

effective dimension

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)

Editors: Konstantina Oliumtsevit (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

The effective dimension of a hypothesis space is its size as seen through the probability distribution of the features: the number of model parameters that make a difference on the data points the distribution produces. For a linear model it is the rank of the second-moment matrix of the feature map, estimated by the rank of the feature matrix, and it never exceeds the number of model parameters. Two hypothesis spaces that differ only outside the support of the distribution have the same effective dimension, and empirical risk minimization (ERM) on them delivers the same predictions on the data. The same hypothesis space can have different effective dimensions under different data distributions. The Rademacher complexity is the distribution-dependent size measure of learning theory, while the Vapnik–Chervonenkis dimension (VC dimension) is its worst case over the whole feature space.

Definition

Consider the 366 days of 2024 at the weather station Krems an der Donau, each a data point with the minimum temperature of the day as its feature x and the maximum temperature as its label y . Two hypothesis spaces are candidates for predicting y from x : the affine linear model $\mathcal{H}^{(1)} = \{w_0 + w_1x\}$ with two model parameters, and the linear model $\mathcal{H}^{(2)} = \{w_0 + w_1x + w_2 \max(0, x - 25)\}$ with three, whose third feature is a rectified linear unit (ReLU) hinge at 25 °C. The minimum temperature of the record ranges from -10.1 to 20.8 °C, so the hinge is zero on every day. On this record, every hypothesis of $\mathcal{H}^{(2)}$ agrees with a hypothesis of $\mathcal{H}^{(1)}$: the third model parameter changes no prediction that the data can test.

The effective dimension $d_{\text{eff}}(\mathcal{H})$ of a hypothesis space $\mathcal{H}$ is its size as seen through the probability distribution $\mathbb{P}^{(x)}$ of the features: it counts the model parameters that make a difference on the data points the distribution produces, not the model parameters that are written down. For a linear model with feature map Φ , the effective dimension is the rank of the second-moment matrix $\mathbb{E}\{\Phi(x)\Phi(x)^\top\}$, which a dataset estimates by the rank of its feature matrix (Hastie et al., 2009, Sect. 7.6). It never exceeds the number of model parameters, and it equals that number only when the features are linearly independent on the support of $\mathbb{P}^{(x)}$. On the Krems record the feature matrix of $\mathcal{H}^{(2)}$ has three

columns and rank 2, since its hinge column is zero, so $d_{\text{eff}}(\mathcal{H}^{(1)}) = d_{\text{eff}}(\mathcal{H}^{(2)}) = 2$. Two hypothesis spaces that differ only outside the support of $\mathbb{P}^{(x)}$ have the same effective dimension.

The consequence for empirical risk minimization (ERM) is visible in Fig. 1(a). Least squares on $\mathcal{H}^{(1)}$ and on $\mathcal{H}^{(2)}$ delivers the same prediction for every day of the record and the same training error of 19.2 squared degrees. The hinge coefficient w_2 is not identifiable: every value gives the same training error, and the two learned hypotheses differ only for a minimum temperature above 25 °C, a region that no day of the record reaches. There the data say nothing, which is the generalization problem in its sharpest form: a hypothesis is tested only where the features lie.

The effective dimension depends on the probability distribution, not only on the hypothesis space. Fig. 1(b) keeps $\mathcal{H}^{(2)}$ and changes the feature to the maximum temperature of the day, with the maximum temperature of the next day as the label. The maximum temperature exceeds 25 °C on 97 of the 365 day pairs, the hinge column of the feature matrix is no longer zero, its rank is 3, and least squares on $\mathcal{H}^{(2)}$ now reaches a smaller training error than on $\mathcal{H}^{(1)}$, if only slightly. The same hypothesis space has effective dimension 2 under one data distribution and 3 under another.

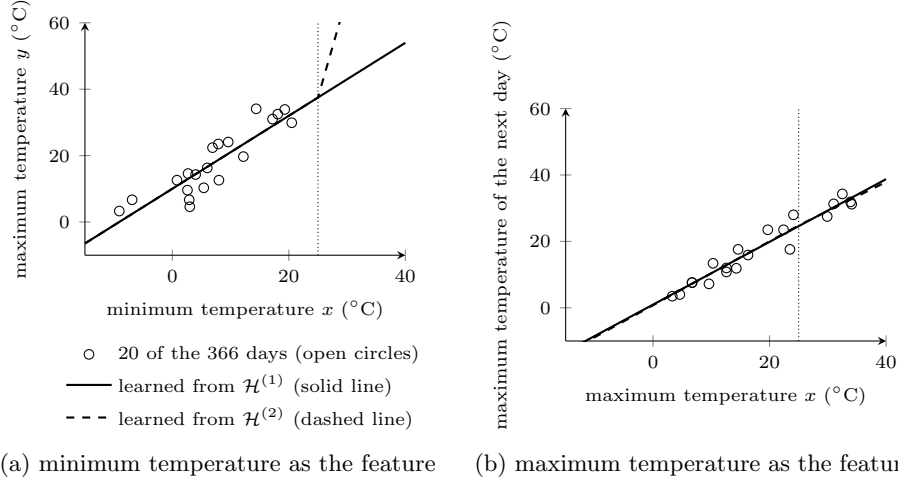


Figure 1: The same two hypothesis spaces under two data distributions at Krems an der Donau, 2024. In (a) the feature is the minimum temperature of the day, which never reaches the hinge at 25 °C (dotted vertical line): the hypotheses learned by least squares from $\mathcal{H}^{(1)}$ (solid line) and from $\mathcal{H}^{(2)}$ (dashed line, drawn with the hinge coefficient set to 5) coincide wherever the data lie and differ only beyond it. In (b) the feature is the maximum temperature of the day and the label the maximum temperature of the next day; the hinge is active on the hot days, and the two learned hypotheses differ, slightly, beyond 25 °C. Open circles are 20 of the days, every 19th in date order. Weather data: GeoSphere Austria (GeoSphere Austria, 2026). Data generated by `pythondemos/effdim.py`

Beyond linear models, the size measures of learning theory also differ by whether they take a worst case over the feature space or an average over the data distribution. The Vapnik–Chervonenkis dimension (VC dimension) is a worst case over all datasets that the feature space admits, so it counts the hinge of $\mathcal{H}^{(2)}$ whether or not the data ever reach it. The Rademacher complexity is an expectation over datasets drawn from $\mathbb{P}^{(x)}$ and is therefore the effective size of $\mathcal{H}$ for the application at hand (Bartlett and Mendelson, 2002). For a method such as ridge regression, whose learned hypothesis is a linear function of the labels, the trace of that linear map is the effective number of model parameters: it decreases from the rank of the feature matrix toward zero as the regularization parameter α grows (Hastie et al., 2009, Sect. 7.6).

Cross References

hypothesis space, model parameter, dimension, linear model, feature map, feature matrix, rank, probability distribution, generalization, VC dimension, Rademacher complexity, ridge regression

References

1. 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>.
2. Bartlett and Mendelson (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Nov), pp. 463–482.
3. Hastie et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer Science+Business Media, New York, NY, USA.