ridge regression

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

Tikhonov regularization, weight decay

Abstract

Ridge regression learns a linear hypothesis map by minimizing the average squared error loss on a training set plus a penalty term proportional to the squared Euclidean norm of the model parameters. The penalty term enforces regularization: it prevents overfitting when the number of features exceeds the number of data points, or when features are nearly collinear. In contrast to least absolute shrinkage and selection operator (Lasso), ridge regression uses an objective function with a unique minimizer, given in closed form by one system of linear equations, for any training set. In the basis of the singular vectors of the feature matrix, ridge regression shrinks each component of the linear regression solution by a factor between 0 and 1 that is small for a small singular value, which bounds the noise amplification of linear regression. The penalty term represents the increase of the average squared error loss under data augmentation with random feature perturbations. Ridge regression on transformed feature vectors constructed from a kernel is kernel ridge regression.

Definition

Predicting the maximum temperature of tomorrow at Krems an der Donau from the 24 hourly temperatures of today is a machine learning (ML) application. A data point is one day, its feature vector $\mathbf{x} \in \mathbb{R}^{24}$ lists the hourly temperatures of the day, and its label y is the maximum temperature of the next day (GeoSphere Austria, 2026). A training set of $m = 20$ days of July 2024 has fewer data points than features, $d = 24$, and the features are nearly collinear: the temperatures at 13:00 and at 14:00 have correlation 0.93. Linear regression then learns a linear model that reproduces the 20 training labels exactly and predicts the maximum temperatures of August 2024, the test set, with a mean squared error (MSE) of 141. Ridge regression learns a linear model from the same training set with a training MSE of 5.4 and an MSE of 8.3 on the test set (see Fig. 1).

Ridge regression learns the model parameters $\mathbf{w}$ of a linear hypothesis map $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^{\top}\mathbf{x}$, as does linear regression. It minimizes the penalized average

squared error loss on a training set $\mathcal{D}^{(\text{train})} = \left\{ \left(\mathbf{x}^{(r)}, y^{(r)}\right) \right\}_{r=1}^{m}$ (Hoerl and Kennard, 1970),

$$\widehat{\mathbf{w}}^{(\alpha)} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{m} \sum_{r=1}^{m} \left(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\right)^2 + \alpha \|\mathbf{w}\|_2^2. \tag{1}$$

The optimization problem can be written more compactly using the feature matrix $\mathbf{X}$ and the label vector $\mathbf{y}$,

$$\mathbf{X} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right)^\top \in \mathbb{R}^{m \times d}, \qquad \mathbf{y} = \left(y^{(1)}, \ldots, y^{(m)}\right)^\top \in \mathbb{R}^m.$$

In terms of $\mathbf{X}$ and $\mathbf{y}$, the ridge regression problem (1) reads

$$\widehat{\mathbf{w}}^{(\alpha)} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{m} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \alpha \|\mathbf{w}\|_2^2.$$

Ridge regression is obtained from linear regression by adding a penalty term to the average squared error loss. The penalty term is the scaled squared Euclidean norm $\alpha \|\mathbf{w}\|_2^2$ with a regularization parameter $\alpha > 0$ (Hastie et al., 2009, Sect. 3.4.1, Eq. (3.41)). The resulting method is an instance of regularized empirical risk minimization (RERM). The purpose of the penalty term is regularization. It prevents overfitting in the high-dimensional regime, where the number of features d exceeds the number of data points m in the training set (Hastie et al., 2009, Sect. 18.1), as it does for the 20 days with 24 hourly temperatures each. An alternative method for regularization is least absolute shrinkage and selection operator (Lasso), which differs from ridge regression in the construction of the penalty term: Lasso uses the ℓ_1-norm $\|\mathbf{w}\|_1$ instead of the squared Euclidean norm $\|\mathbf{w}\|_2^2$.

In contrast to Lasso, whose objective can have several minimizers, the ridge regression objective function has, for any feature matrix $\mathbf{X}$ and label vector $\mathbf{y}$, the unique minimizer

$$\widehat{\mathbf{w}}^{(\alpha)} = \left(\mathbf{X}^\top \mathbf{X} + \alpha m \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{2}$$

The normal equations of plain linear regression, by comparison, can have infinitely many solutions, as they do for the 20 days of July: every one of them reproduces the training labels. Training is a one-shot computation: ridge regression requires no iterative optimization method, and hence no fixed-point interpretation, only the solution of the $d \times d$ system of linear equations $\left(\mathbf{X}^\top \mathbf{X} + \alpha m \mathbf{I}\right)\mathbf{w} = \mathbf{X}^\top \mathbf{y}$ with an invertible matrix, the zero-gradient condition of the penalized objective function. For many features, gradient descent (GD) on the penalized objective function is the alternative, and the penalty term adds αm to every eigenvalue of $\mathbf{X}^\top \mathbf{X}$, which lowers the condition number that sets the convergence rate of GD.

The singular value decomposition (SVD) of the feature matrix shows why the linear regression solution amplifies the noise in the labels and how the penalty term bounds the amplification. Write $\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^\top$ with singular values $\lambda_1 \geq$

$\ldots \geq \lambda_p > 0$, where p is the rank of $\mathbf{X}$, with left singular vectors $\mathbf{v}^{(j)}$ and right singular vectors $\mathbf{u}^{(j)}$. The solution of linear regression with the smallest Euclidean norm is

$$\widehat{\mathbf{w}} = \sum_{j=1}^{p} \frac{(\mathbf{v}^{(j)})^\top \mathbf{y}}{\lambda_j} \, \mathbf{u}^{(j)}.$$

Two properties of this solution cause its large test error. A component of $\mathbf{y}$ along $\mathbf{v}^{(j)}$ is divided by λ_j, so a small singular value amplifies the noise in $\mathbf{y}$ by the factor $1/\lambda_j$. Small singular values arise from nearly collinear features: the singular values of the 20 days of hourly temperatures run from 504 down to 0.21. And a direction $\mathbf{u}^{(j)}$ with $j > p$, which exists whenever $d > m$, leaves $\mathbf{X}\mathbf{w}$ unchanged, so the training set does not determine the component of $\mathbf{w}$ along it.

Inserting the SVD into the closed form (2) gives (Hastie et al., 2009, Sect. 3.4.1, Eq. (3.47))

$$\widehat{\mathbf{w}}^{(\alpha)} = \sum_{j=1}^{p} \frac{\lambda_j^2}{\lambda_j^2 + \alpha m} \cdot \frac{(\mathbf{v}^{(j)})^\top \mathbf{y}}{\lambda_j} \, \mathbf{u}^{(j)}.$$

Ridge regression multiplies each component of $\widehat{\mathbf{w}}$ in the basis of the right singular vectors by the shrinkage factor $\lambda_j^2/(\lambda_j^2 + \alpha m)$, which lies strictly between 0 and 1. For $\lambda_j^2 \gg \alpha m$ the factor is close to 1 and the component is kept. For $\lambda_j^2 \ll \alpha m$ the component is shrunk toward 0. The product of the factor and $1/\lambda_j$ is $\lambda_j/(\lambda_j^2 + \alpha m)$, which never exceeds $1/(2\sqrt{\alpha m})$, so the noise amplification is bounded: for the 20 days and $\alpha = 1$, the bound is 0.11, where $1/\lambda_j$ reaches 4.9. The undetermined components with $j > p$ are set to 0, which is why the minimizer is unique for any feature matrix. Fig. 1 shows the effect on the learned model parameters: those of linear regression swing between -4.1 and 4.9 from one hour to the next, those of ridge regression stay within ± 0.4.

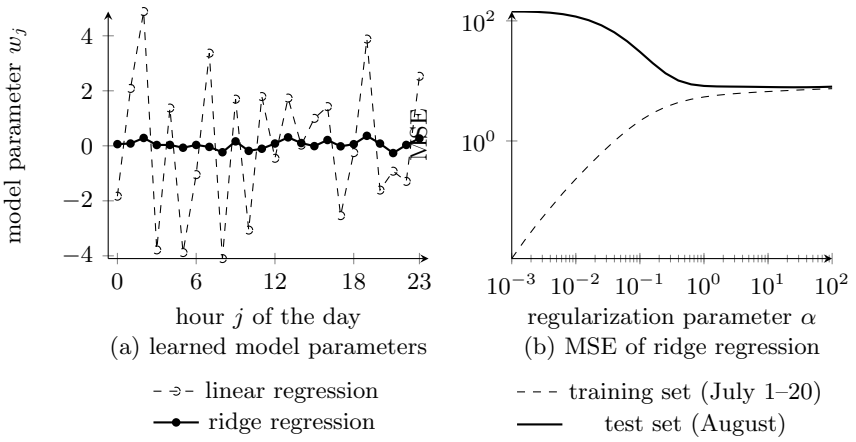


Figure 1: Ridge regression for the maximum temperature of the next day at Krems an der Donau, learned from the 24 hourly temperatures of $m = 20$ days of July 2024. Panel (a) shows the 24 model parameters of the linear model learned by linear regression (dashed line, open circles) and by ridge regression with $\alpha = 1$ (solid line, dots). Panel (b) shows the MSE of ridge regression on the training set (dashed curve) and on the 31 days of August 2024 (solid curve) over the regularization parameter α. Data generated by `pythondemos/ridgeregression.py`

The penalty term $\alpha\|\mathbf{w}\|_2^2$ has an interpretation as a form of data augmentation (Bishop, 1995, Eqs. (10)–(11)). In particular, adding the penalty term to the average squared error loss of the hypothesis $h^{(\mathbf{w})}$ is equivalent to computing its average squared error loss on an augmented training set (see Fig. 2). This augmented training set is obtained from the original training set. For each data point $\left(\mathbf{x}^{(r)}, y^{(r)}\right)$, $r = 1, \ldots, m$, the feature vector $\mathbf{x}^{(r)}$ is replaced by realizations of infinitely many independent and identically distributed (i.i.d.) random variables (RVs) $\mathbf{x}^{(r)} + \boldsymbol{\varepsilon}^{(r)}$ with Gaussian random variables (Gaussian RVs) $\boldsymbol{\varepsilon}^{(r)} \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I})$, i.e., each feature is perturbed independently with variance α. The label $y^{(r)}$ is left unchanged.

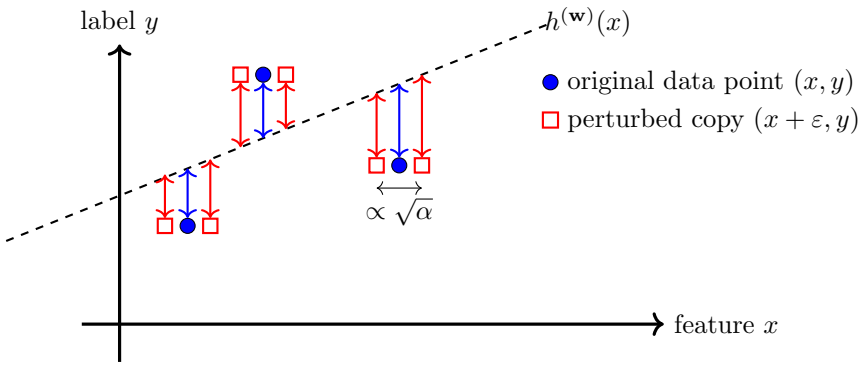


Figure 2: For a linear model with one feature, adding the penalty term $\alpha\|\mathbf{w}\|_2^2$ to the average squared error loss is equivalent to computing the average squared error loss on an augmented training set: each original data point (blue circle) is replaced by copies (red squares) whose feature is perturbed by a Gaussian RV ε with variance α, keeping the label. The vertical arrows show the error $y - h^{(\mathbf{w})}(x)$ of each data point

The penalty term $\alpha\|\mathbf{w}\|_2^2$ is exactly the amount by which the average squared error loss of $h^{(\mathbf{w})}$ over the infinitely many perturbed copies of a data point exceeds its squared error loss on the original data point. Since $\mathbb{E}\{\boldsymbol{\varepsilon}^{(r)}\} = \mathbf{0}$, the increase is

$$\mathbb{E}\{\left(\mathbf{w}^\top \boldsymbol{\varepsilon}^{(r)}\right)^2\} = \mathbf{w}^\top \mathbb{E}\{\boldsymbol{\varepsilon}^{(r)}\left(\boldsymbol{\varepsilon}^{(r)}\right)^\top\}\mathbf{w} = \alpha\|\mathbf{w}\|_2^2.$$

This increase of the average loss under data augmentation is an estimate for the generalization gap (Hastie et al., 2009, Sect. 7.4). Ridge regression applied to transformed feature vectors constructed from a kernel is kernel ridge regression, which learns a nonlinear hypothesis in closed form; the perturbation then acts on the transformed feature vectors.

Cross References

regression, linear regression, Lasso, regularization, penalty term, SVD, condition number, data augmentation, kernel ridge regression

References

1. Bishop (1995). Training with Noise is Equivalent to Tikhonov Regularization. Neural Computation, 7(1), pp. 108–116.

2. GeoSphere Austria (2026). GeoSphere Austria climate station archive. Data hub, station historical resource klima-v2-1d. URL: https://dataset.api.hub.geosphere.at/v1/station/v2-1d.

3. Hoerl and Kennard (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), pp. 55–67.

4. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.