ORIE 4741
Learning with Big Messy Data

Instructor: Madeleine Udell — Discussion: Regularization and Scaling
TA: Chengrun Yang — Nov 6, 2017

1 Regularized Underdetermined Least Squares

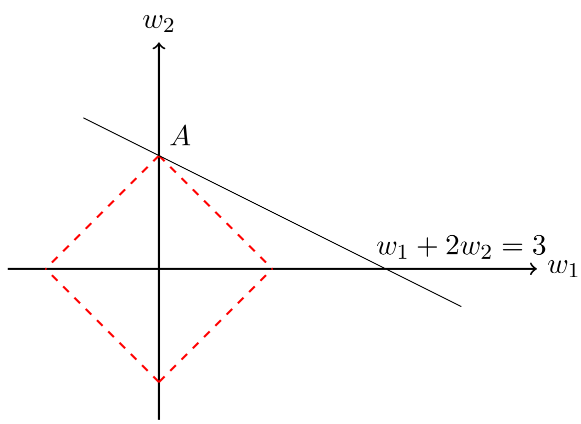
Question. Given $X = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, $y = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$, find solutions to

(i) $\text{argmin}_w ||y - Xw||_2^2 + \lambda ||w||_1$ (LASSO)
(ii) $\text{argmin}_w ||y - Xw||_2^2 + \lambda ||w||_2^2$ (ridge)

when

(i) $\lambda \to 0$
(ii) $\lambda \to +\infty$

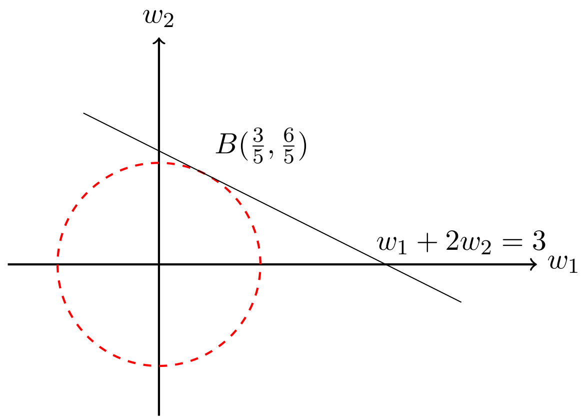
Solution. Notice that X does not have full column rank, thus the non-regularized least squares problem does not have a unique solution for w. Its solutions $w = (w_1, w_2)^T$ satisfy $w_1 + 2w_2 = 3$.



In the above coordinate system, the red dashed diamond is the 1-norm ball with $|w_1| + |w_2| = \frac{3}{2}$. It is tangent to the line $w_1 + 2w_2 = 3$ at $A(0, \frac{3}{2})$, meaning all other points on $w_1 + 2w_2 = 3$ have larger 1-norm than the w vector at point A. Thus A is the solution to the non-regularized setting of this problem with smallest 1-norm.

When $\lambda \to 0$, the regularized least squares problem tends to be non-regularized, hence the solution to LASSO problem tends to $(0, \frac{3}{2})$.

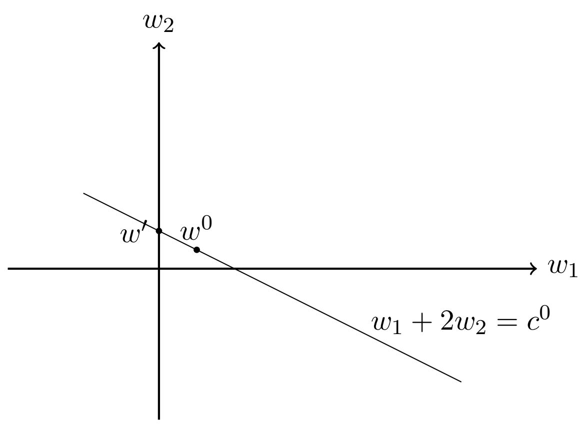
Following the same logic, for the ridge regression problem, we draw 2-norm ball with $w_1 + 2w_2 = 3$ as tangent line in the following figure. The solution to the ridge regression problem tends to $(\frac{3}{5}, \frac{6}{5})$.



When $\lambda \to +\infty$, the regularization term penalizes the objective funtion so much that even a w which is slightly farther from origin will give rise to an extremely large regularization term. Thus the solutions of both LASSO and ridge regression problem tend to $(0, 0)$.

Claim. *$\forall\lambda$, the solution to the above LASSO regression problem lies on the w_2 axis; the solution to the above LASSO regression problem lies on straight line $w_2 = 2w_1$.*

Proof. By contradiction, if the solution w^0 to the LASSO regression problem with a specific λ^0 does not lie on the w_2 axis. It must lie on a straight line $w_1 + 2w_2 = c^0$, in which c^0 is a specific constant.



Denote the intersection of this straight line and w_2 axis as w'. This point has the loss same value as w^0 and has smaller regularization value, which means w^0 is not the optimal point.

Proof for the ridge regression case follows the same logic. The straight line $w_1 + 2w_2 = c$ being tangent line of 2-norm ball indicates that the radius that connects origin with the tangent point lies on the straight line $w_2 = 2w_1$.

$\square$

In demo, change λ to have a look at how solutions to w change, and experimentally verify the above results.

2 Regularized Underdetermined Least Squares

Question: Given problem

$$\text{minimize} \quad \ell(w) + \lambda \cdot r(w)$$

Will variance be larger or smaller when λ becomes larger? How about bias?
Recall:

(i) $\text{var}(x) = \mathbb{E}_D[(g^D(x) - \bar{g}(x))^2]$

(ii) $\text{bias}^2(x) = (f(x) - \bar{g}(x))^2$

Answer: When λ increases, variance becomes smaller and bias becomes larger. See demo for animations.

3 Effect of Scaling when Outlier Exists

When "bad" outlier exists, scaling might make the result worse. See demo for animations.