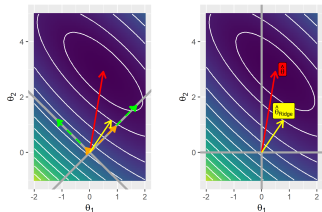
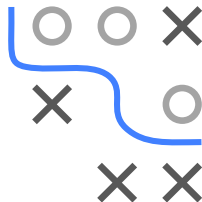


Introduction to Machine Learning

Regularization

Geometry of L2 Regularization



Learning goals

- Approximate transformation of unregularized minimizer to regularized
- Principal components of Hessian influence where parameters are decayed

GEOMETRIC ANALYSIS OF L_2 REGULARIZATION

Quadratic Taylor approx of the unregularized objective $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ around its minimizer $\hat{\boldsymbol{\theta}}$:

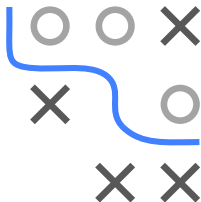
$$\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) + \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) \cdot (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

where \mathbf{H} is the Hessian of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$

We notice:

- First-order term is 0, because gradient must be 0 at minimizer
- \mathbf{H} is positive semidefinite, because we are at the minimizer

$$\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$



GEOMETRIC ANALYSIS OF L_2 REGULARIZATION

/ 2

The minimum of $\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$ occurs where $\nabla_{\boldsymbol{\theta}} \tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ is 0.
Now we L_2 -regularize $\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$, such that

$$\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) = \tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$$

and solve this approximation of \mathcal{R}_{reg} for the minimizer $\hat{\boldsymbol{\theta}}_{\text{ridge}}$:

$$\nabla_{\boldsymbol{\theta}} \tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) = 0$$

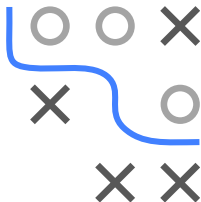
$$\lambda \boldsymbol{\theta} + \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 0$$

$$(\mathbf{H} + \lambda \mathbf{I})\boldsymbol{\theta} = \mathbf{H}\hat{\boldsymbol{\theta}}$$

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

We see: minimizer of L_2 -regularized version is (approximately!)
transformation of minimizer of the unpenalized version.

Doesn't matter whether the model is an LM – or something else!



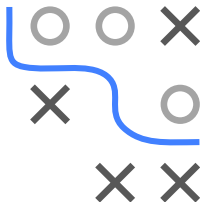
GEOMETRIC ANALYSIS OF L_2 REGULARIZATION

/ 3

- As λ approaches 0, the regularized solution $\hat{\theta}_{\text{ridge}}$ approaches $\hat{\theta}$. What happens as λ grows?
- Because \mathbf{H} is a real symmetric matrix, it can be decomposed as $\mathbf{H} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top$, where $\mathbf{\Sigma}$ is a diagonal matrix of eigenvalues and \mathbf{Q} is an orthonormal basis of eigenvectors.
- Rewriting the transformation formula with this:

$$\begin{aligned}\hat{\theta}_{\text{ridge}} &= \left(\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top + \lambda \mathbf{I} \right)^{-1} \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top \hat{\theta} \\ &= \left[\mathbf{Q}(\mathbf{\Sigma} + \lambda \mathbf{I})\mathbf{Q}^\top \right]^{-1} \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top \hat{\theta} \\ &= \mathbf{Q}(\mathbf{\Sigma} + \lambda \mathbf{I})^{-1} \mathbf{\Sigma}\mathbf{Q}^\top \hat{\theta}\end{aligned}$$

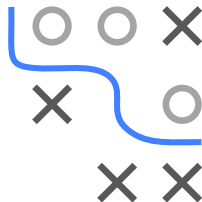
- So: We rescale $\hat{\theta}$ along axes defined by eigenvectors of \mathbf{H} . The component of $\hat{\theta}$ that is associated with the j -th eigenvector of \mathbf{H} is rescaled by factor of $\frac{\sigma_j}{\sigma_j + \lambda}$, where σ_j is eigenvalue.



GEOMETRIC ANALYSIS OF L_2 REGULARIZATION

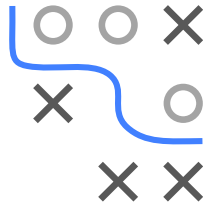
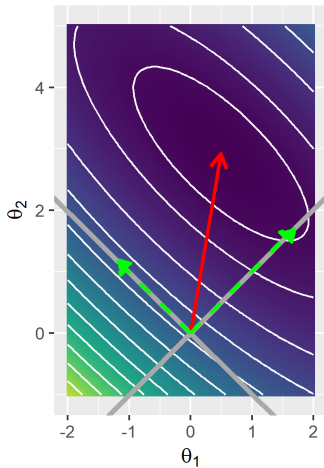
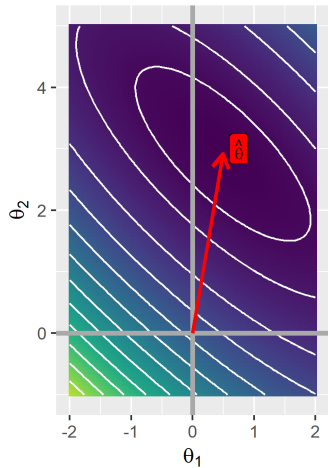
/ 4

First, $\hat{\theta}$ is rotated by \mathbf{Q}^\top , which we can interpret as projection of $\hat{\theta}$ on rotated coord system defined by principal directions of \mathbf{H} :



GEOMETRIC ANALYSIS OF L_2 REGULARIZATION

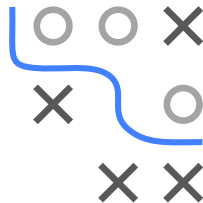
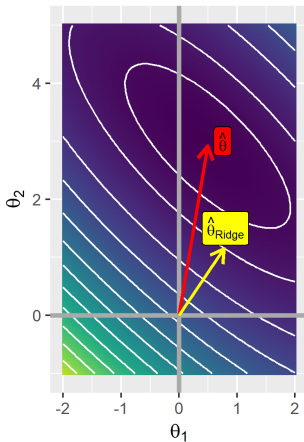
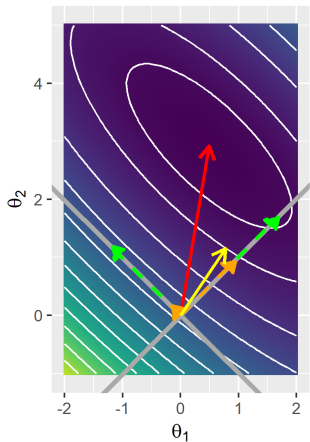
/ 5



GEOMETRIC ANALYSIS OF L_2 REGULARIZATION

/ 6

j -th (new) axis is rescaled by $\frac{\sigma_j}{\sigma_j + \lambda}$ before we rotate back.



GEOMETRIC ANALYSIS OF L_2 REGULARIZATION

/ 7

- Decay: $\frac{\sigma_j}{\sigma_j + \lambda}$
- Along directions where eigenvals of \mathbf{H} are relatively large, e.g., $\sigma_j \gg \lambda$, effect of regularization is small.
- Components / directions with $\sigma_j \ll \lambda$ are strongly shrunk.
- So: Directions along which parameters contribute strongly to objective are preserved relatively intact.
- In other directions, small eigenvalue of Hessian means that moving in this direction will not decrease objective much. For such unimportant directions, corresponding components of θ are decayed away.

