\mathbf{X})}{\partial \mathbf{X}} =$	$\mathbf{I}$	
$\mathbf{U} = \mathbf{U}(\mathbf{X})$, $\mathbf{V} = \mathbf{V}(\mathbf{X})$	$\frac{\partial \operatorname{tr}(\mathbf{U}+\mathbf{V})}{\partial \mathbf{X}} =$	$\frac{\partial \operatorname{tr}(\mathbf{U})}{\partial \mathbf{X}} + \frac{\partial \operatorname{tr}(\mathbf{V})}{\partial \mathbf{X}}$	
a is not a function of $\mathbf{X}$, $\mathbf{U} = \mathbf{U}(\mathbf{X})$	$\frac{\partial \operatorname{tr}(a\mathbf{U})}{\partial \mathbf{X}} =$	$a\frac{\partial \operatorname{tr}(\mathbf{U})}{\partial \mathbf{X}}$	
$\mathbf{g}(\mathbf{X})$ is any polynomial with scalar coefficients, or any matrix function defined by an infinite polynomial series (e.g. $e^{\mathbf{X}}$, $\sin(\mathbf{X})$, $\cos(\mathbf{X})$, $\ln(\mathbf{X})$, etc. using a Taylor series); $g(x)$ is the equivalent scalar function, $g'(x)$ is its derivative, and $\mathbf{g}'(\mathbf{X})$ is the corresponding matrix function	$\frac{\partial \operatorname{tr}(\mathbf{g}(\mathbf{X}))}{\partial \mathbf{X}} =$	$\mathbf{g}'(\mathbf{X})$	$(\mathbf{g}'(\mathbf{X}))^{\mathrm{T}}$
$\mathbf{A}$ is not a function of $\mathbf{X}$	[7] $\frac{\partial \operatorname{tr}(\mathbf{AX})}{\partial \mathbf{X}} = \frac{\partial \operatorname{tr}(\mathbf{XA})}{\partial \mathbf{X}} =$	$\mathbf{A}$	$\mathbf{A}^{\mathrm{T}}$
$\mathbf{A}$ is not a function of $\mathbf{X}$	[6] $\frac{\partial \operatorname{tr}(\mathbf{AX}^{\mathrm{T}})}{\partial \mathbf{X}} = \frac{\partial \operatorname{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{A})}{\partial \mathbf{X}} =$	$\mathbf{A}^{\mathrm{T}}$	$\mathbf{A}$
$\mathbf{A}$ is not a function of $\mathbf{X}$	[6] $\frac{\partial \operatorname{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{AX})}{\partial \mathbf{X}} =$	$\mathbf{X}^{\mathrm{T}}(\mathbf{A}+\mathbf{A}^{\mathrm{T}})$	$(\mathbf{A}+\mathbf{A}^{\mathrm{T}})\mathbf{X}$
$\mathbf{A}$ is not a function of $\mathbf{X}$	[6] $\frac{\partial \operatorname{tr}(\mathbf{X}^{-1}\mathbf{A})}{\partial \mathbf{X}} =$	$-(\mathbf{X}^{-1})^{\mathrm{T}}\mathbf{A}(\mathbf{X}^{-1})^{\mathrm{T}}$	$-\mathbf{X}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{X}^{-1}$
$\mathbf{A}$, $\mathbf{B}$ are not functions of $\mathbf{X}$	$\frac{\partial \operatorname{tr}(\mathbf{AXB})}{\partial \mathbf{X}} = \frac{\partial \operatorname{tr}(\mathbf{BAX})}{\partial \mathbf{X}} =$	$\mathbf{BA}$	$\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}$
$\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are not functions of $\mathbf{X}$	$\frac{\partial \operatorname{tr}(\mathbf{AXBX}^{\mathrm{T}}\mathbf{C})}{\partial \mathbf{X}} =$	$\mathbf{BX}^{\mathrm{T}}\mathbf{CA} + \mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}$	$\mathbf{A}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{XB}^{\mathrm{T}} + \mathbf{CAXB}$
n is a positive integer	[6] $\frac{\partial \operatorname{tr}(\mathbf{X}^n)}{\partial \mathbf{X}} =$	$n\mathbf{X}^{n-1}$	$n(\mathbf{X}^{n-1})^{\mathrm{T}}$
$\mathbf{A}$ is not a function of $\mathbf{X}$, n is a positive integer	[6] $\frac{\partial \operatorname{tr}(\mathbf{AX}^n)}{\partial \mathbf{X}} =$	$\sum_{i=0}^{n-1}\mathbf{X}^i\mathbf{AX}^{n-i-1}$	$\sum_{i=0}^{n-1}(\mathbf{X}^i\mathbf{AX}^{n-i-1})^{\mathrm{T}}$
	[6] $\frac{\partial \operatorname{tr}(e^{\mathbf{X}})}{\partial \mathbf{X}} =$	$e^{\mathbf{X}}$	$(e^{\mathbf{X}})^{\mathrm{T}}$
	[6] $\frac{\partial \operatorname{tr}(\sin(\mathbf{X}))}{\partial \mathbf{X}} =$	$\cos(\mathbf{X})$	$(\cos(\mathbf{X}))^{\mathrm{T}}$
	[8] $\frac{\partial \lvert\mathbf{X}\rvert}{\partial \mathbf{X}} =$	$\operatorname{cofactor}(X)^{\mathrm{T}} = \lvert\mathbf{X}\rvert\mathbf{X}^{-1}$	$\operatorname{cofactor}(X) = \lvert\mathbf{X}\rvert(\mathbf{X}^{-1})^{\mathrm{T}}$
a is not a function of $\mathbf{X}$	[6] $\frac{\partial \ln\lvert a\mathbf{X}\rvert}{\partial \mathbf{X}} =$ [9]	$\mathbf{X}^{-1}$	$(\mathbf{X}^{-1})^{\mathrm{T}}$
$\mathbf{A}$, $\mathbf{B}$ are not functions of $\mathbf{X}$	[6] $\frac{\partial \lvert\mathbf{AXB}\rvert}{\partial \mathbf{X}} =$	$\lvert\mathbf{AXB}\rvert\mathbf{X}^{-1}$	$\lvert\mathbf{AXB}\rvert(\mathbf{X}^{-1})^{\mathrm{T}}$
n is a positive integer	[6] $\frac{\partial \lvert\mathbf{X}^n\rvert}{\partial \mathbf{X}} =$	$n\lvert\mathbf{X}^n\rvert\mathbf{X}^{-1}$	$n\lvert\mathbf{X}^n\rvert(\mathbf{X}^{-1})^{\mathrm{T}}$
(see pseudo-inverse)	[6] $\frac{\partial \ln\lvert\mathbf{X}^{\mathrm{T}}\mathbf{X}\rvert}{\partial \mathbf{X}} =$	$2\mathbf{X}^{+}$	$2(\mathbf{X}^{+})^{\mathrm{T}}$
(see pseudo-inverse)	[6] $\frac{\partial \ln\lvert\mathbf{X}^{\mathrm{T}}\mathbf{X}\rvert}{\partial \mathbf{X}^{+}} =$	$-2\mathbf{X}$	$-2\mathbf{X}^{\mathrm{T}}$
$\mathbf{A}$ is not a function of $\mathbf{X}$, $\mathbf{X}$ is square and invertible	$\frac{\partial \lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert}{\partial \mathbf{X}} =$	$2\lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert\mathbf{X}^{-1}$	$2\lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert(\mathbf{X}^{-1})^{\mathrm{T}}$
$\mathbf{A}$ is not a function of $\mathbf{X}$, $\mathbf{X}$ is non-square, $\mathbf{A}$ is symmetric	$\frac{\partial \lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert}{\partial \mathbf{X}} =$	$2\lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert(\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$	$2\lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert\mathbf{AX}(\mathbf{X}^{\mathrm{T}}\mathbf{AX})^{-1}$
$\mathbf{A}$ is not a function of $\mathbf{X}$, $\mathbf{X}$ is non-square, $\mathbf{A}$ is non-symmetric	$\frac{\partial \lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert}{\partial \mathbf{X}} =$	$\lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert((\mathbf{X}^{\mathrm{T}}\mathbf{AX})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A} + (\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}})$	$\lvert\mathbf{X}^{\mathrm{T}}\mathbf{AX}\rvert(\mathbf{AX}(\mathbf{X}^{\mathrm{T}}\mathbf{AX})^{-1} + \mathbf{A}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{X})^{-1})$

Identities: matrix-by-scalar $\frac{\partial \mathbf{Y}}{\partial x}$

Condition	Expression	Numerator layout, i.e. by Y
$\mathbf{U} = \mathbf{U}(x)$	$\frac{\partial a\mathbf{U}}{\partial x} =$	$a\frac{\partial \mathbf{U}}{\partial x}$
$\mathbf{A}$, $\mathbf{B}$ are not functions of x, $\mathbf{U} = \mathbf{U}(x)$	$\frac{\partial \mathbf{AUB}}{\partial x} =$	$\mathbf{A}\frac{\partial \mathbf{U}}{\partial x}\mathbf{B}$
$\mathbf{U} = \mathbf{U}(x)$, $\mathbf{V} = \mathbf{V}(x)$	$\frac{\partial (\mathbf{U}+\mathbf{V})}{\partial x} =$	$\frac{\partial \mathbf{U}}{\partial x} + \frac{\partial \mathbf{V}}{\partial x}$
$\mathbf{U} = \mathbf{U}(x)$, $\mathbf{V} = \mathbf{V}(x)$	$\frac{\partial (\mathbf{UV})}{\partial x} =$	$\mathbf{U}\frac{\partial \mathbf{V}}{\partial x} + \frac{\partial \mathbf{U}}{\partial x}\mathbf{V}$
$\mathbf{U} = \mathbf{U}(x)$, $\mathbf{V} = \mathbf{V}(x)$	$\frac{\partial (\mathbf{U}\otimes\mathbf{V})}{\partial x} =$	$\mathbf{U}\otimes\frac{\partial \mathbf{V}}{\partial x} + \frac{\partial \mathbf{U}}{\partial x}\otimes\mathbf{V}$
$\mathbf{U} = \mathbf{U}(x)$, $\mathbf{V} = \mathbf{V}(x)$	$\frac{\partial (\mathbf{U}\circ\mathbf{V})}{\partial x} =$	$\mathbf{U}\circ\frac{\partial \mathbf{V}}{\partial x} + \frac{\partial \mathbf{U}}{\partial x}\circ\mathbf{V}$
$\mathbf{U} = \mathbf{U}(x)$	$\frac{\partial \mathbf{U}^{-1}}{\partial x} =$	$-\mathbf{U}^{-1}\frac{\partial \mathbf{U}}{\partial x}\mathbf{U}^{-1}$
$\mathbf{U} = \mathbf{U}(x, y)$	$\frac{\partial^2 \mathbf{U}^{-1}}{\partial x \partial y} =$	$\mathbf{U}^{-1}\left(\frac{\partial \mathbf{U}}{\partial x}\mathbf{U}^{-1}\frac{\partial \mathbf{U}}{\partial y} - \frac{\partial^2 \mathbf{U}}{\partial x \partial y} + \frac{\partial \mathbf{U}}{\partial y}\mathbf{U}^{-1}\frac{\partial \mathbf{U}}{\partial x}\right)\mathbf{U}^{-1}$
$\mathbf{A}$ is not a function of x, $\mathbf{g}(\mathbf{X})$ is any polynomial with scalar coefficients, or any matrix function defined by an infinite polynomial series (e.g. $e^{\mathbf{X}}$, $\sin(\mathbf{X})$, $\cos(\mathbf{X})$, $\ln(\mathbf{X})$, etc.); $g(x)$ is the equivalent scalar function, $g'(x)$ is its derivative, and $\mathbf{g}'(\mathbf{X})$ is the corresponding matrix function	$\frac{\partial \mathbf{g}(x\mathbf{A})}{\partial x} =$	$\mathbf{A}\mathbf{g}'(x\mathbf{A}) = \mathbf{g}'(x\mathbf{A})\mathbf{A}$
$\mathbf{A}$ is not a function of x	$\frac{\partial e^{x\mathbf{A}}}{\partial x} =$	$\mathbf{A}e^{x\mathbf{A}} = e^{x\mathbf{A}}\mathbf{A}$

Identities: scalar-by-scalar, with vectors involved

Condition	Expression	Any layout (assumes dot product ignores row vs. column layout)
$\mathbf{u} = \mathbf{u}(x)$	$\frac{\partial g(\mathbf{u})}{\partial x} =$	$\frac{\partial g(\mathbf{u})}{\partial \mathbf{u}} \cdot \frac{\partial \mathbf{u}}{\partial x}$
$\mathbf{u} = \mathbf{u}(x)$, $\mathbf{v} = \mathbf{v}(x)$	$\frac{\partial (\mathbf{u}\cdot\mathbf{v})}{\partial x} =$	$\mathbf{u}\cdot\frac{\partial \mathbf{v}}{\partial x} + \frac{\partial \mathbf{u}}{\partial x}\cdot\mathbf{v}$

四.Matlab 编程

L2 和 Fro 范数：平方和开根号

故有：

$\|A\|_2$ => norm(A,' fro')^2

$\sum_i \sum_j W_{ij}\|F_i - F_j\|^2 = 2 * trace(F^T(D-W)F)$=>故通常为：$\frac{1}{2}\sum_i \sum_j W_{ij}\|F_i - F_j\|^2 = trace(F^T(D-W)F)$

$\sum_i \|F_i - Y_i\|^2 = trace((F-Y)^T(F-Y))$

五.参考文献

Web: http://en.wikipedia.org/wiki/Matrix_calculus

http://blog.csdn.net/zouxy09/article/details/24971995

http://blog.csdn.net/abcjennifer/article/details/7797502

Paper: