Introduction to Deep Learning

Chapter 3: Basic Regularization

Bernd Bischl

Department of Statistics – LMU Munich

WS 2021/2022

REGULARIZATION

- Any technique that is designed to reduce the test error possibly at the expense of increased training error can be considered a form of regularization.
- Regularization is important in DL because NNs can have extremely high capacity (millions of parameters).

REVISION: REGULARIZED RISK MINIMIZATION

- The goal of regularized risk minimization is to penalize the complexity of the model to minimize the chances of overfitting.
- By adding a parameter norm penalty term $J(\boldsymbol{\theta})$ to the empirical risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ we obtain a regularized cost function:

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda J(\boldsymbol{\theta})$$

  with hyperparamater $\lambda \in [0, \infty)$, that weights the penalty term, relative to the unconstrained objective function $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$.
- Therefore, instead of pure **empirical risk minimization**, we add a penalty for complex (read: large) parameters $\boldsymbol{\theta}$.
- Declaring $\lambda = 0$ obviously results in no penalization.
- We can choose between different parameter norm penalties $J(\boldsymbol{\theta})$.
- In general, we do not penalize the bias.

L2-REGULARIZATION / WEIGHT DECAY

Let us optimize the L2-regularized risk of a model $f(\mathbf{x} \mid \boldsymbol{\theta})$

$$\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

by gradient descent. The gradient is

$$\nabla \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \nabla \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}.$$

We iteratively update $\boldsymbol{\theta}$ by step size α times the negative gradient

$$\boldsymbol{\theta}^{[\text{new}]} = \boldsymbol{\theta}^{[\text{old}]} - \alpha \left(\nabla \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^{[\text{old}]}\right) = \boldsymbol{\theta}^{[\text{old}]}(1 - \alpha\lambda) - \alpha \nabla \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

$\rightarrow$ The term $\lambda \boldsymbol{\theta}^{[old]}$ causes the parameter (**weight**) to **decay** in proportion to its size (which gives rise to the name).

L2-REGULARIZATION / WEIGHT DECAY

Weight decay can be interpreted **geometrically**. Let us make a

quadratic approximation of the unregularized objective $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ in the neighborhood of its minimizer $\hat{\boldsymbol{\theta}}$,

$$\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) + \nabla \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) \cdot (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

where $\boldsymbol{H}$ is the Hessian matrix of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$.
Because $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$,

- the first order term is 0 in the expression above because the gradient is 0, and,
- $\boldsymbol{H}$ is positive semidefinite.

Source: Goodfellow et al. (2016), ch. 7

L2-REGULARIZATION / WEIGHT DECAY

The minimum of $\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$ occurs where $\nabla_{\boldsymbol{\theta}}\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ is 0.
Adding the weight decay gradient $\lambda\boldsymbol{\theta}$, we get the regularized version of

$\tilde{\mathcal{R}}_{\text{emp}}(\boldsymbol{\theta})$. We solve it for the minimizer $\hat{\boldsymbol{\theta}}_{\text{Ridge}}$:

$$\lambda\boldsymbol{\theta} + \boldsymbol{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 0$$
$$(\boldsymbol{H} + \lambda\boldsymbol{I})\boldsymbol{\theta} = \boldsymbol{H}\hat{\boldsymbol{\theta}}$$
$$\hat{\boldsymbol{\theta}}_{\text{Ridge}} = (\boldsymbol{H} + \lambda\boldsymbol{I})^{-1}\boldsymbol{H}\hat{\boldsymbol{\theta}}$$

where $\boldsymbol{I}$ is the identity matrix. As λ approaches 0, the regularized

solution $\hat{\boldsymbol{\theta}}_{\text{Ridge}}$ approaches $\hat{\boldsymbol{\theta}}$. What happens as λ grows?

L2-REGULARIZATION / WEIGHT DECAY

- Because $\boldsymbol{H}$ is a real symmetric matrix, it can be decomposed as $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top$ where $\boldsymbol{\Sigma}$ is a diagonal matrix of eigenvalues and $\boldsymbol{Q}$ is an orthonormal basis of eigenvectors.
- Rewriting the equation on the previous slide using the eigendecomposition above,

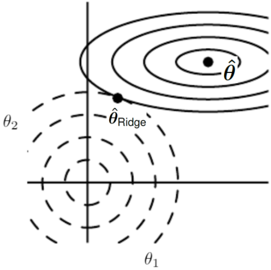
$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{Ridge}} &= \left(\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top + \lambda\boldsymbol{I}\right)^{-1} \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\hat{\boldsymbol{\theta}} \\ &= \left[\boldsymbol{Q}(\boldsymbol{\Sigma} + \lambda\boldsymbol{I})\boldsymbol{Q}^\top\right]^{-1} \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\hat{\boldsymbol{\theta}} \\ &= \boldsymbol{Q}(\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\hat{\boldsymbol{\theta}}\end{aligned}$$

- Therefore, the weight decay rescales $\hat{\boldsymbol{\theta}}$ along the axes defined by the eigenvectors of $\boldsymbol{H}$. The component of $\hat{\boldsymbol{\theta}}$ that is aligned with the i-th eigenvector of $\boldsymbol{H}$ is rescaled by a factor of $\frac{\sigma_i}{\sigma_i+\lambda}$, where σ_i is the corresponding eigenvalue.

L2-REGULARIZATION / WEIGHT DECAY

- Along directions where the eigenvalues of $\boldsymbol{H}$ are relatively large, for example, where $\sigma_i >> \lambda$, the effect of regularization is quite small.
- On the other hand, components with $\sigma_i << \lambda$ will be shrunk to have nearly zero magnitude.
- In other words, only directions along which the parameters contribute significantly to reducing the objective function are preserved relatively intact.
- In the other directions, a small eigenvalue of the Hessian means that moving in this direction will not significantly increase the gradient. For such unimportant directions, the corresponding components of $\boldsymbol{\theta}$ are decayed away.

L2-REGULARIZATION / WEIGHT DECAY

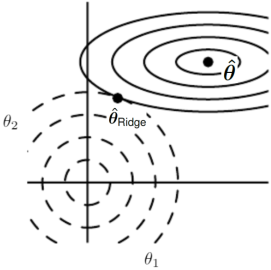


Credit: Goodfellow et al. (2016), ch. 7

Figure: The solid ellipses represent the contours of the unregularized objective and the dashed circles represent the contours of the L2 penalty. At $\hat{\boldsymbol{\theta}}_{\text{Ridge}}$, the competing objectives reach an equilibrium.

In the first dimension, the eigenvalue of the Hessian of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is small. The objective function does not increase much when moving horizontally away from $\hat{\boldsymbol{\theta}}$. Therefore, the regularizer has a strong effect on this axis and θ_1 is pulled close to zero.

L2-REGULARIZATION / WEIGHT DECAY



Credit: Goodfellow et al. (2016), ch. 7

Figure: The solid ellipses represent the contours of the unregularized objective and the dashed circles represent the contours of the L2 penalty. At $\hat{\boldsymbol{\theta}}_{\text{Ridge}}$, the competing objectives reach an equilibrium.

In the second dimension, the corresponding eigenvalue is large indicating high curvature. The objective function is very sensitive to movement along this axis and, as a result, the position of θ_2 is less affected by the regularization.

L1-REGULARIZATION

- The L1-regularized risk of a model $f(\mathbf{x} \mid \boldsymbol{\theta})$ is

$$\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1$$

  and the (sub-)gradient is:

$$\nabla_{\theta} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \lambda \operatorname{sign}(\boldsymbol{\theta}) + \nabla_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

- Note that, unlike in the case of L2, the contribution of the L1 penalty to the gradient doesn't scale linearly with each θ_i. Instead, it is a constant factor with a sign equal to $\operatorname{sign}(\theta_i)$.
- Let us now make a quadratic approximation of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$. To get a clean algebraic expression, we assume the Hessian of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is diagonal, i.e. $\boldsymbol{H} = \operatorname{diag}([H_{1,1}, \ldots, H_{n,n}])$, where each $H_{i,i} > 0$.
- This assumption holds, for example, if the input features for a linear regression task have been decorrelated using PCA.

L1-REGULARIZATION

- The quadratic approximation of $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ decomposes into a sum over the parameters:

$$\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) + \sum_i \left[\frac{1}{2} H_{i,i}(\theta_i - \hat{\theta}_i)^2\right] + \sum_i \lambda |\theta_i|$$

  where $\hat{\boldsymbol{\theta}}$ is the minimizer of the unregularized risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$.
- The problem of minimizing this approximate cost function has an analytical solution (for each dimension i), with the following form:

$$\hat{\theta}_{\text{Lasso},i} = \text{sign}(\hat{\theta}_i) \max\left\{|\hat{\theta}_i| - \frac{\lambda}{H_{i,i}}, 0\right\}$$

- If $0 < \hat{\theta}_i \leq \frac{\lambda}{H_{i,i}}$, the optimal value of θ_i (for the regularized risk) is 0 because the contribution of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ to $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ is overwhelmed by the L1 penalty, which forces it to be 0.

L1-REGULARIZATION

- If $0 < \frac{\lambda}{H_{i,i}} < \hat{\theta}_i$, the $L1$ penalty shifts the optimal value of θ_i toward 0 by the amount $\frac{\lambda}{H_{i,i}}$.
- A similar argument applies when $\hat{\theta}_i < 0$.
- Therefore, the L1 penalty induces sparsity in the parameter vector.

EQUIVALENCE TO CONSTRAINED OPTIMIZATION

Norm penalties can be interpreted as imposing a constraint on the weights. One can show that

$$\arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda J(\boldsymbol{\theta})$$

is equvilalent to

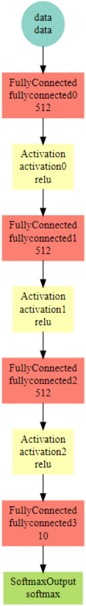
$$\begin{aligned} &\arg\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) \\ &\text{subject to} \quad J(\boldsymbol{\theta}) \leq k \end{aligned}$$

for some value k that depends on λ the nature of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$.

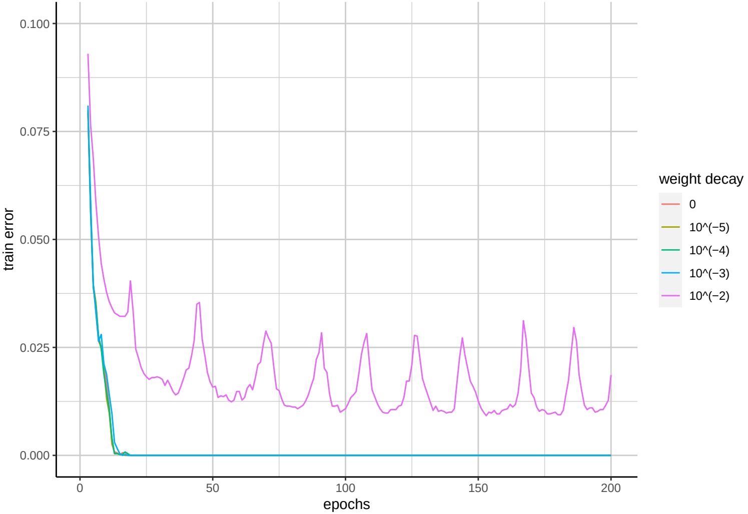
(Goodfellow et al. (2016), ch. 7.2)

EXAMPLE: WEIGHT DECAY

- We fit the huge neural network on the right side on a smaller fraction of MNIST (5000 train and 1000 test observations)
- Weight decay: $\lambda \in (10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0)$

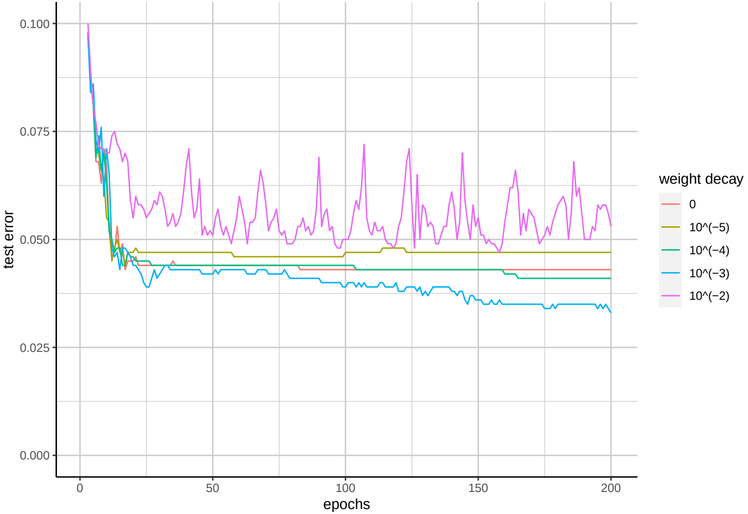


EXAMPLE: WEIGHT DECAY



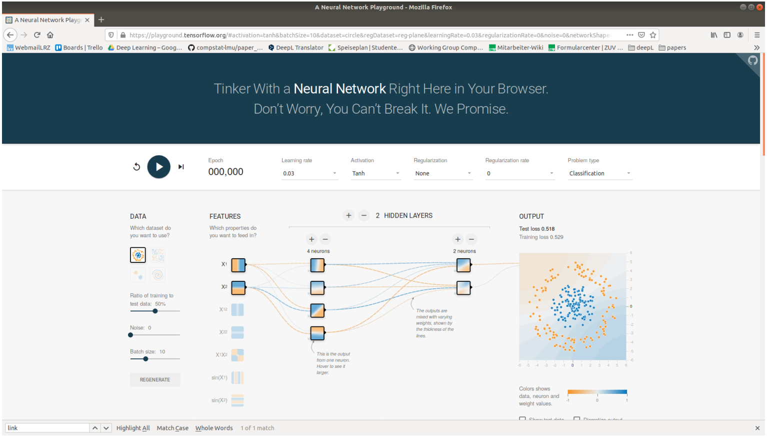
A high weight decay of 10^{-2} leads to a high error on the training data.

EXAMPLE: WEIGHT DECAY



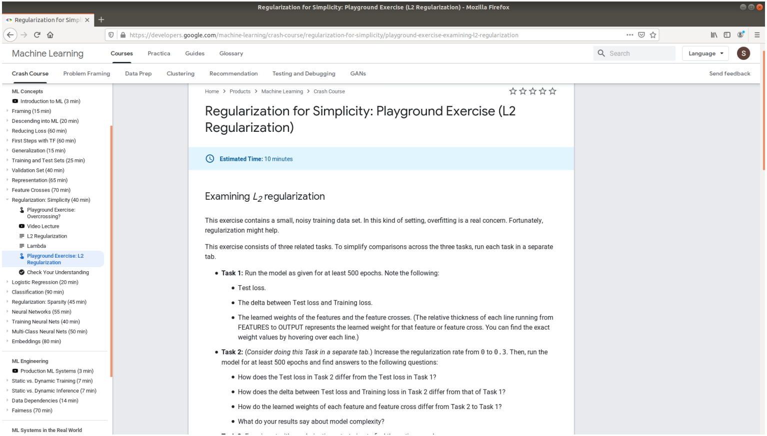
Second strongest weight decay leads to the best result on the test data.

TENSORFLOW PLAYGROUND



https://playground.tensorflow.org/

TENSORFLOW PLAYGROUND - EXERCISE



https://developers.google.com/machine-learning/crash-course/regularization-for-simplicity/playground-exercise-examining-l2-regularization