1 point

1. **The best fit model of size 5 (i.e., with 5 features) always contains the set of features from best fit model of size 4.**

- ( ) True
- (•) False

---

1 point

2. **Given 20 potential features, how many models do you have to evaluate in the all subsets algorithm?**

1048576

---

1 point

3. **Given 20 potential features, how many models do you have to evaluate if you are running the forward stepwise greedy algorithm? Assume you run the algorithm all the way to the full feature set.**

210

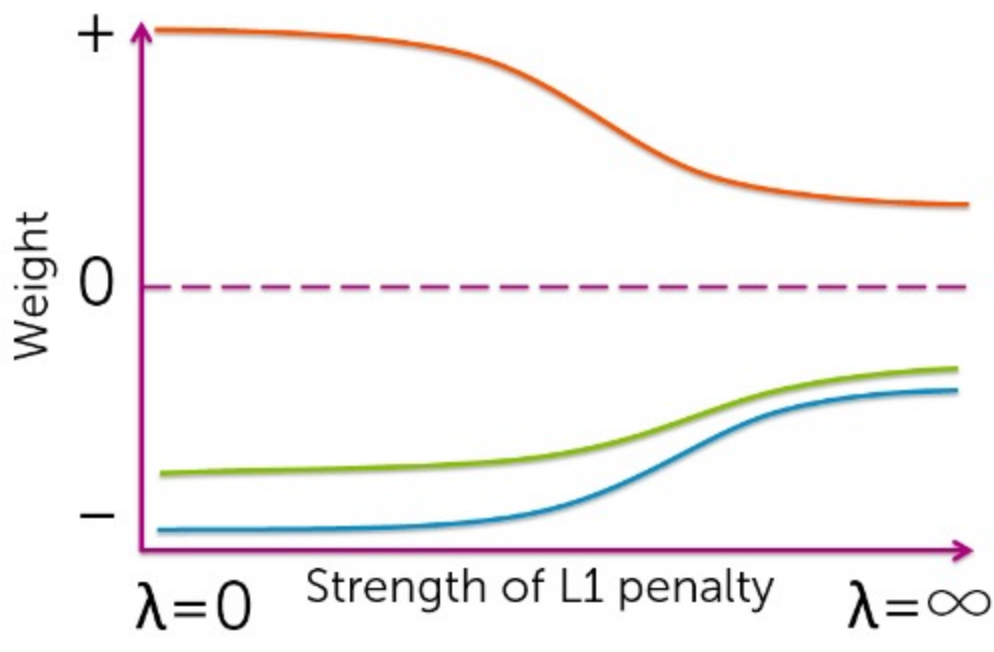
---

1 point

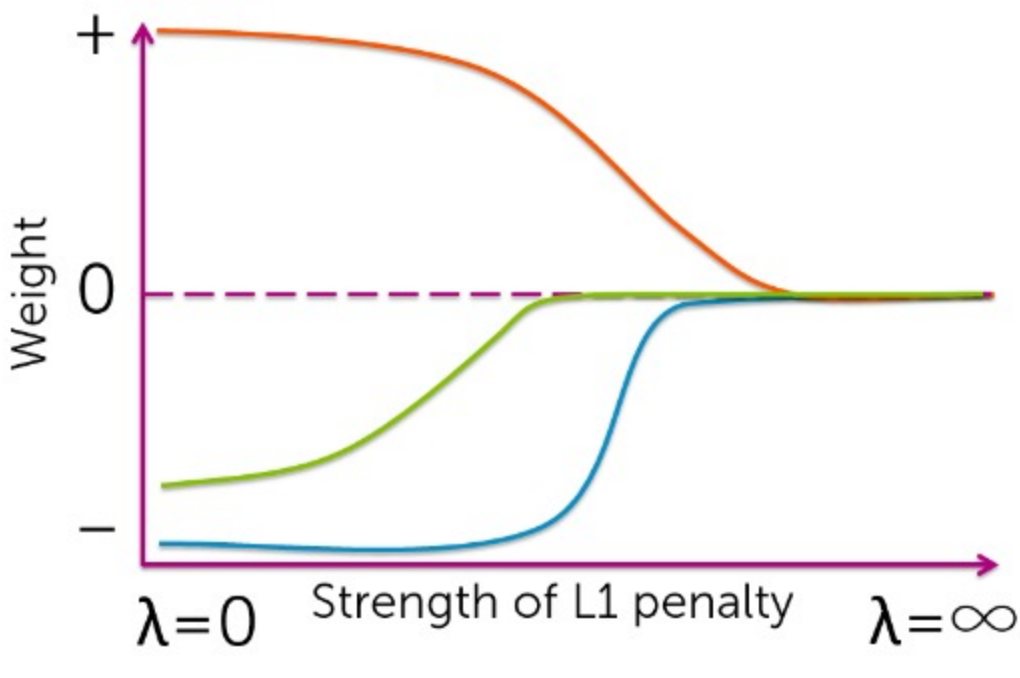
4. **Which of the plots could correspond to a lasso coefficient path? Select ALL that apply.**

Hint: notice $\lambda = \infty$ in the bottom right of the plots. How should coefficients behave eventually as λ goes to infinity?

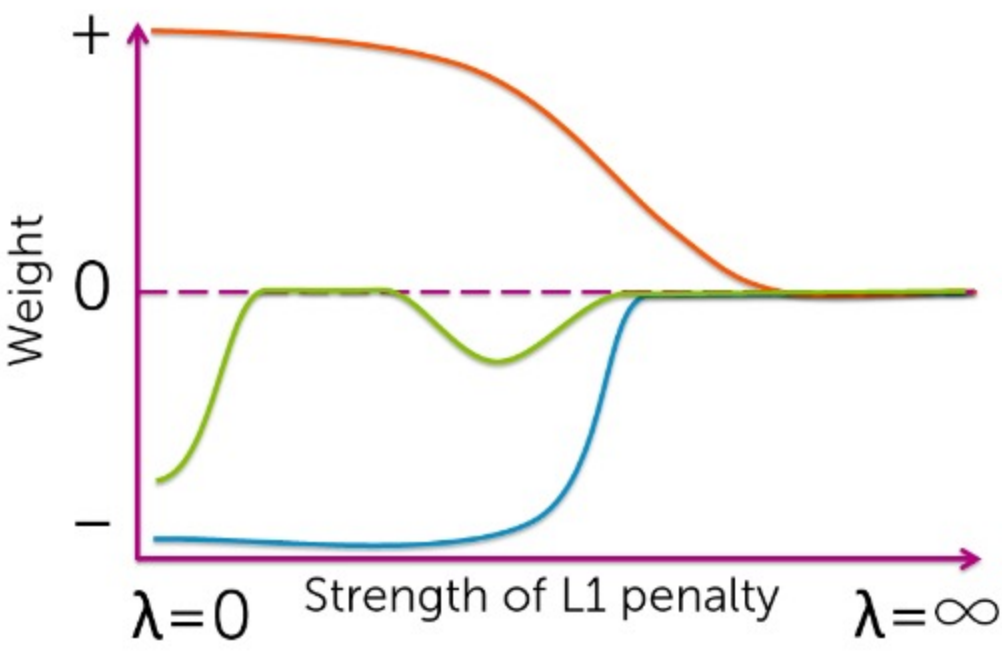
- [ ]



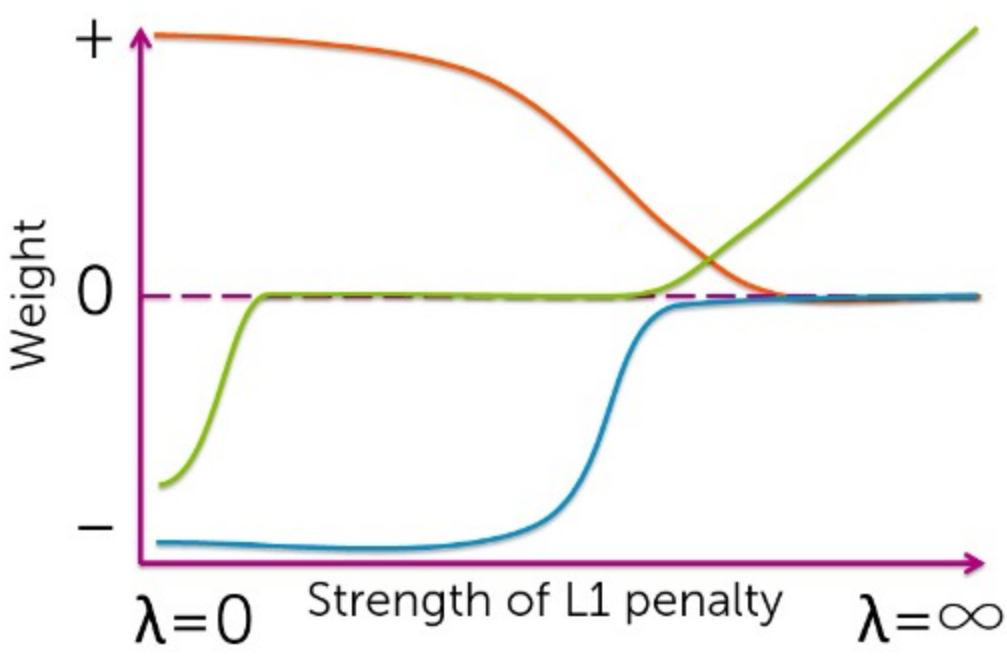
- [■]



- [ ]



- [ ]



---

1 point

5. **Which of the following statements about coordinate descent is true? (Select all that apply.)**

- [ ] A small enough step size should be chosen to guarantee convergence.
- [■] To test the convergence of coordinate descent, look at the size of the maximum step you take as you cycle through coordinates.
- [ ] Coordinate descent cannot be used to optimize the ordinary least squares objective.
- [ ] Coordinate descent is always less efficient than gradient descent, but is often easier to implement.

---

1 point

6. **Using normalized features, the ordinary least squares coordinate descent update for feature j has the form (with ρ_j defined as in the videos):**

- (•) $\hat{w}_j = \rho_j$
- ( ) $\hat{w}_j = (\rho_j)^2$
- ( ) $\hat{w}_j = \rho_j - \lambda$
- ( ) $\hat{w}_j = \rho_j/2 - \lambda$

---

1 point

7. **Using normalized features, the ridge regression coordinate descent update for feature j has the form (with ρ_j defined as in the videos):**

- ( ) $\hat{w}_j = \rho_j - \lambda$
- ( ) $\hat{w}_j = \rho_j/2 - \lambda$
- (•) $\hat{w}_j = \rho_j/(\lambda + 1)$
- ( ) $\hat{w}_j = \rho_j$