General concept of regularization

Regularization is a technique that imposes constraints on model parameters to control the solution space. By limiting where parameters can be found, it effectively shrinks the space of possible solutions. This reduction in parameter flexibility not only enhances model generalizability but also helps prevent overfitting by focusing on simpler solutions.

The term derives from Latin “regula” (rule) and “regularis” (in accordance with rules), reflecting its role in establishing systematic constraints on model behavior.

Through these controlled parameter constraints and reduced solution space, regularization helps create simpler, more robust models by reducing their sensitivity to noise in the training data.

The regularization method traces back to A.N. Tikhonov’s work in 1963, who proposed it for solving ill-posed problems where formal mathematical solutions are meaningless.

A linear system $\boldsymbol{y} = X\boldsymbol{\beta}$ has no solution when X is singular ($\det X = 0$), rank-deficient ($\operatorname{rank} X < k$), or when data contains errors preventing $\boldsymbol{y} = X\boldsymbol{\beta}$ from being satisfied:

$$\|y - X\beta\|_2^2 \overset{?}{\to} \min_{\beta}.$$

Adding constraints on model parameters $\boldsymbol{\beta}$ through a regularizer $R(\boldsymbol{\beta})$ term shrinks the solution space, making it possible to find (sometimes) a practically useful approximate solution:

$$\|y - X\beta\|_2^2 + R(\beta) \overset{\text{OK}}{\to} \min_{\beta}.$$

Probabilistic interpretation of regularization

Probabilistic framework. Consider a joint distribution of data $\boldsymbol{x} \in \mathbb{R}^k, y \in \mathbb{R}$ and model's parameters $\boldsymbol{\theta} \in \mathbb{R}$:

$$\boldsymbol{x}, y, \boldsymbol{\theta} \sim X, Y, \Theta. \tag{1}$$

1. The prior distribution of $\boldsymbol{x}$ is independent of parameters $\boldsymbol{\theta}$ and can assumed to be uniform and ignored in the model:

$$X \sim f_X(\boldsymbol{x}|\boldsymbol{\theta}) \;\coloneqq\; f_X(\boldsymbol{x}) \sim \mathcal{U}. \tag{2}$$

2. The posterior distribution of responses y depends on parameters $\boldsymbol{\theta}$ and specific data point $\boldsymbol{x}'$, following a semi-probabilistic model formalism. The model is specified by defining the conditional distribution of responses $\mathbb{P}[y|\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{\theta}]$ given a specific $\boldsymbol{x} = \boldsymbol{x}^*$ and model parameters $\boldsymbol{\theta}$. When the parameters are fitted, we make predictions for new data points $\boldsymbol{x}'$ by maximizing the probability of a response y given $\boldsymbol{x}'$:

$$a_{\boldsymbol{\theta}}(\boldsymbol{x}') = \arg \max_{y \in \operatorname{supp} Y} \underbrace{f_Y\left(y|\boldsymbol{x} = \boldsymbol{x}', \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}\right)}_{\text{model}}. \tag{3}$$

Support of a random variable Y is the set of all possible values $y_1^*, y_2^*, \ldots$ that Y can take with non-zero probability:

$$\operatorname{supp} Y = \{y_1^*, y_2^*, \ldots\}$$

3. The prior distribution of parameters $\boldsymbol{\theta}$ is assumed to be known and defined by the hyperparameter vector $\boldsymbol{\gamma}$:

$$f_\Theta(\boldsymbol{\theta}) \;\coloneqq\; f_\Theta(\boldsymbol{\theta}|\boldsymbol{\gamma}) \sim \Theta(\boldsymbol{\gamma}) \tag{4}$$

Applying MAP. The joint distribution of data and parameters can be rewritten as a product of conditional pdfs:

$$\begin{aligned} f(\boldsymbol{x}, y, \boldsymbol{\theta}) &= f(y|\boldsymbol{x}, \boldsymbol{\theta}) \cdot f(\boldsymbol{x}, \boldsymbol{\theta}) \\ &= f(y|\boldsymbol{x}, \boldsymbol{\theta}) \cdot \cancel{f(\boldsymbol{x}|\boldsymbol{\theta})} \cdot f(\boldsymbol{\theta}|\boldsymbol{\gamma}) \\ &= f(y|\boldsymbol{x}, \boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}|\boldsymbol{\gamma}) \end{aligned} \tag{5}$$

We omit random variables X, Y, Θ in the pdf's underscripts for brevity. Just look at the arguments before the bar to understand to which random variable the pdf refers: *e.g.*, $f(x, y|\theta)$ means $f_{X,Y}(\boldsymbol{x}, y|\boldsymbol{\theta})$.

$$\begin{aligned} \mathbb{P}[y|x, \theta] &= \mathbb{P}[\{Y = y\}|\{X = x\}, \{\Theta = \theta\}] \\ &= \mathbb{P}[\{Y = y\}|\{X = x\}\{\Theta = \theta\}] \\ &= \frac{\mathbb{P}[\{Y = y\}\{X = x\}\{\Theta = \theta\}]}{\mathbb{P}[\{X = x\}\{\Theta = \theta\}]} \\ &= \frac{\mathbb{P}[x, y, \theta]}{\mathbb{P}[x, \theta]} \end{aligned}$$

As it was mentioned, the canceled prior distribution of data $f(\boldsymbol{x}|\boldsymbol{\theta})$ is independent of the model parameters $\boldsymbol{\theta}$. We ignore it (or assume uniform).

Still, we didn't ignore the prior distribution of parameters $\boldsymbol{\theta}$, which is $f(\boldsymbol{\theta}|\boldsymbol{\gamma})$. Because of that, it's MAP (Maximum a Posteriori) estimation, not MLE (Maximum Likelihood Estimation).

Finding parameters. For specific training samples $y^*, \boldsymbol{x}^*$ and predefined hyperparameters $\boldsymbol{\gamma}^*$, we write the joint likelihood of data and model parameters and maximize it:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log \prod_{(\boldsymbol{x}^*, y^*) \in (X,Y)^\ell} \underbrace{f(y = y^*|\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}|\boldsymbol{\gamma} = \boldsymbol{\gamma}^*)}_{\text{MAP}} \\ &= \sum_{(\boldsymbol{x}^*, y^*) \in (X,Y)^\ell} \{\log f(y = y^*|\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{\theta}) + \log f(\boldsymbol{\theta}|\boldsymbol{\gamma} = \boldsymbol{\gamma}^*)\} \\ &= \sum_{(\boldsymbol{x}^*, y^*) \in (X,Y)^\ell} \underbrace{\log f(y = y^*|\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{\theta})}_{\text{log-likelihood}} + \underbrace{\lambda \cdot \log f(\boldsymbol{\theta}|\boldsymbol{\gamma} = \boldsymbol{\gamma}^*),}_{\text{prior regularizer}}) \to \max_{\boldsymbol{\theta}} \end{aligned} \tag{6}$$

A pdf $f(y|x, \theta)$ becomes a likelihood function when we consider it as a function of arguments behind the bar, *e.g.*

- $h(y) \coloneqq f(y|x = x^*, \theta = \theta^*)$ is still a pdf of y given $x = x^*$ and $\theta = \theta^*$.
- $g(\theta) \coloneqq f(y = y^*|x = x^*, \theta)$ is already a likelihood function of θ given $x = x^*$ and $y = y^*$.

The second term is the regularizer, its strength is defined by constant λ and hyperparameters $\boldsymbol{\gamma}$. Regularizer narrows the space in which the parameters can be found. The more narrow the space, the more constrained the model is.

After finding the parameter vector estimate $\hat{\boldsymbol{\theta}}$, predictions for a new data point $\boldsymbol{x}'$ can be made by substituting the estimate $\hat{\boldsymbol{\theta}}$ into the model $f_Y\left(y|\boldsymbol{x} = \boldsymbol{x}', \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}\right)$:

$$a_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}') = \arg \max_{y \in \operatorname{supp} Y} f_Y\left(y|\boldsymbol{x} = \boldsymbol{x}', \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}\right) \tag{7}$$

Loss-function. Probabilistic regularizer (6) can be rewritten as the empirical risk where it becomes an additional loss function:

$$R(\boldsymbol{\theta}) = \underbrace{\sum_{\boldsymbol{x} \in X^\ell} \mathcal{L}(\boldsymbol{x})}_{\text{data fitting}} + \underbrace{\lambda \cdot \mathcal{L}_{\text{reg}(\boldsymbol{\theta})}}_{\substack{\text{parameters} \\ \text{regularization}}} \to \min_{\boldsymbol{\theta}} \tag{8}$$

L_2-norm regularization (Tikhonov regularization)

On a privious step, we found the general form of the regularazer assuming prior distribution of parameters $f_\Theta(\boldsymbol{\theta}|\gamma)$:

$$R(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}\in X^\ell} \mathcal{L}(\boldsymbol{x}) + \lambda \cdot \mathcal{L}_{\text{reg}}(\boldsymbol{\theta}) \to \min_{\boldsymbol{\theta}}. \tag{9}$$

Model:. Here we make specific choices of the prior distributions parameters $f_\Theta(\boldsymbol{\theta}|\gamma)$ and the data:

1. All parameters $\boldsymbol{\theta} \coloneqq \boldsymbol{\beta}$ are independent and linear, so the joint distribution is a product of the individual distributions:

$$f_\Theta(\boldsymbol{\theta} = \boldsymbol{\beta}|\gamma) = \prod_{j=1}^{k} f(\beta_j|\gamma) \tag{10}$$

These distributions impose prior constraints on the model coefficients, effectively reducing the solution space

2. Each parameter β_j follows a Gaussian distribution with two hyperparameters common for all individual distributions: mean $\gamma_1 \coloneqq \mu = 0$ and standard deviation $\gamma_2 \coloneqq \tau$:

$$f(\beta_j|\gamma) \coloneqq f(\beta_j|\mu,\tau) = \frac{1}{\sqrt{2\pi}\tau} e^{-\beta_j^2/2\tau^2} \sim N(\mu,\tau). \tag{11}$$

3. The data is generated by a linear model with Gaussian noise $N(0,\sigma)$:

$$y(\boldsymbol{x}) = \beta^{\mathsf{T}}\,\boldsymbol{x} + \varepsilon(\boldsymbol{x}), \quad \varepsilon(\boldsymbol{x}) \sim N(0,\sigma). \tag{12}$$

Both errors $\varepsilon \sim N(\mathbf{0},\sigma I)$ and model's parameters $\beta \sim N(\mathbf{0},\tau I)$ follow multivariate Gaussian distributions with zero mean and different covariance matrices σI and τI respectively.

Dimensions of vectors ε and β are different:

$$\dim \varepsilon = \ell, \quad \dim \beta = k.$$

All corresponding distribution parameters have appropriate dimensions:

$$\dim \mathbf{0}_\varepsilon = \ell, \quad \dim \mathbf{0}_\beta = k,$$
$$\dim \sigma I = \ell \times \ell, \quad \dim \tau I = k \times k.$$

Here we denote pdf underscripts with letters corresponding to random variables

Applying MAP. For any arbitrary model we can estimate the error term $\hat{\boldsymbol{\varepsilon}}$ as difference between the predicted $\hat{\boldsymbol{y}}$ and the actual $\boldsymbol{y}$ responses. The posterior distribution of parameters $\boldsymbol{\beta}$:

$$\begin{aligned} f_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\gamma,\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}}) &= \frac{f_{\boldsymbol{\beta},\gamma,\boldsymbol{\varepsilon}}(\boldsymbol{\beta},\gamma,\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}})}{f_{\gamma,\boldsymbol{\varepsilon}}(\gamma,\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}})} \\ &= \frac{f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}}|\boldsymbol{\beta},\gamma)\cdot f_{\boldsymbol{\beta},\gamma}(\boldsymbol{\beta},\gamma)}{f_{\gamma,\boldsymbol{\varepsilon}}(\gamma,\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}})} \\ &= \frac{f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}}|\boldsymbol{\beta},\gamma)\cdot f_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\gamma)\cdot f_{\gamma}(\gamma)}{f_{\gamma,\boldsymbol{\varepsilon}}(\gamma,\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}})} \\ &= \frac{f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}}|\boldsymbol{\beta})\cdot f_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\gamma)\cdot \cancel{f_{\gamma}(\gamma)}}{\cancel{f_{\gamma}(\gamma)}\cdot f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}})} \to \max_{\boldsymbol{\beta}} \end{aligned} \tag{13}$$

As $f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}})$ is independent of $\boldsymbol{\beta}$, we cancel it out:

$$\boxed{f_{\varepsilon}\,(\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}}|\boldsymbol{\beta})\cdot f_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \to \max_{\boldsymbol{\beta}}.} \tag{14}$$

Independence. We applied MAP and wrote the optimization problem, now continue with substituting the specific distributions:

$$f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}}|\boldsymbol{\beta}) = \prod_{\boldsymbol{x}^*\in X^\ell} e^{-\varepsilon(\boldsymbol{x}=\boldsymbol{x}^*|\boldsymbol{\beta})^2/2\sigma^2} \tag{15}$$

The error estimates are directly related to the data:

$$\hat{\varepsilon}(\boldsymbol{x}=\boldsymbol{x}^*|\boldsymbol{\beta}) = \hat{y}(\boldsymbol{x}^*|\boldsymbol{\beta}) - y(\boldsymbol{x}^*) \tag{16}$$

Let's write the prior distribution of parameters:

$$f_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \prod_{j=1}^{k} e^{-\frac{\beta_j^2}{2}\tau^2} \tag{17}$$

The data distribution can be written through the error distribution:

$$f_Y(\boldsymbol{y}=\boldsymbol{y}^*|\boldsymbol{\beta}) = f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}=\boldsymbol{y}-\boldsymbol{y}^*|\boldsymbol{\beta}) \tag{18}$$

Let's write the posterior distribution of parameters:

$$f_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\varepsilon}=\hat{\boldsymbol{\varepsilon}}) \coloneqq \prod_{\boldsymbol{x}\in X^\ell} e^{-\hat{\varepsilon}\frac{(\boldsymbol{x}|\boldsymbol{\beta})^2}{2}\sigma^2} \cdot \prod_{j=1}^{k} e^{-\frac{\beta_j^2}{2}\tau^2} \tag{19}$$

Let's write the likelihood function (log-loss):

$$\ell(\boldsymbol{\varepsilon}, \boldsymbol{\beta}|X) \coloneqq -\sum_{\boldsymbol{x}\in X^\ell} \frac{\varepsilon(\boldsymbol{x}|\boldsymbol{\beta})^2}{2\sigma^2} - \sum_{j=1}^{k} \frac{\beta_j^2}{2\tau^2} \to \max_{\boldsymbol{\beta}} \tag{20}$$

Let's rewrite it as empirical risk minimization:

$$Q(\boldsymbol{\beta}) = \sum_{\boldsymbol{x}\in X^\ell} (\hat{y}(\boldsymbol{x}) - y(\boldsymbol{x}))^2 + \frac{2\sigma^2}{2\tau^2} \sum_{j=1}^{k} \beta_j^2 \to \min_{\boldsymbol{\beta}} \tag{21}$$

$$Q(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \cdot \|\boldsymbol{\beta}\|_2^2 \to \min_{\boldsymbol{\beta}} \tag{22}$$

In L_1 regularization, everything is similar, but the errors are described by the Laplace distribution.

L_1-norm regularization

$$Q(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \cdot \|\boldsymbol{\beta}\|_1 \to \min_{\boldsymbol{\beta}} \tag{23}$$

Unlike L_2 regularization, LASSO assumes model errors follow the Laplace distribution, characterized by heavy tails and a sharp peak:

$$\varepsilon \sim \text{pdf}(\varepsilon|\mu, b) = \frac{1}{2b} \exp \frac{-|\varepsilon - \mu|}{b} \tag{24}$$

Using MAP for parameter estimation:

$$\begin{aligned} f_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\varepsilon} = \hat{\boldsymbol{\varepsilon}}) &= \frac{f_{\varepsilon}(\boldsymbol{\varepsilon} = \hat{\boldsymbol{\varepsilon}}|\boldsymbol{\beta}) \cdot f_{\boldsymbol{\beta}}(\boldsymbol{\beta})}{\cancel{f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon} = \hat{\boldsymbol{\varepsilon}})}} \\ &= \prod_{\boldsymbol{x}^* \in X^\ell} e^{-\frac{\varepsilon(\boldsymbol{x}=\boldsymbol{x}^*|\boldsymbol{\beta})^2}{2}\sigma^2} \cdot \prod_{j=1}^{k} \frac{1}{2b} e^{-\frac{|\beta_j|}{b}} \\ &\to \max_{\boldsymbol{\beta}} \end{aligned} \tag{25}$$

Let's write the likelihood function (log-loss):

$$\ell(\boldsymbol{\varepsilon}, \boldsymbol{\beta}|X) := -\sum_{\boldsymbol{x} \in X^\ell} \frac{\varepsilon(\boldsymbol{x}|\boldsymbol{\beta})^2}{2\sigma^2} - \sum_{j=1}^{k} \frac{|\beta_j|}{b} - \cancel{k \cdot \ln 2b} \to \max_{\boldsymbol{\beta}} \tag{26}$$

Let's rewrite it as empirical risk minimization:

$$Q(\boldsymbol{\beta}) = \sum_{\boldsymbol{x} \in X^\ell} (\hat{y}(\boldsymbol{x}) - y(\boldsymbol{x}))^2 + \frac{2\sigma^2}{b} \sum_{j=1}^{k} |\beta_j| \to \min_{\boldsymbol{\beta}} \tag{27}$$

L_0 regularization

Geometric interpretation of regularization

Ridge, LASSO, and Elastic Net regression