# Regualrization and Its Application in Data Mining

**Wenbao Li**

**liwenbao930116@gmail.com**

**University of Electronic Science and Technology of China**

# Outline

- **The concept of regularization**
- **The theory of regularization**
- **The application of regularization**
  - In data mining/machine learning
    - In multi-task learning

# What is regularization

- **Two problems:**
  - Ill-posed inverse problem

According to Hadamard,1915 : Given mapping $A\!:\!X \rightarrow Y$, equation

$$Ax = y$$

is well - posed provided

- a solution exists for each $y \in Y, \exists x \in X$ such that $Ax = y$

- the solution is unique i.e. $Ax_1 = Ax_2 \Rightarrow x_1 = x_2$

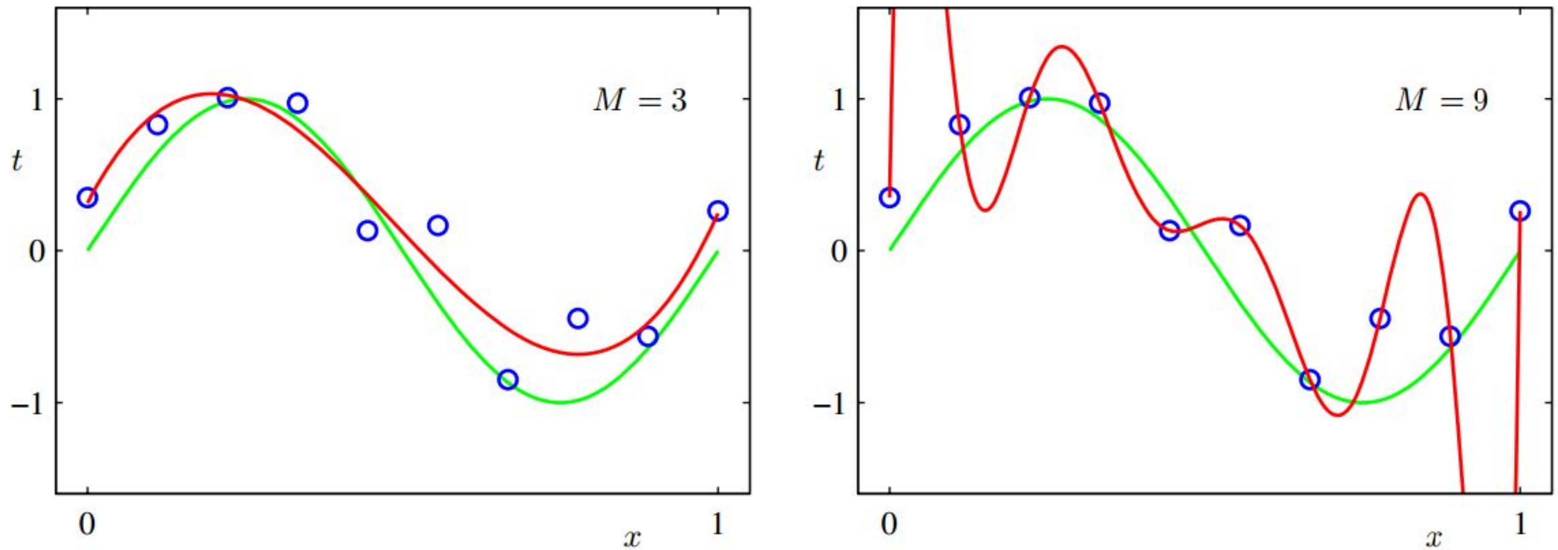
- the solution is stable i.e. $A^{-1}$ is continuous

An equation is *ill-posed* if it is not *well-posed.*

  - So,how do we solute such problem $Ax=y$ which is ill-posed

# What is regularization

- **Two problems:**
  - Overfitting in machine learning regression problems:




$$t = \sum_{i=0}^{m} \theta_i x^i$$

  - So,how do we decide which model is to be selected?

# What is regularization

- **Definition:**
  - Regularization was first introduced in the context of solving integral equation numerically by Tikhonov(1943).
  - (Wikipedia)Regularization, in mathematics and statistics and particularly in the fields of machine learning and *inverse problems*, refers to *a process of introducing additional prior information in order to solve an ill-posed problem or to prevent overfitting*.
  - (Inverse problems)Informally,Regularization is defined as it "*Imposes stability on an ill-posed problem in a manner that yields accurate approximate solutions*,often by incorporating prior information".
  - One simple form of regularization is

$$\min_{x} \|Ax - y\| + \gamma \|x\|$$

# What is regularization

- **Definition:**
  - Regularization provides methods for
    - finding approximate and stable solutions of the ill-posed inverse problems.
    - preventing overfitting or ensure the smoothness of regression function or solution.
  - It was first designed for solving the ill-posed inverse problem,but later give rise to regularized learning algorithms.

# The theory of regularization

- **The generalized regularization form**
  - Linear System

$$origi: \min_x \|Ax - y\|$$
$$regularized: \min_x \|Ax - y\| + \gamma \|x\|$$

  - Learning algorithm system

$$\min_W L[h(W, X), Y] + \lambda \|g(W)\|, \quad eg: \ h(X) = W^T X,$$

- The first term make sure that the measurement of fitting or the degree of consistence with the training examples.
- The second term make sure the simpler model or not the extreme solutions.
- So,here are two parameters $\lambda$ and $g(W)$ to be decided.

# The theory of regularization

- **Typical regularization method—L1-norm regularization**

$$\min \|Ax - b\|_2 + \lambda \|x\|_1$$

  – By varying the parameter $\delta$ we can sweep out the optimal trade-off curve between $\|Ax-b\|_2$ and $\|x\|_1$ ,which serves as an approximation of the optimal trade-off curve between $\|Ax-b\|_2$ and the sparsity or cardinality card($x$) of the vector x,i.e.,the number of nonzero elements.

# The theory of regularization

- **Typical regularization method—Tikhonov regularization**

$$\min \|Ax-b\|_2^2 + \lambda \|\Gamma x\|_2^2$$

$$x = \left(A^T A + \lambda \Gamma^T \Gamma\right)^{-1} A^T b$$

  - The penalty term is the form of squared L2 norm of x.
  - $\Gamma$ is the tikhonov matrix or tikhonov operator.When $\Gamma = I$ ,it becomes the standard form. In many cases, $\Gamma = \alpha I$
  - $\lambda > 0$ is the regularization parameter.

- **Typical regularization method—Smooth regularization method(Special case of Tikhonov regularization)**

$$\min \|Ax-b\|_2^2 + \delta\|\Delta x\|_2^2$$

$$\min \|Ax-b\|_2^2 + \delta\|\Delta x\|_2^2 + \eta\|x\|_2^2$$

$\Delta$ is typically the discretization of a derivative operator of first or second order.And the interpretation of it is the smoothness of x.

# The theory of regularization

- **Typical regularization method—Iterative Tikhonov regularization**
  - Once we have computed the Tikhonov solution , we may find a better approximation by applying Tikhonov regularization again using the previous finding soloution as initial solution.

$$\min \|Ax-b\|_2^2 + \lambda \|x\|_2^2$$

$$x_0 = \mathbf{0},\quad x_k = \left(A^T A + \lambda I\right)^{-1} (A^T b + \lambda x_{k-1}), \text{for k} = 1,2,\ldots \text{t - 1}$$

  - Parameter: $\lambda$ and $t$
  - Advantages:

# The theory of regularization

- **Typical regularization method—Landweber iteration**

$$\min J(x) = \min \|Ax - b\|_2^2 + \lambda \|x\|_2^2$$

  - use gradient descent

$$\frac{1}{2}\nabla J(x) = A^T(Ax - b) + \lambda x = (A^T A + \lambda I)x - A^T b$$

$$x_0 = \mathbf{0}, \quad x_k = x_{k-1} - \frac{\mu}{2}\nabla J(x_{k-1})$$

$$= x_{k-1} - \mu((A^T A + \lambda I)x_{k-1} - A^T b) \text{for k} = 1,2,\ldots \text{t - 1}$$

  - use induction method we can derive

$$x_n = \sum_{j=0}^{n-1}\left[(1 - \mu\lambda)I - \mu A^T A\right]^j A^T b$$

# The theory of regularization

- **Typical regularization method—Bregman Iterative regularization(used for image restoring when first proposed)**
  – Probelm: $\min_{u} J(u)+H(u), J(u)$ is regularizer

Require: $J(\bullet), H(\bullet)$

1.Initialize: $k=0, u^0=\mathbf{0}, p^0=\mathbf{0}.$

2.**while** "not converge" **do**

3.$u^{k+1} \leftarrow \arg\min_{u} D_J^{p^k}(u,u^k)+H(u)$

where $D_J^p(u,v)=J(u)-J(v)-\langle p,u-v\rangle$

4.$p^{k+1} \leftarrow p^k-\nabla H(u^{k+1}) \in \partial J(u^{k+1})$

5.$k \leftarrow k+1$

6.**end while**

- **Typical regularization method—Truncated SVD**
  - Idea:Cut off components corresponding to small singular values.

$A \in R^{m\times n}$ has singular value decomposition $A = U\Sigma V^T$

$U,V \in R^{m\times n}$ are orthogonal.$U = (u_1, u_2, \cdots, u_n), V = (v_1, v_2, \cdots, v_n)$

$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_n)$,

$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_k > \sigma_{k+1} = \cdots = \sigma_n = 0$

$rank(A) = k$

  - The definition of TSVD of A is

$$A_k = U\Sigma_k V^T = \sum_{i=1}^{k} u_i \sigma_i v_i^T, \Sigma_k = diag(\sigma_1, ..., \sigma_k, 0, ..., 0) \in R^{m\times n}$$

  - The TSVD solution of $\min\|Ax-b\|_2$

$$x_k = Vdiag(\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_k}, 0, \cdots, 0)U^T b = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i$$

# The theory of regularization

- **Typical regularization method—Truncated SVD regularization**
  - Consider now regularization in standard form

$$\min \|Ax-b\|_2^2 + \lambda \|x\|_2^2$$
$$x = \left(A^T A + \lambda I\right)^{-1} A^T b$$

  - The definition of TSVD of A is

$$A_k = U\Sigma_k V^T = \sum_{i=1}^{k} u_i \sigma_i v_i^T, \Sigma_k = diag(\sigma_1,\ldots,\sigma_k,0,\ldots,0) \in R^{m\times n}$$

  - The TSVD solution of $\min \|Ax-b\|_2^2 + \lambda \|x\|_2^2$

$$x_k = V diag\left(\frac{\sigma_1}{\sigma_1^2+\lambda},\cdots,\frac{\sigma_k}{\sigma_k^2+\lambda},0,\cdots,0\right)U^T b = \sum_{i=1}^{k} \frac{\sigma_i u_i^T b}{\sigma_i^2+\lambda} v_i$$

# The theory of regularization

- **Typical regularization method—Truncated SVD regularization(cont.)**
  - filter out the contributions to the solution corresponding to the smallest singular values

$$x_k = V diag(\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_k}, 0, \cdots, 0) U^T b = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i$$

$$x_k = V diag(\frac{\sigma_1}{\sigma_1^2 + \lambda}, \cdots, \frac{\sigma_k}{\sigma_k^2 + \lambda}, 0, \cdots, 0) U^T b = \sum_{i=1}^{k} \frac{\sigma_i u_i^T b}{\sigma_i^2 + \lambda} v_i$$

  - The filter function can be shown as following

$$f_i = \begin{cases} 1/\sigma_i, \sigma_i \geq \sigma_k \\ 0, \sigma_i < \sigma_k \end{cases} \qquad f_i = \frac{\sigma_i}{\sigma_i^2 + \lambda}, i = 1, 2, \cdots, n$$

# The theory of regularization

- **The relation of regularization in linear system and in learning algorithm system**
  - training set $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
  - X is the n by d input matrix.
  - $Y=(Y_1,\ldots,Y_n)$ is the output vector.
  - k denotes the kernel function, K is the n by n kernel matrix with entries $K_{ij}=k(X_i,X_j)$ and H is the RKHS with kernel k.
  - RLS estimator solves

$$\min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_H^2$$

  - And we know the solution is

$$f_S^\lambda(x) = \sum_{i=1}^{n} c_i k(x, x_i) \text{ with } (K + n\lambda I)c = Y$$

- **ERM**
  - Similarly we can prove that the solution of empirical risk minimization

$$\min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

  - can be written as

$$f_S(x) = \sum_{i=1}^{n} c_i k(x, x_i) \text{ with } Kc = Y, \quad c = (c_1, c_2, \dots, c_n)$$

  - So,what we should do is solving the problem Kc = Y

- **The role of regularization**
  - We observed that adding a penalization term can be interpreted as way to to control smoothness and avoid overfitting.

*Learning System*

$$\min_{f\in H}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \Rightarrow \min_{f\in H}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda\|f\|_H^2.$$

  - From a numerical point of view:

$$Kc = Y \Rightarrow (K + n\lambda I)c = Y$$

*Linear System*

    - It stabilizes a possibly ill-conditioned matrix inversion problem.
    - This is the point of view of regularization for (ill-posed) inverse problems.

# The theory of regularization

- **Regularization as a filter**
  - Goal:solve $Kc = Y$
  - In the finite-dimensional case, the main problem is numerical stability. For example, let the kernel matrix have

$$K = Q\Sigma Q^T, \Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_n), \sigma \geq \sigma_2 \geq \ldots \geq 0$$

$Q = (q_1, q_2, \ldots, q_n), q_i$ is the corresponding eigenvectors of $K$

then

$$c = K^{-1}Y = (Q\Sigma Q^T)^{-1}Y = Q\Sigma^{-1}Q^T Y = \sum_{i=1}^{n} \frac{1}{\sigma_i} \langle q_i, Y \rangle q_i.$$

  - But $K^{-1}$ doesn't always exist .That is terms in this sum with small eigenvalues $\sigma_i$ give rise to numerical instability. For instance, if there are eigenvalues of zero, the matrix will be impossible to invert. As eigenvalues tend toward zero, the matrix tends toward rank-deficiency, and inversion becomes less stable. Statistically, this will correspond to high variance of the coefficients $c_i$.

- **Regularization as a filter(cont.)**
  - So,we take regularization into account.For example,tikhonov regularization

$$(K+n\lambda I)c=Y$$

then

$$c=(K+n\lambda I)^{-1}Y=(Q(\Sigma+n\lambda I)Q^{T})^{-1}Y=Q(\Sigma+n\lambda I)^{-1}Q^{T}Y=\sum_{i=1}^{n}\frac{1}{\sigma_i+n\lambda}\langle q_i,Y\rangle q_i.$$

  - This shows that regularization as the effect of suppressing the influence of small eigenvalues in computing the inverse. In other words, regularization flters out the undesired components.

# The theory of regularization

- **Regularization as a filter(cont.)**
  - So,we can define more general flters.Let $G_\lambda(\sigma)$ be a function on the kernel matrix.We can eigendecompose $K$ to define

$$G_\lambda(K) = QG_\lambda(\Sigma)Q^T$$

  - meaning

$$G_\lambda(K)Y = \sum_{i=1}^{n} G_\lambda(\sigma_i)\langle q_i, Y\rangle q_i.$$

  - For Tikhonov Regularization

$$G_\lambda(\sigma) = \frac{1}{\sigma + n\lambda}$$

# The theory of regularization

- **Regularization as a filter(cont.)**
  - For Landweber Iteration

$$c = \mu \sum_{i=0}^{t-1} (I - \mu K)^i Y$$

$$G_\lambda(\sigma) = \mu \sum_{i=0}^{t-1} (I - \mu\sigma)^i Y$$

  - For TSVD

$$G_\lambda(\sigma) = \begin{cases} 1/\sigma & ,\sigma > \mathrm{n}\lambda \\ 0 & ,\mathrm{otherwise} \end{cases}$$

• • •

# The theory of regularization

- **Regularization parameter selection criterion(for solving the inverse problem)**
  - Gfrerer / Raus method

$$\lambda^3 b^T (AA^T + \lambda I)^{-3} b = \|e\|^2$$

  - Morozov's discrepancy principle(Ask for the norm of the residual to be equal to the norm of the noise vector)

$$\left\| b - A(A^T A + \lambda I)^{-1} A^T b \right\| = \|e\|$$

  - The quasi-optimality criterion

$$\min\left[\lambda^2 b^T A(A^T A + \lambda I)^{-4} A^T b\right]$$

  - Wahba:generalized cross validation
  - Hansen:L-curve

# The theory of regularization

- **Regularization parameter selection criterion**
  - Gfrerer / Raus method

$$\lambda^3 b^T (AA^T + \lambda I)^{-3} b = \|e\|^2$$

- **Regularization parameter selection criterion**
  - Morozov's discrepancy principle
    - Ask for the norm of the residual to be equal to the norm of the noise vector(take tikhonov regularization as example)

$$\|Ax_\lambda - b\| = \|e\|$$

$$\|A(A^T A + \lambda I)^{-1} A^T b - b\| = \|e\|$$

# The theory of regularization

[Frank Bauer, Markus Reiÿ:Regularization independent of the noise level:an analysis of quasi-optimality]

- **Regularization parameter selection criterion**
  - The quasi-optimality criterion
    - take tikhonov regularization as example $x_\lambda = (A^T A + \lambda I)^{-1} A^T b$
    - Idea:choose parameter $\lambda > 0$ such that

$$\left\| \lambda \frac{dx_\lambda}{d\lambda} \right\| \to \min_\lambda$$

$$\frac{dx_\lambda}{d\lambda} = -(A^T A + \lambda I)^{-2} A^T b$$

$$\left\| \lambda \frac{dx_\lambda}{d\lambda} \right\| = \lambda^2 \left( -(A^T A + \lambda I)^{-2} A^T b \right)^T \left( -(A^T A + \lambda I)^{-2} A^T b \right)$$

$$= \lambda^2 b^T A (A^T A + \lambda I)^{-4} A^T b$$

$$\min \left[ \lambda^2 b^T A (A^T A + \lambda I)^{-4} A^T b \right]$$

# The theory of regularization

[Golub;Heath;Wahba:generalized cross validation as a method for choosing a good ridge parameter]

- **Regularization parameter selection criterion**
  - Wahba:GCV
    - For ridge regression problem

$$y = X\beta + \varepsilon$$

    - The ridge estimate is

$$\hat{\beta}(\lambda) = (X^T X + n\lambda I)^{-1} X^T y$$

    - The GCV estimate of the parameter $\lambda$ is the minimizer of $V(\lambda)$

$$V(\lambda) = \frac{\frac{1}{n}\left\| (I - X(X^T X + n\lambda I)^{-1} X^T) y \right\|_2^2}{\left[ \frac{1}{n} Trace(I - X(X^T X + n\lambda I)^{-1} X^T) \right]^2}$$

# The theory of regularization

[P.C. Hansen:The L-curve and its use in the numerical treatment of inverse problems]

- **Regularization parameter selection criterion**
  - Hansen:L-curve
    - For a regularization problem such as tikhonov regularization,there are two parts to be minimize,the regularization solution norm and the residual norm

$$\min \|Ax-b\|_2^2+\lambda\|L(x-x_0)\|_2^2 (generalized\ \ form)$$

    - L-curve is actually the plot of these two quantities versus each other,i.e.,as a curve

$$\left(\|Ax_\lambda-b\|_2, \|L(x_\lambda-x_0)\|_2\right)$$

# The theory of regularization

[P.C. Hansen:The L-curve and its use in the numerical treatment of inverse problems]

- **Regularization parameter selection criterion**
  - Hansen:L-curve(cont.)

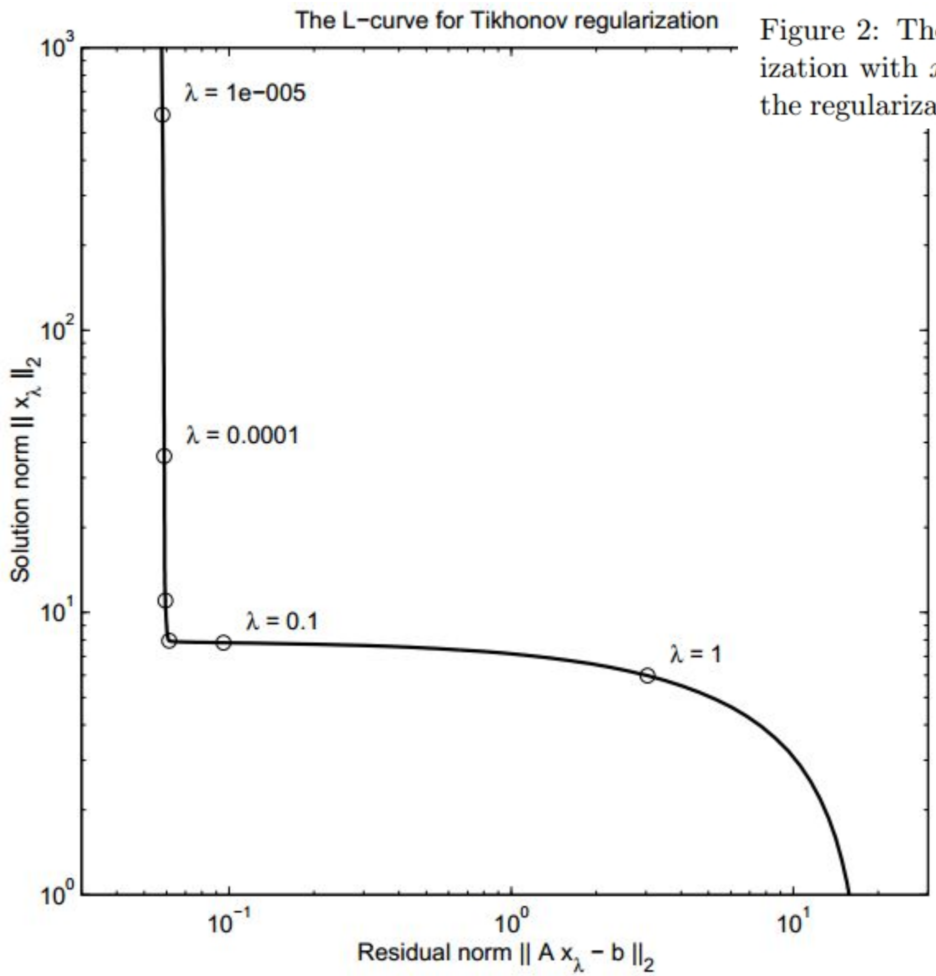



Figure 2: The generic L-curve for standard-form Tikhonov regularization with $x_0 = 0$; the points marked by the circles correspond to the regularization parameters $\lambda = 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$ and 1.

$$L = I, x_0 = 0$$

The corner point is what we want

[P.C. Hansen:The L-curve and its use in the numerical treatment of inverse problems]

- **Regularization parameter selection criterion**
  - Hansen:L-curve(cont.)
    - The definition of corner of L-curve
      - the point on the L-curve $(\hat{\rho}/2, \hat{\eta}/2)$ with maximum curvature $\kappa$ given by equation

$$\kappa = 2\frac{\eta\rho}{\eta'}\frac{\lambda^2\eta'\rho + 2\lambda\eta\rho + \lambda^4\eta\eta'}{(\lambda^2\eta^2 + \rho^2)^{3/2}}$$

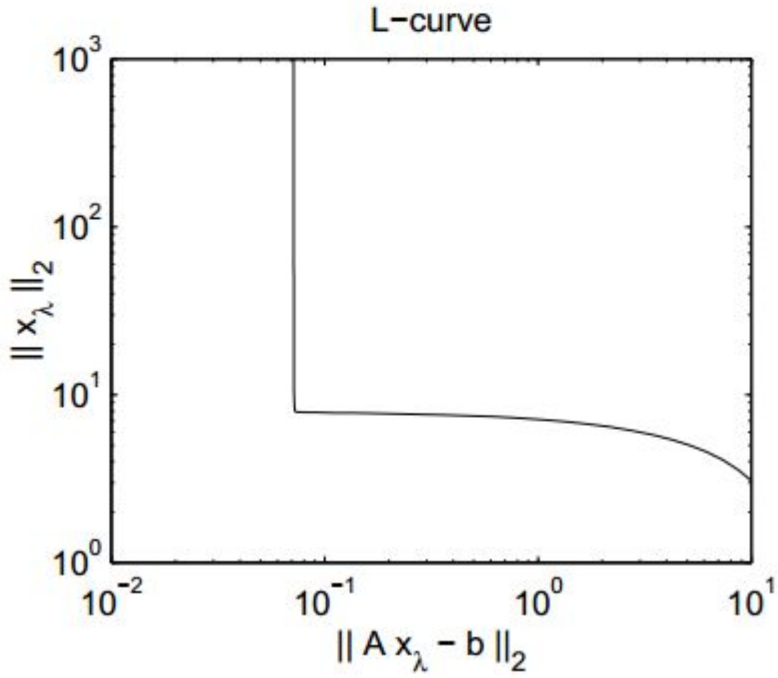
      - where

$$\eta = \|x_\lambda\|_2^2, \rho = \|Ax_\lambda - b\|_2^2$$
$$\hat{\eta} = \log\eta, \hat{\rho} = \log\rho$$
$$\eta' = -\frac{4}{\lambda}\sum_{i=1}^{n}(1-f_i)f_i^2\frac{(u_i^T)}{\sigma_i^2}, f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}$$

# The theory of regularization

[P.C. Hansen:The L-curve and its use in the numerical treatment of inverse problems]

- **Regularization parameter selection criterion**
  – Hansen:L-curve(cont.)



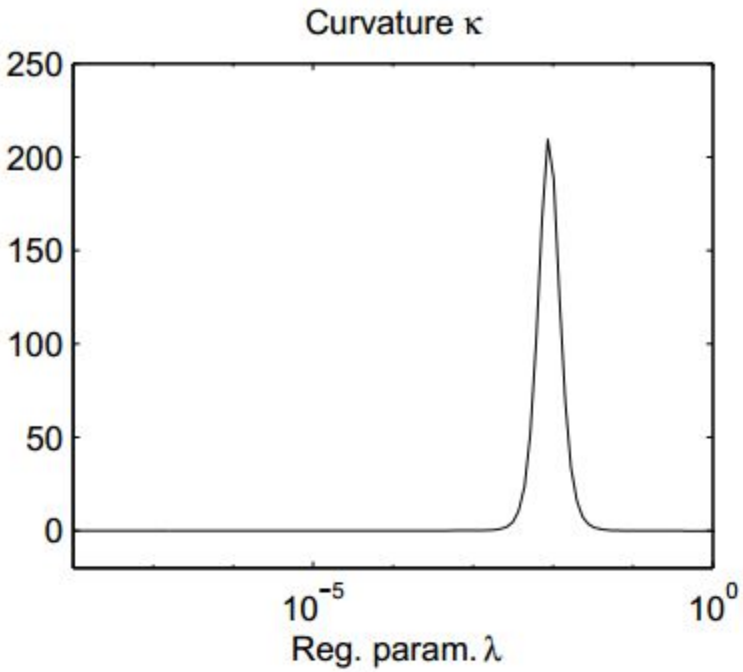




Figure 3: A typical L-curve (left) and a plot (right) of the corresponding curvature $\kappa$ as a function of the regularization parameter.

# The application of regularization

- **Application in machine learning—Ridge regression**
  - In the context of linear regression,n is the number of training examples,p is the number of features。
  - Problems encountered when imposing generalized least squared error in linear regression.
    - if n >> p,there's smaller error in least squared regression
    - if n ≈ p,it's easy to produce overfitting.
    - if n << p,least squared regression doesn't make sense about the result.

# The application of regularization

- **Application in machine learning—Ridge regression(cont.)**
  - The above problem can be shown by the variance and its bias of error,which can be modeld by the following diagram.

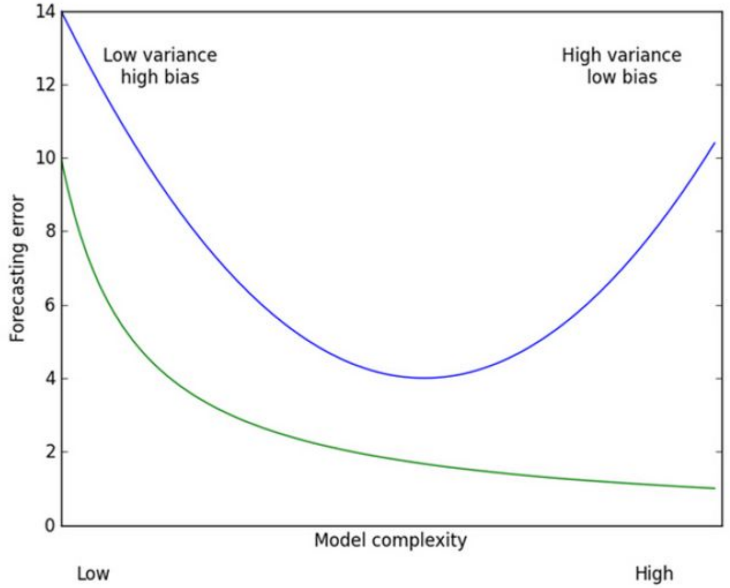



**Figure 8.8** **The bias variance tradeoff illustrated with test error and training error. The training error is the top curve, which has a minimum in the middle of the plot. In order to create the best forecasts, we should adjust our model complexity where the test error is at a minimum.**



  - So,we need to find the trade-off of variance and bias.

# The application of regularization

- **Application in machine learning—Ridge regression(cont.)**
  - With the complex model,the training examples are not enough to do regression.So,we need to do feature selection.
  - There are two solutions,one of which is ridge regression

$$\min \left\| w^T x - y \right\|_2^2 + \lambda \left\| w \right\|_2^2, \lambda > 0$$

$$\min \sum_{i=1}^{n} \left( y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2, \lambda > 0$$

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

# The application of regularization

- **Application in machine learning—Lasso regression**
  - Based on the previous problem,another solution is lasso regression

$$\min\left\|w^T x - y\right\|_2^2 + \lambda\left\|w\right\|_1, \lambda > 0$$

$$\min\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} w_j x_{ij}\right)^2 + \lambda\sum_{j=0}^{p}\left|w_j\right|, \lambda > 0$$

- There is no analytical solution.But provide sparsity for solution.

- **Application in machine learning**
  – Regularized linear regression
  – Regularized logistic regression

# The application of regularization

- **Application in multi-task learning-Regularization-based MLT**
  - MTL:learning multiple task simutanously so as to get better learning performance which comes from the related tasks.
  - Key point:The relatedness among tasks.Different methods modeling the relatedness produce different algorithms.
  - Regularization-based MTL:Take the relatedness among tasks as a prori of models then adding to the objective function as a regularizer.

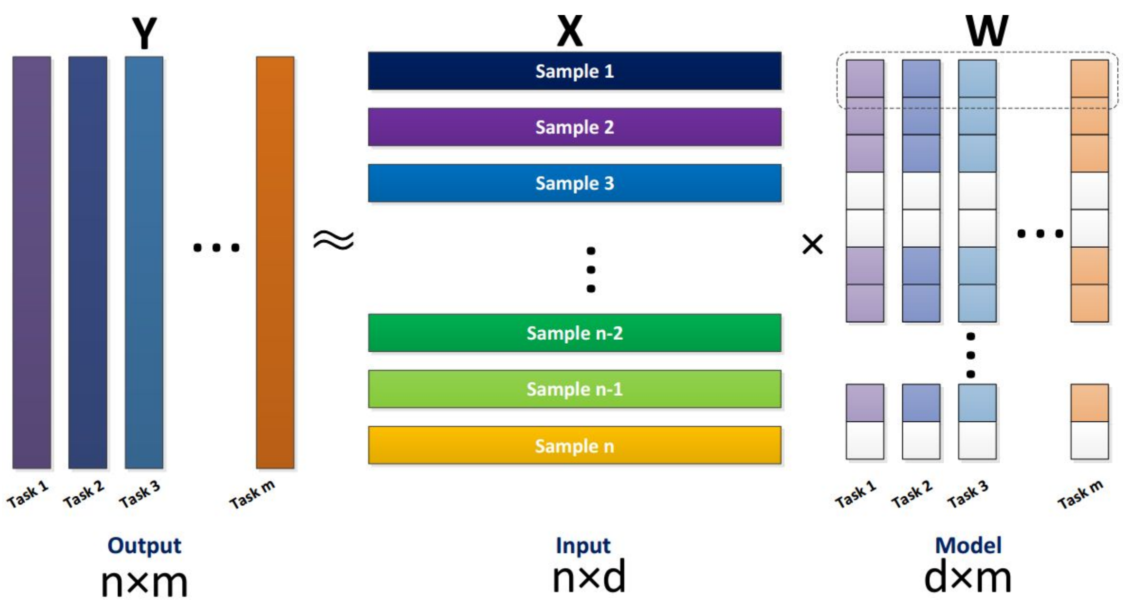
# The application of regularization

- **Application in multi-task learning-Regularization-based MLT(Examples)**
  - Mean-Regularized Multi-Task Learning(Evgeniou & Pontil, 2004 KDD)
    - Assumption: task parameter vectors of all tasks are close to each other.
    - Advantage: simple, intuitive, easy to implement
    - Disadvantage:may not hold in real applications.
    - Regularization:penalizes the deviation of each task from the mean

$$\min_{W} \frac{1}{2}\|XW - Y\|_F^2 + \lambda \sum_{i=1}^{m} \left\| W_i - \frac{1}{m} \sum_{s=1}^{m} W_s \right\|_2^2$$

# The application of regularization

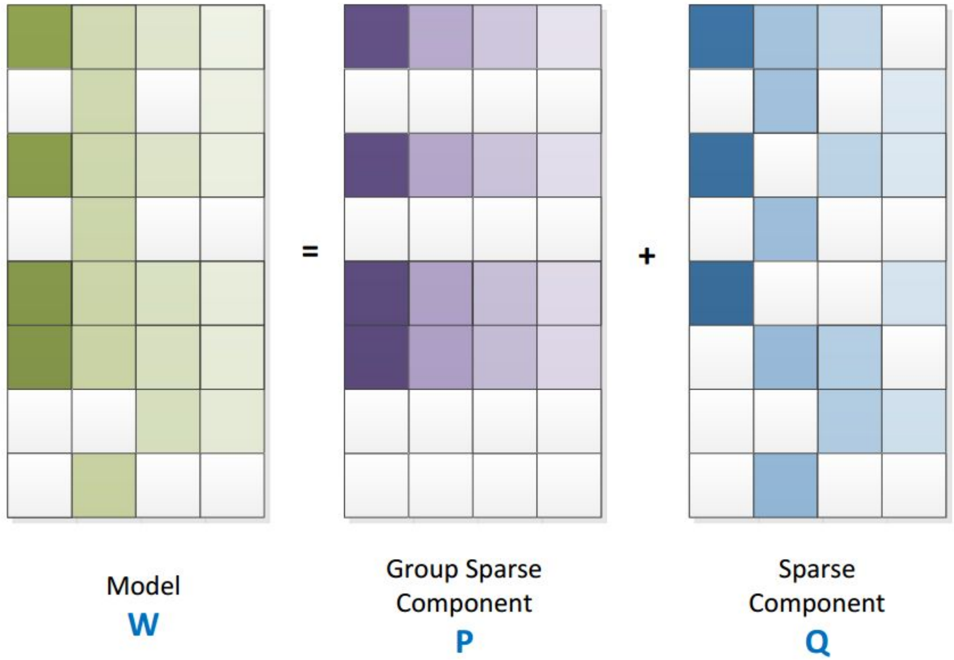
- **Application in multi-task learning-Regularization-based MLT(Examples)**
  - Multi-Task Learning with Joint Feature Learning(Obozinski et. al. 2009 Stat Comput, Liu et. al. 2010 Technical Report)
    - Using group sparsity: $l_1 / l_q -$ norm regularization $\|W\|_{1,q} = \sum_{i=1}^{d} \|w_i\|_q$
    - When q>1 we have group sparsity




$$\min_{W} \frac{1}{2}\|XW - Y\|_F^2 + \lambda\|W\|_{1,q}$$

# The application of regularization

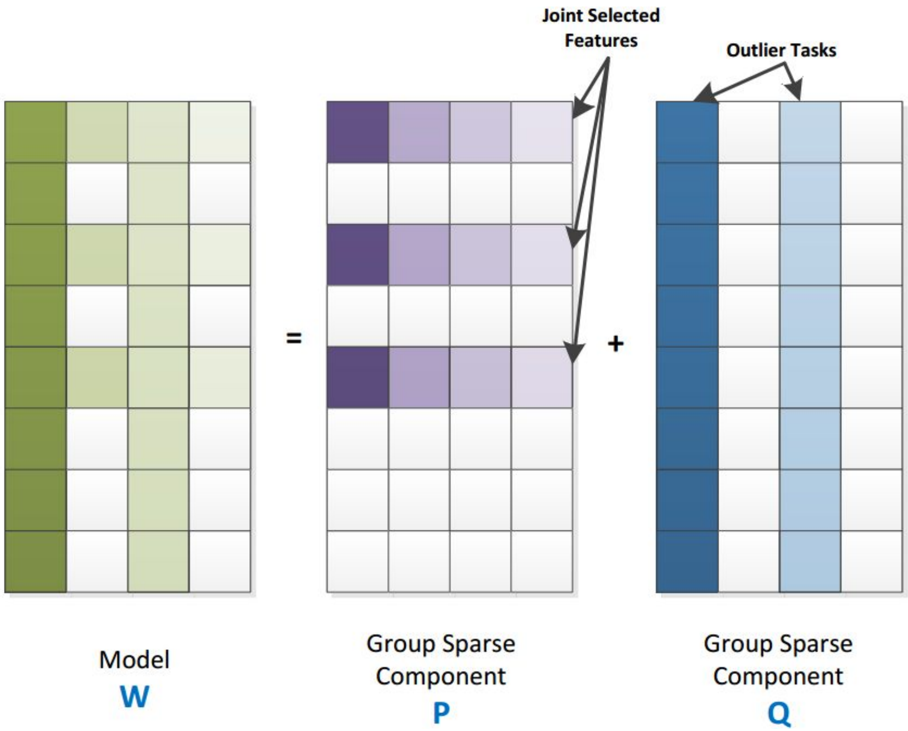
- **Application in multi-task learning-Regularization-based MLT(Examples)**
  - Dirty Model for Multi-Task Learning(Jalali et. al. 2010 NIPS)
    - In practical applications, it is too restrictive to constrain all tasks to share a single shared structure




$$\min_{P,Q} \frac{1}{2}\|X(P+Q)-Y\|_F^2 + \lambda_1\|P\|_{1,q} + \lambda_2\|Q\|_1$$

# The application of regularization

- **Application in multi-task learning-Regularization-based MLT(Examples)(outlier tasks)**
  - Robust Multi-Task Feature Learning(Gong et. al. 2012 Submitted)
    - Simultaneously captures a common set of features among relevant tasks and identifies outlier tasks




$$\min_{P,Q} \frac{1}{2}\left\|X(P+Q)-Y\right\|_F^2 + \lambda_1\left\|P\right\|_{1,q} + \lambda_2\left\|Q^T\right\|_{1,q}$$

# The application of regularization

- **Application in multi-task learning-Regularization-based MLT(Examples)**

AND SO ON…

# References


1. Convex Optimization.(Stephen Boyd,2006)
2. http://en.wikipedia.org/wiki/Regularization_(mathematics)
3. http://en.wikipedia.org/wiki/Tikhonov_regularization
4. The truncated SVD as a method for regularization.(Pet Christian Hansen,1986)
5. An iterative regularization method for total variation-based image restoration(S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin,2005,SIAM)
6. Bregman Iterative Algorithms for L1-Minimization with Applications to Compressed Sensing(Wotao Yin, Stanley Osher , Donald Goldfarb, and Jerome Darbon,2008,SIAM)
7. Spectral regularization(Lorenzo Rosasco,2009)
8. The L-curve and its use in the numerical treatment of inverse problems(P.C. Hansen)
9. Generalized cross validation as a method for choosing a good ridge parameter(Golub,Heath,Wahba,1979)
10. And so on...

# Acknowledgement

Thanks for your listening