

Hand In 1 – Regularization

TIF345 – ADVANCED SIMULATION AND MACHINE LEARNING
CHALMERS UNIVERSITY OF TECHNOLOGY

OSCAR STOMMENDAL[†], FALL 2025

In this assignment, we consider an i.i.d. (independent and identically distributed) data likelihood with a linear model design matrix Φ and parameters θ ,

$$p(\mathcal{D}_i|\theta) \sim \mathcal{N}(\mathcal{D}_i | [\Phi\theta]_i, \sigma^2). \quad (1)$$

Maximum a Posteriori Estimator with Gaussian Prior

In this case, we consider a Gaussian prior on the parameters θ ,

$$p(\theta) \sim \prod_{i=1}^{N_p} \mathcal{N}(\theta_i | 0, \sigma_0^2). \quad (2)$$

The maximum a posteriori (MAP) estimator is given by

$$\theta_{MAP} = \arg \max_{\theta} p(\theta | \mathcal{D}) = \arg \max_{\theta} p(\mathcal{D} | \theta) p(\theta). \quad (3)$$

Here, we have used Bayes' theorem and dropped the evidence term $p(\mathcal{D})$ since it does not depend on θ . Taking the logarithm of Eq. 3, we have

$$\theta_{MAP} = \arg \max_{\theta} \log p(\mathcal{D} | \theta) + \log p(\theta). \quad (4)$$

For Gaussian distributions, i.e. the likelihood $p(\mathcal{D} | \theta)$ and the prior $p(\theta)$, we have that

$$p(\mathbf{x} | \mu) \sim \mathcal{N}(\mathbf{x} | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{x} - \mu)^T(\mathbf{x} - \mu)}{2\sigma^2}\right), \quad (5)$$

for i.i.d. data with variance σ^2 . Applying this to the likelihood and prior, we have

$$p(\mathcal{D}, \theta) \sim \mathcal{N}(\mathcal{D} | \Phi\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathcal{D} - \Phi\theta)^T(\mathcal{D} - \Phi\theta)}{2\sigma^2}\right), \quad (6)$$

$$p(\theta) \sim \prod_{i=1}^{N_p} \mathcal{N}(\theta_i | 0, \sigma_0^2) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^{N_p} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^{N_p} \theta_i^2\right). \quad (7)$$

Taking the logarithm of these expressions, we have

$$\theta_{MAP} = \arg \max_{\theta} \left[-\frac{1}{2\sigma^2} (\mathcal{D} - \Phi\theta)^T (\mathcal{D} - \Phi\theta) - \frac{1}{2\sigma_0^2} \sum_{i=1}^{N_p} \theta_i^2 + C \right], \quad (8)$$

where C is a constant that does not depend on θ coming from the normalization constants. Multiplying by $2\sigma^2$, dropping the constant term, and converting the maximization to a minimization by multiplying by -1 , we have

$$\theta_{MAP} = \arg \min_{\theta} \left[(\mathcal{D} - \Phi\theta)^T (\mathcal{D} - \Phi\theta) + \frac{\sigma^2}{\sigma_0^2} \sum_{i=1}^{N_p} \theta_i^2 \right]. \quad (9)$$

[†]oscarsto@chalmers.se

This is the Ridge regression estimator with regularization parameter $\lambda = \frac{\sigma^2}{\sigma_0^2}$. A large data variance σ^2 means that the data is noisy, and we should therefore trust it less, leading to a larger λ (yielding smaller estimated parameters). Conversely, a large prior variance σ_0^2 means that we are less certain about our prior knowledge of $\boldsymbol{\theta}$, leading to a smaller λ , allowing larger estimated parameters. So, λ can be interpreted as a penalty term that balances the fit to the data, influenced by our confidence in the data and prior.

Maximum a Posteriori Estimator with Laplace Prior

Now, we instead consider a Laplace prior on the parameters $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta}) \sim \prod_{i=1}^{N_p} \mathcal{L}(\theta_i | 0, \sigma_0) = \left(\frac{1}{2\sigma_0} \right)^{N_p} \exp \left(-\frac{1}{\sigma_0} \sum_{i=1}^{N_p} |\theta_i| \right). \quad (10)$$

The MAP estimator is again given by Eq. 4. For the likelihood, we have the same expression as in Eq. 6. Taking the logarithm of the likelihood and prior, we have

$$\boldsymbol{\theta}_{MAP} = \arg \max_{\boldsymbol{\theta}} \left[-\frac{1}{2\sigma^2} (\mathcal{D} - \Phi\boldsymbol{\theta})^T (\mathcal{D} - \Phi\boldsymbol{\theta}) - \frac{1}{\sigma_0} \sum_{i=1}^{N_p} |\theta_i| + C \right], \quad (11)$$

where C again is a constant that does not depend on $\boldsymbol{\theta}$. Following the steps as before: multiplying by $2\sigma^2$, dropping the constant term, and multiplying by -1 , we have

$$\boldsymbol{\theta}_{MAP} = \arg \min_{\boldsymbol{\theta}} \left[(\mathcal{D} - \Phi\boldsymbol{\theta})^T (\mathcal{D} - \Phi\boldsymbol{\theta}) + \frac{2\sigma^2}{\sigma_0} \sum_{i=1}^{N_p} |\theta_i| \right]. \quad (12)$$

This is the LASSO regression estimator with regularization parameter $\lambda = \frac{2\sigma^2}{\sigma_0}$. Similarly to the Ridge regression case, the regularization parameter can be interpreted as a penalty term that balances the fit to the data, influenced by our confidence in the data and prior knowledge. However, λ now scales with σ_0 instead of σ_0^2 , reflecting the different nature of the Laplace prior compared to the Gaussian prior.

Concluding Discussion

Considering our expressions for the maximum a posteriori estimators with Gaussian and Laplace priors, we see that these add regularization terms to the ordinary least squares minimization problem. When we have $\sigma_0^2 \rightarrow \infty$ or $\sigma_0 \rightarrow \infty$, the regularization terms vanish, and we recover the ordinary least squares estimator for the two respective cases.

Regularization (or using priors) helps prevent a model from fitting the noise in the data [1]. By adding a penalty (the regularization term) to large coefficients, the method keeps the parameter estimates smaller and more stable. In the Bayesian view, this penalty comes directly from a prior belief that the parameters should not be too large. A Gaussian prior leads to Ridge regression, while a Laplace prior leads to LASSO. Overall, regularization can improve prediction when data are noisy or when there are many correlated or irrelevant features, and it makes the model easier to interpret.

References

- [1] J. Murel and E. Kavakoglu, *What is regularization?* <https://www.ibm.com/think/topics/regularization>, Accessed: 2025-11-10, IBM.