CSI 436/536
Introduction to Machine Learning

Regularized LLSE

Professor Siwei Lyu
Computer Science
University at Albany, State University of New York

When over-fitting always happens

- For an under-complete least squares problem

$$\min_{\mathbf{w}} \left\| Y - X w \right\|^2$$

  - we have infinite number of optimal solutions
  - a similar situation: given two numbers a and b, and we know that $a \cdot b = 6$, determine a, b
- ill-posed problem has **too many** solutions, and overfitting always occurs in such problems
- In the context of machine learning, this means that there are multiple (sometimes infinite number of) models that can fit the data

Principle of Parsimony

- When multiple model can fit the data equally well, choose the ***simplest*** model
  - Also known as Occam's razor
- Complexity is measured for different model
  - For linear models like $f_w(x) = w^T g(x)$, the complexity is related with the L2 norm of w based on the learning theory
  - Simpler model may also mean w has many zeros, i.e., it is sparse
- The idea is to put a complexity term in the learning objective, together with the loss function

learning objective = **loss** + **regularizer**

ridge regression

- Solution: augment the least squares problem with an extra term as

$$\min_w ||y - X^T w||^2 + \lambda ||w||^2$$

  - this form of least squares problem is known as ***ridge regression***
  - $\lambda||\mathbf{w}||^2$ is known as the ***Tikhonov regularizer***
  - λ controls the trade-off between the error and the regularizer
- solution given by linear equation

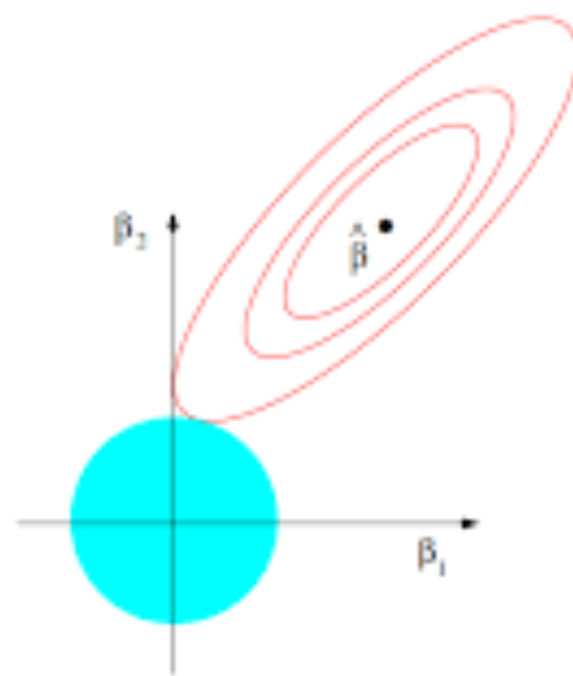
$$w = (XX^T + \lambda I)^{-1} Xy$$

Regularization

- Numerically, regularization stabilizes the solution to the normal equation $XX^T w = Xy$
- Recall that if X is not full ranked, then the correlation matrix XX^T is not invertible
- however, note that $XX^T + \lambda I$ is always invertible for $\lambda > 0$, so we get a ***stabilized*** solution to the normal equation
  - So this is also known as stabilization theory
- Unlike penalty method, where the penalty has no direct relation with the model parameter, regularizer is directly implied on the model parameter w
- We can also choose other types of regularizer, such as L1 norm

Equivalence with norm constraints

- Consider the following optimization problem $\min_w \|y - X^T w\|^2, \quad \text{s.t.} \quad \|w\|^2 = \rho$
- This problem is equivalent with the ridge regression objective function, where λ is the corresponding Lagrangian multiplier
- So regularization is essentially the same as putting constraints on the ill-posed problem
- Geometric picture
  - A parabola intersecting with a unit ball w.r.t. L2 norm for w



Choosing the weight on regularizer

- Small λ puts more weights on the error term, if λ = 0, reduces back to ordinary LLSE
- Large λ penalizes larger regularizer term, if λ = ∞, what is the solution to the problem

$$\min_w \|y - X^T w\|^2 + \lambda \|w\|^2$$

- In practice, it is a hyper-parameter that can be chosen with cross-validation
- λ needs to be larger than the smallest negative eigenvalue to work

LASSO

- Sometimes out of the infinite number of possible solutions to an ill-posed problem we prefer ***sparse*** ones
  - Sparseness is related with the number of zero elements in a vector
  - L0 norm is defined as the number of non-zero elements in a vector (is it a norm?)
- Sparsity is useful to identify a small set of factors that contribute to the observations
- In practice, it is usually replaced with L1 norm

sparsity regression (LASSO)

- we can also choose l_1 norm as the regularizer

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

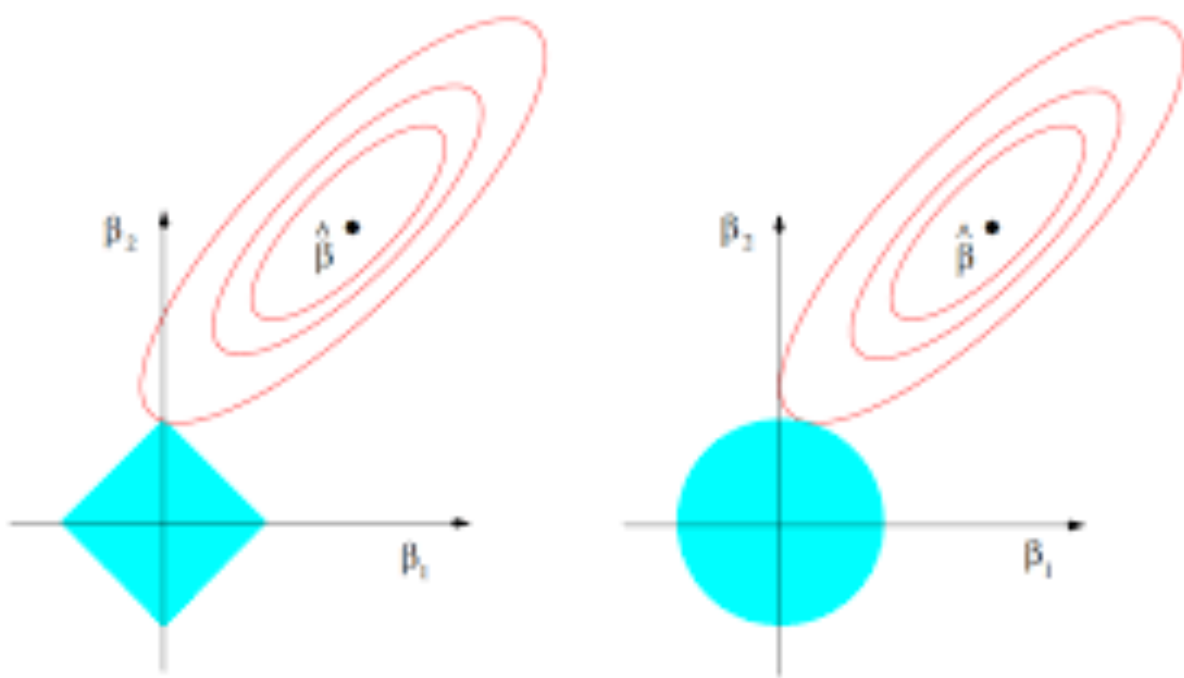
- however, this objective function is **not differentiable**
  - we cannot use the previous method (differentiation based) to solve it
  - many methods have been proposed, a very active research area and we will focus on one simple and efficient method

LASSO regression

- we can also choose l_1 norm as the regularizer

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

- using l_1 norm encourages **sparsity**



LASSO

Original objective function:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

Introduce new variable:

$$\begin{aligned} \min\nolimits_{\mathbf{x}} \quad & \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{z}\|_1 \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{z}. \end{aligned}$$

New objective function:

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\beta}{2}\|\mathbf{z} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{z}\|_1.$$

As $\beta \to \infty$, the solution to the objective function becomes the solution to the original problem.

- the new term is known as the ***penalty function***, i.e., it penalizes solutions that violate the equality constraint

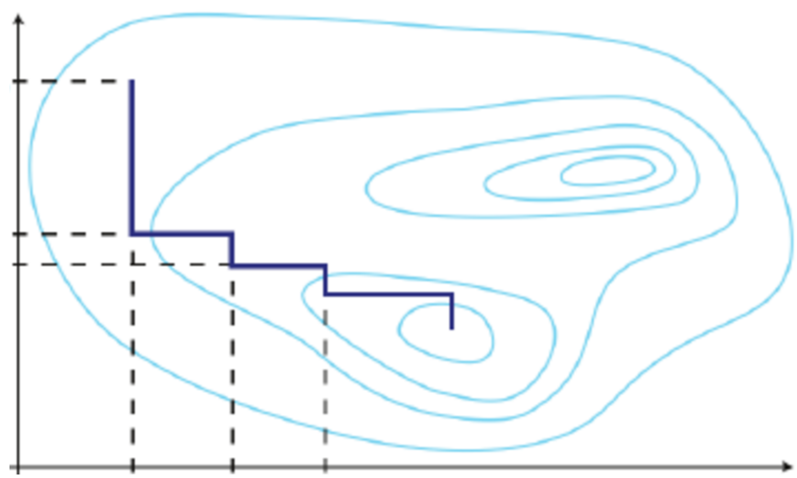
block coordinate descent

- alternating between optimizing x and z

Optimize

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\beta}{2}\|\mathbf{z} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{z}\|_1.$$

- algorithm is guaranteed to converge
- this is a convex problem so converge globally



Solving the sub-problems

- The x sub-problem

$$\min_x \frac{1}{2}\|y - Ax\|^2 + \frac{\beta}{2}\|x - z\|^2$$

- This can be solved with a ridge regression problem

$$\min_{\tilde{x}} \frac{1}{2}\|\tilde{y} - A\tilde{x}\|^2 + \frac{\beta}{2}\|\tilde{x}\|^2$$

where $\tilde{x} = x - z, \tilde{y} = y - Az$

- The z sub-problem

$$\min_z \frac{\beta}{2}\|z - x\|^2 + \lambda\|z\|_1$$

this problem separates in individual variables
solve individual 1D problems

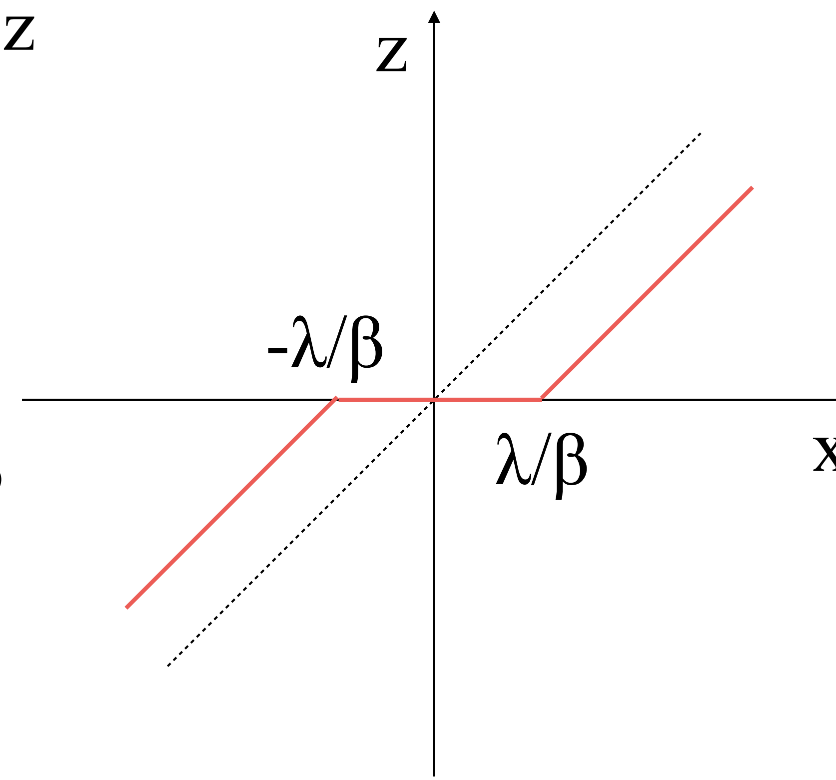
$$\min_z \frac{\beta}{2}(z - x)^2 + \lambda|z|$$

soft thresholding

- solve 1D problem

$$\min_z \frac{\beta}{2}(z-x)^2 + |z|$$

  - $z > 0$, minimize $0.5\beta(z-x)^2 + \lambda z$, $z = \max(x - \lambda/\beta, 0)$
  - $z < 0$, minimize $0.5\beta(z-x)^2 - \lambda z$, $z = \min(x+\lambda/\beta, 0)$
  - $z = 0$, do nothing
- $z = \max(x - \lambda/\beta,0) + \min(x+\lambda/\beta,0)$
- creating a "dead zone between $-\lambda/\beta$ and λ/β
- compare to hard thresholding



Overall algorithm

Optimize

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\beta}{2}\|\mathbf{z} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{z}\|_1 .$$

Repeat until convergence:

- Iterate between two steps until convergence:
  - Optimize **x**: ridge regression.
  - Optimize **z**: soft-thresholding.
- Increase β, e.g., $\beta \leftarrow 2\beta$.