Regularization

- We want to keep the complexity of the model down, so we can penalize the cost function for large coefficients:

$$L(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\beta^T x^{(i)} - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \beta_j^2 = \frac{1}{m}\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \quad (L2 \text{ reg.})$$

$$L(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\text{———})^2 + \lambda \sum_{j=1}^{n} |\beta_j| = \frac{1}{m}\|X\beta - y\|_2^2 + \lambda \|\beta\|_1 \quad (L1 \text{ reg.})$$

(leads to sparsity in features)

From a statistical pov, minimizing this is equivalent to solving for the MAP estimates w/ a Gaussian and Laplacian prior on β respectively:

$$P(\beta|X,y) \overset{\beta}{\propto} P(y|X,\beta)P(\beta)$$

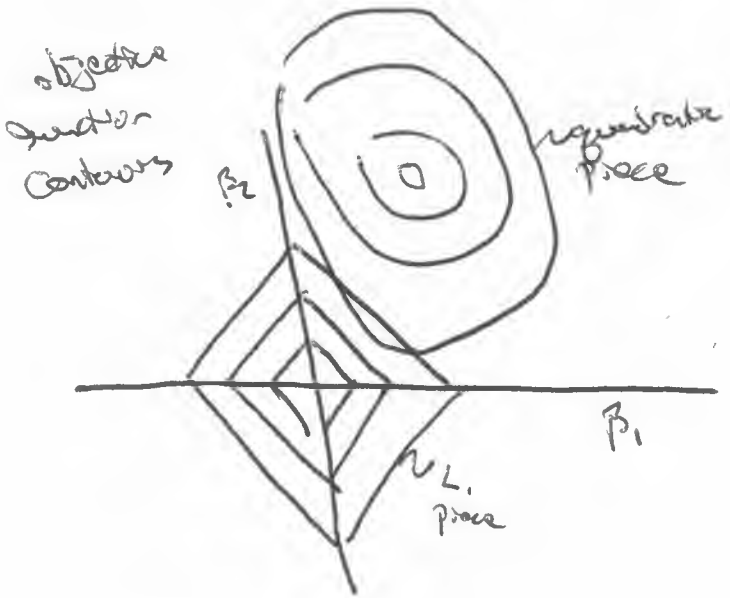
$$\Rightarrow \hat{\beta}_{MAP} = \underset{\beta \in \mathbb{R}^n}{\arg\max} \{ \ln P(y|X,\beta) + \ln P(\beta) \}$$

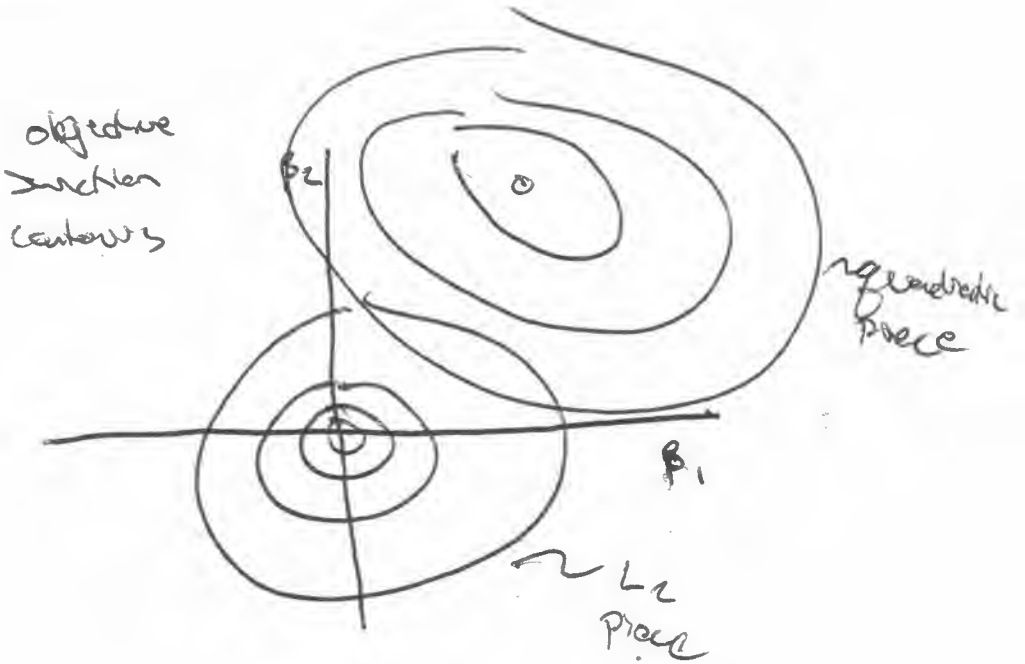
$\ln P(y|X,\beta) \propto \|X\beta - y\|^2$ as found before

$\ln P(\beta) \propto \|\beta\|_1 \lambda$ for a laplacian, or $\|\beta\|_2^2 \lambda$ " " gaussian

- the fact that most of the prior mass is centered around zero is why the posterior leads to a soln. w/ smaller β values

- Qualitatively, regularization affects the minimization problem as follows. Our goal is the find the minimum of the sum of the 2 pieces.

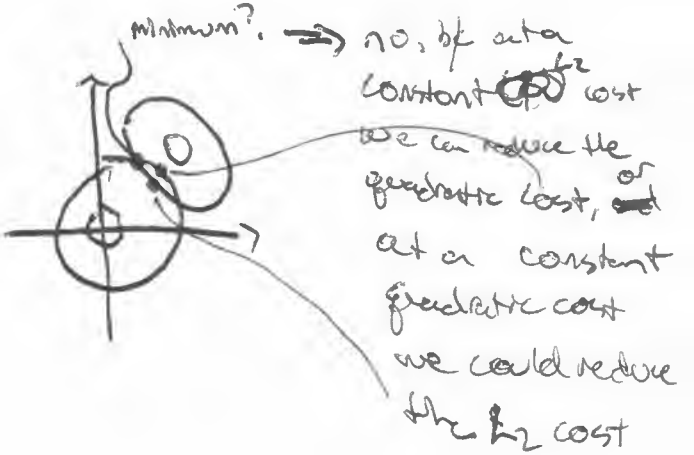




- minimum occurs where the 2 contours are $\parallel$, otherwise.

qualitatively

- this is why L_1 promotes sparsity, b/c there are infinitely many more ways at the vertices for the 2 pieces to be tangent.



- This can also be turned into a constrained minimization problem, w/ some Lagrangian theory, which can gain further insight into the problem.