RIDGE REGRESSION / L2 PENALTY

Intuitive measure of model complexity is deviation from 0-origin; coeffs then have no or a weak effect. So we measure $J(\theta)$ through a vector norm, shrinking coeffs closer to 0.

$$\begin{aligned}\hat{\theta}_{\text{ridge}} &= \arg\min_{\theta} \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \mathbf{x}^{(i)}\right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \\ &= \arg\min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_2^2\end{aligned}$$

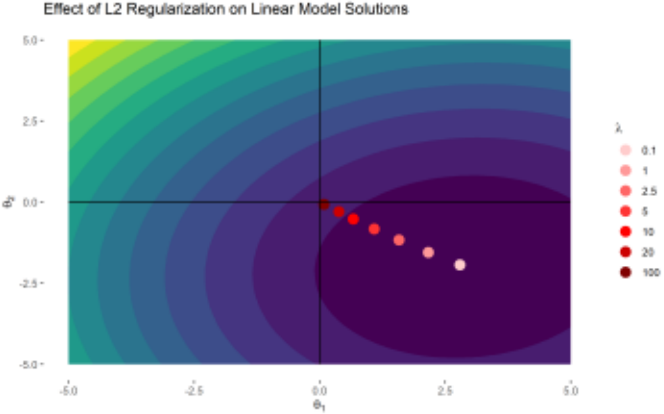
Can still analytically solve this:

$$\hat{\theta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda \boldsymbol{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Name: We add pos. entries along the diagonal "ridge" of $\mathbf{X}^T\mathbf{X}$

RIDGE REGRESSION / L2 PENALTY / 2

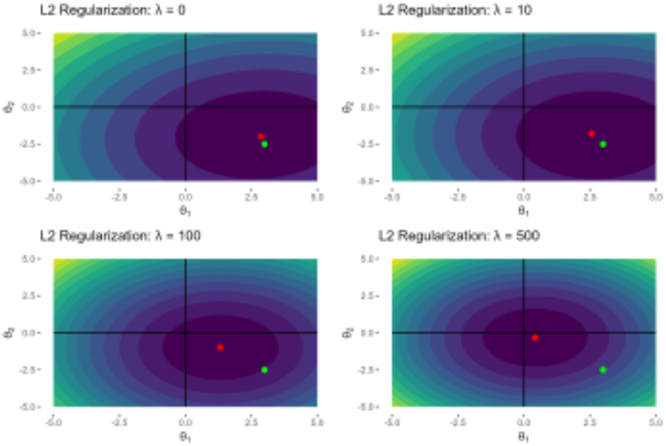
Let $y = 3x_1 - 2x_2 + \epsilon, \epsilon \sim N(0,1)$. The true minimizer is $\theta^* = (3, -2)^T$, with $\hat{\theta}_{\text{ridge}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$.



With increasing regularization, $\hat{\theta}_{ridge}$ is pulled back to the origin (contour lines show unregularized objective).

RIDGE REGRESSION / L2 PENALTY / 3

Contours of regularized objective for different λ values.
$\hat{\theta}_{\text{ridge}} = \arg\min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda\|\theta\|^2$.



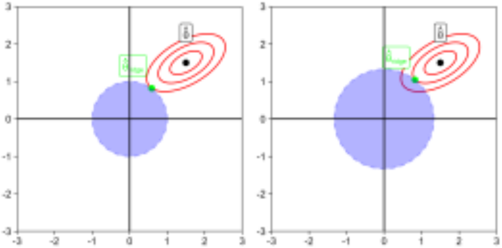
Green = true coefs of the DGP and red = ridge solution.

RIDGE REGRESSION / L2 PENALTY / 4

We understand the geometry of these 2 mixed components in our regularized risk objective much better, if we formulate the optimization as a constrained problem (see this as Lagrange multipliers in reverse).

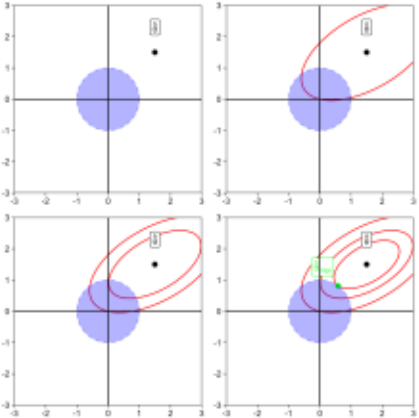
$$\min_{\theta} \quad \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \theta\right)\right)^2$$

$$\text{s.t.} \quad \|\theta\|_2^2 \leq t$$

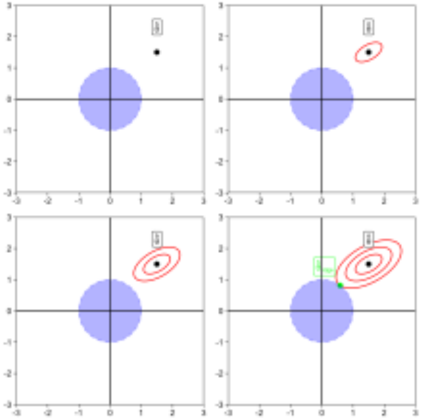


NB: There is a bijective relationship between λ and t: $\lambda \uparrow \Rightarrow t \downarrow$ and vice versa.

RIDGE REGRESSION / L2 PENALTY / 5

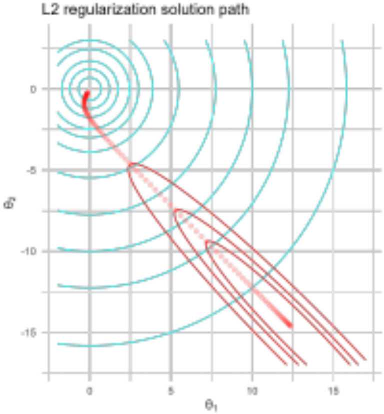


- Inside constraints perspective: From origin, jump from contour line to contour line (better) until you become infeasible, stop before.
- We still optimize the $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$, but cannot leave a ball around the origin.
- $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ grows monotonically if we move away from $\hat{\theta}$ (elliptic contours).
- Solution path moves from origin to border of feasible region with minimal L_2 distance.



- Outside constraints perspective:
  From $\hat{\theta}$, jump from contour line to contour line (worse) until you become feasible, stop then.
- So our new optimum will lie on the boundary of that ball.
- Solution path moves from unregularized estimate to feasible region of regularized objective with minimal L_2 distance.

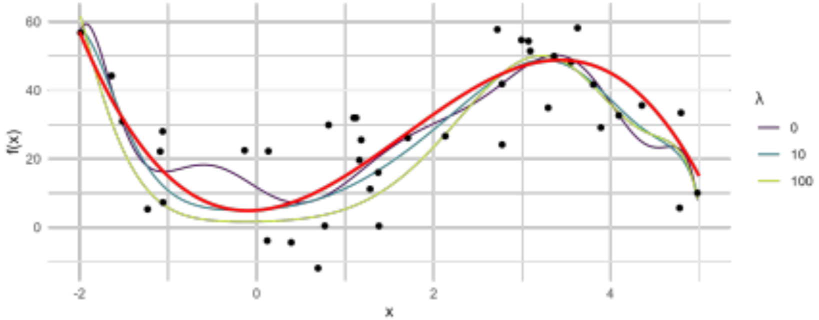
RIDGE REGRESSION / L2 PENALTY



- Here we can see entire solution path for ridge regression
- Cyan contours indicate feasible regions induced by different λs
- Red contour lines indicate different levels of the unreg. objective
- Ridge solution (red points) gets pulled toward origin for increasing λ

EXAMPLE: POLYNOMIAL RIDGE REGRESSION / 2

With an $L2$ penalty we can now select d "too large" but regularize our model by shrinking its coefficients. Otherwise we have to optimize over the discrete d.



λ	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
0.00	12.00	-16.00	4.80	23.00	-5.40	-9.30	4.20	0.53	-0.63	0.13	-0.01
10.00	5.20	1.30	3.70	0.69	1.90	-2.00	0.47	0.20	-0.14	0.03	-0.00
100.00	1.70	0.46	1.80	0.25	1.80	-0.94	0.34	-0.01	-0.06	0.02	-0.00