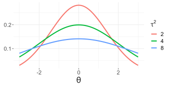
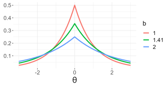
Introduction to Machine Learning

Regularization
Bayesian Priors

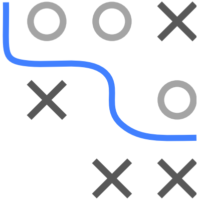




Learning goals

- RRM is same as MAP in Bayes
- Gaussian/Laplace prior corresponds to $L2/L1$ penalty

RRM VS. BAYES



We already created a link between max. likelihood estimation and ERM.

Now we will generalize this for RRM.

Assume we have a parameterized distribution $p(y|\boldsymbol{\theta}, \mathbf{x})$ for our data and a prior $q(\boldsymbol{\theta})$ over our param space, all in Bayesian framework.

From Bayes theorem:

$$p(\boldsymbol{\theta}|\mathbf{x}, y) = \frac{p(y|\boldsymbol{\theta}, \mathbf{x})q(\boldsymbol{\theta})}{p(y|\mathbf{x})} \propto p(y|\boldsymbol{\theta}, \mathbf{x})q(\boldsymbol{\theta})$$

EXAMPLE: BAYESIAN L2 REGULARIZATION

We can easily see the equivalence of *L*2 regularization and a Gaussian prior:

- Gaussian prior $\mathcal{N}_d(\mathbf{0}, diag(\tau^2))$ with uncorrelated components for $\boldsymbol{\theta}$:

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{d} \phi_{0,\tau^2}(\theta_j) = (2\pi\tau^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2\tau^2}\sum_{j=1}^{d}\theta_j^2\right)$$

- MAP:

$$\begin{aligned}
\hat{\theta}^{\text{MAP}} &= \arg\min_{\boldsymbol{\theta}} \left(-\log p\left(y \mid \boldsymbol{\theta}, \mathbf{x}\right) - \log q(\boldsymbol{\theta})\right) \\
&= \arg\min_{\boldsymbol{\theta}} \left(-\log p\left(y \mid \boldsymbol{\theta}, \mathbf{x}\right) + \tfrac{d}{2}\log(2\pi\tau^2) + \frac{1}{2\tau^2}\sum_{j=1}^{d}\theta_j^2\right) \\
&= \arg\min_{\boldsymbol{\theta}} \left(-\log p\left(y \mid \boldsymbol{\theta}, \mathbf{x}\right) + \frac{1}{2\tau^2}\|\boldsymbol{\theta}\|_2^2\right)
\end{aligned}$$

- We see how the inverse variance (precision) $1/\tau^2$ controls shrinkage