Lecture 14 notes:
MAP inference for linear-Gaussian model
Tues, Oct 20

1 Regularization with priors: quick refresher

1.1 MAP inference

We have previously discussed the idea of adding a prior (or equivalently, a penalty) to regularize weights in a GLM or other regression model.

That is, we seek the maximum a posteriori (MAP) estimate:

$$\hat{w}_{\text{map}} = \arg \max_{\vec{w}} \log p(\vec{w}|Y, X, \theta) = \arg \max_{\vec{w}} \left[ \log p(Y|X, \vec{w}) + \log p(\vec{w}|\theta) \right].$$ (1)

Maximizing the log-posterior is equivalent to maximizing the rightmost expression (the log-likelihood plus log-prior) because the denominator in Bayes’ rule is a constant does not depend on $\vec{w}$.

1.2 Ridge and Smoothing Priors

The priors we have considered so far are the ridge prior, $\vec{w} \sim \mathcal{N}(0, \frac{1}{\lambda} I)$, which has log-prior (or penalty) equal to

$$\text{penalty} = \log p(\vec{w}|\lambda) = -\frac{1}{2} \lambda \vec{w}^T \vec{w}, \quad (2)$$

and the graph Laplacian smoothing prior, which has log-prior (or penalty) equal to

$$\text{penalty} = \log p(\vec{w}|\lambda) = -\frac{1}{2} \lambda \vec{w}^T L \vec{w}, \quad (3)$$

where $L$ is the graph Laplacian.

Here we have written the prior so that $\lambda$ is equal to the inverse variance of the prior. Thus $\lambda = 0$ corresponds to a prior with infinitely broad variance (in which case we have no regularization, so the MAP estimate is equal to the maximum likelihood estimate). At the other extreme, $\lambda = \infty$ corresponds to an infinitely strong penalty, or a prior that is infinitely concentrated at zero (that is, a Dirac delta function).

Note we previously use $\theta$ to parametrize the prior, where $\theta$ denoted the prior variance but we switch here to $\lambda = 1/\theta$ simply because this notation is more common!
2 MAP inference for linear-Gaussian model

For the linear-Gaussian model we can compute the MAP estimate in closed form. We have the model now specified by

\[ \vec{w} \sim \mathcal{N}(0, C) \]  
\[ Y|X, \vec{w} \sim \mathcal{N}(X\vec{w}, \sigma^2 I), \]

where \( C \) is a prior covariance matrix (parametrized by \( \theta \)) and \( \sigma^2 \) is the variance of the additive noise in \( Y \).

The log-posterior is given by:

\[ \log p(Y|X, \vec{w}) + \log p(\vec{w}|\theta) + \text{const} = \frac{1}{2\sigma^2} (Y - X\vec{w})^\top (Y - X\vec{w}) - \frac{1}{2} \vec{w} C^{-1} \vec{w} + \text{const} \]
\[ = -\frac{1}{2} \vec{w}^\top \left( \frac{1}{\sigma^2} X^\top X \right) \vec{w} - \frac{1}{2} \vec{w} C^{-1} \vec{w} - \vec{w} \left( \frac{1}{\sigma^2} X^\top Y \right) + \text{const} \]
\[ = -\frac{1}{2} \vec{w}^\top \left( \frac{1}{\sigma^2} X^\top X + C^{-1} \right) \vec{w} - \vec{w} \left( \frac{1}{\sigma^2} X^\top Y \right) + \text{const}, \]

where \( \text{const} \) is a catch-all that collects terms that do not contain \( \vec{w} \).

Differentiating with respect to \( \vec{w} \), setting to zero and solving gives:

\[ \hat{\vec{w}}_{\text{map}} = (X^\top X + \sigma^2 C^{-1})^{-1} X^\top Y \]

In the case of ridge regression, where prior covariance \( C = \theta I \), this corresponds to

\[ \hat{\vec{w}}_{\text{ridge}} = (X^\top X + \frac{\sigma^2}{\theta} I)^{-1} X^\top Y. \]

Thus the stimulus covariance \( X^\top X \) is added to a diagonal matrix with diagonal value \( \sigma^2/\theta \), the ratio of noise variance to prior variance. The larger this value, the more the coefficients of \( \hat{\vec{w}}_{\text{map}} \) are shrunk towards zero.

Note: although we have this handy closed-form solution for the linear-Gaussian model, for the other GLMs we’ve considered (Bernoulli and Poisson GLM), we must find the maximum via numerical optimization.

3 The key question: how to set \( \lambda \)?

The question we have (so far) neglected is: how do we choose the strength of the regularizer, or in Bayesian terms, the width of the prior? We will consider two possible approaches: (1) Cross-validation; (2) maximum marginal likelihood / evidence optimization.

To be continued....