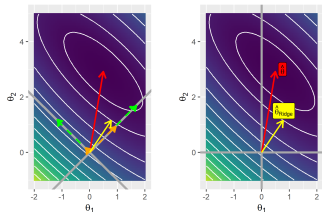# Introduction to Machine Learning

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 I**

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

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

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

We notice:

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

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

GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION I

The minimum of $\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$ occurs where $\nabla_{\boldsymbol{\theta}}\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta} - \hat{\theta})$ 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 $\mathcal{R}_{\text{reg}}$ for the minimizer $\hat{\boldsymbol{\theta}}_{\text{ridge}}$:
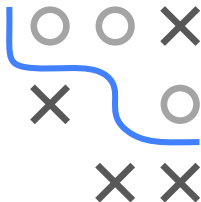
$$\nabla_{\boldsymbol{\theta}}\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) = 0$$
$$\lambda\boldsymbol{\theta} + \boldsymbol{H}(\boldsymbol{\theta} - \hat{\theta}) = 0$$
$$(\boldsymbol{H} + \lambda\boldsymbol{I})\boldsymbol{\theta} = \boldsymbol{H}\hat{\theta}$$
$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = (\boldsymbol{H} + \lambda\boldsymbol{I})^{-1}\boldsymbol{H}\hat{\theta}$$

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 $L2$ REGULARIZATION I

- As λ approaches 0, the regularized solution $\hat{\theta}_{\text{ridge}}$ approaches $\hat{\theta}$. What happens as λ grows?

- Because H is a real symmetric matrix, it can be decomposed as $H = Q\Sigma Q^\top$, where Σ is a diagonal matrix of eigenvalues and Q is an orthonormal basis of eigenvectors.

- Rewriting the transformation formula with this:

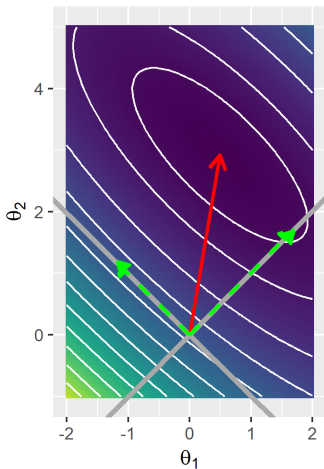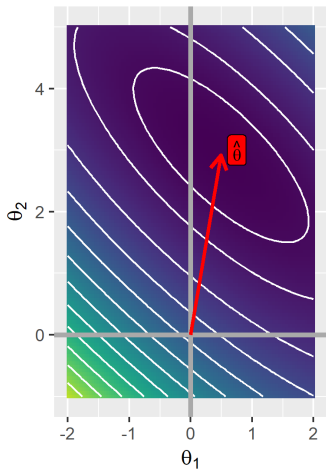$$\hat{\theta}_{\text{ridge}} = \left(Q\Sigma Q^\top + \lambda I \right)^{-1} Q\Sigma Q^\top \hat{\theta}$$
$$= \left[Q(\Sigma + \lambda I)Q^\top \right]^{-1} Q\Sigma Q^\top \hat{\theta}$$
$$= Q(\Sigma + \lambda I)^{-1}\Sigma Q^\top \hat{\theta}$$

- So: We rescale $\hat{\theta}$ along axes defined by eigenvectors of H. The component of $\hat{\theta}$ that is associated with the j-th eigenvector of H is rescaled by factor of $\frac{\sigma_j}{\sigma_j + \lambda}$, where σ_j is eigenvalue.
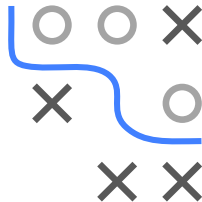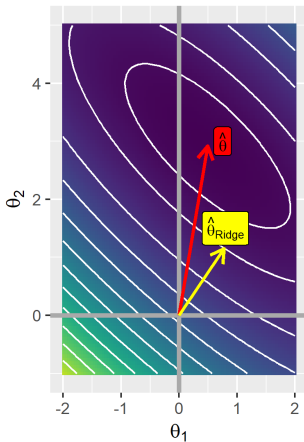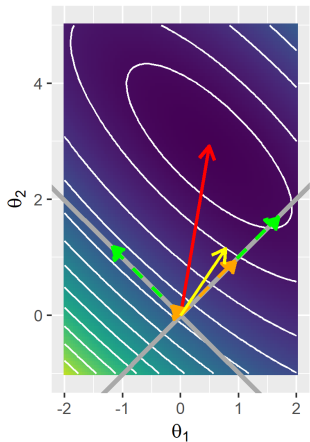
GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION I

First, $\hat{\theta}$ is rotated by $\boldsymbol{Q}^\top$, which we can interpret as projection of $\hat{\theta}$ on rotated coord system defined by principal directions of $\boldsymbol{H}$:
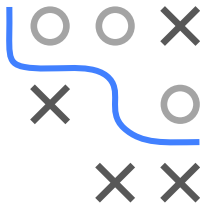
GEOMETRIC ANALYSIS OF *L*2 REGULARIZATION I

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

GEOMETRIC ANALYSIS OF L2 REGULARIZATION I

- Decay: $\frac{\sigma_j}{\sigma_j + \lambda}$

- Along directions where eigenvals of **H** are relatively large, e.g., $\sigma_j >> \lambda$, effect of regularization is small.

- Components / directions with $\sigma_j << \lambda$ are strongly shrunken.

- 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 $\boldsymbol{\theta}$ are decayed away.