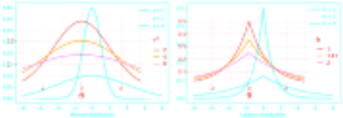
Introduction to Machine Learning

Regularization

Bayesian Priors



Learning goals

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

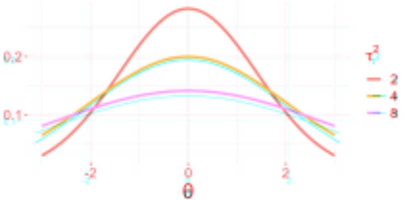
RRM VS. BAYES / 2

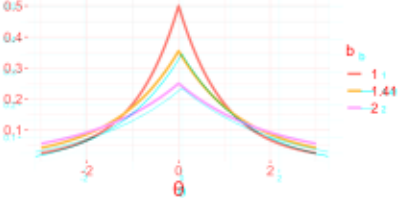
The maximum a posteriori (MAP) estimator of θ is now the minimizer of

$$-\log p\left(y \mid \theta, \mathbf{x}\right) - \log q(\theta).$$

- Again, we identify the loss $L\left(y, f(\mathbf{x} \mid \theta)\right)$ with $-\log(p(y|\theta, \mathbf{x}))$.
- If $q(\theta)$ is constant (i.e., we used a uniform, non-informative prior), the second term is irrelevant and we arrive at ERM.
- If not, we can identify $J(\theta) \propto -\log(q(\theta))$, i.e., the log-prior corresponds to the regularizer, and the additional λ, which controls the strength of our penalty, usually influences the peakedness / inverse variance / strength of our prior.

RRM VS. BAYES / 3





- *L2* regularization corresponds to a zero-mean Gaussian prior with constant variance on our parameters: $\theta_j \sim \mathcal{N}(0, \tau^2)$
- *L1* corresponds to a zero-mean Laplace prior: $\theta_j \sim \textit{Laplace}(0, b)$. $\textit{Laplace}(\mu, b)$ has density $\frac{1}{2b}\exp(-\frac{|\mu - x|}{b})$, with scale parameter b, mean μ and variance $2b^2$.
- In both cases, regularization strength increases as variance of prior decreases: more prior mass concentrated around 0 encourages shrinkage.
- Elastic-net regularization corresponds to a compromise between Gaussian and Laplacian priors ▸ Zou and Hastie 2005 ▸ Hans 2011

EXAMPLE: BAYESIAN L2 REGULARIZATION

We can easily see the equivalence of $L2$ 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:

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

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

EXAMPLE: BAYESIAN L2 REGULARIZATION / 2

- DGP $y = \theta + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 1)$ and $\theta = 1$;
  with Gaussian prior on θ, so $\mathcal{N}(0, \tau^2)$ for $\tau \in \{0.25, 0.5, 2\}$
- For $n = 20$, posterior of θ and MAP can be calculated analytically
- Plotting the $L2$ regularized empirical risk $\mathcal{R}_{reg}(\theta) = \sum_{i=1}^{n}(y_i - \theta)^2 + \lambda\theta^2$ with $\lambda = 1/\tau^2$ shows that ridge solution is identical with MAP
- In our simulation, the empirical mean is $\bar{y} = 0.94$, with shrinkage toward 0 induced in the MAP