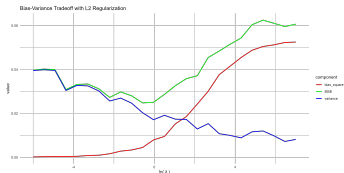


Introduction to Machine Learning

Regularization

Intuition for L2 Regularization in Non-Linear Models



Learning goals

- Understand how regularization and parameter shrinkage can be beneficial to non-linear models

SUMMARY: REGULARIZED RISK MINIMIZATION

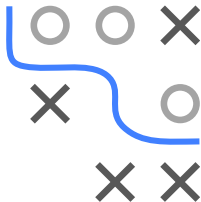
If we should define (supervised) ML in only one line, this might be it:

$$\min_{\theta} \mathcal{R}_{\text{reg}}(\theta) = \min_{\theta} \left(\sum_{i=1}^n L(y^{(i)}, f(\mathbf{x}^{(i)} | \theta)) + \lambda \cdot J(\theta) \right)$$

We can choose for a task at hand:

- the **hypothesis space** of f , which determines how features can influence the predicted y
- the **loss** function L , which measures how errors should be treated
- the **regularization** $J(\theta)$, which encodes our inductive bias and preference for certain simpler models

By varying these choices one can construct a huge number of different ML models. Many ML models follow this construction principle or can be interpreted through the lens of regularized risk minimization.



REGULARIZATION IN NEURAL NETWORKS

For neural networks, the regularized loss function is:

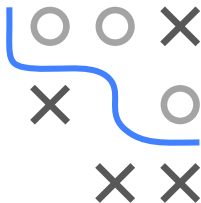
$$\mathcal{R}_{\text{reg}}(\theta) = \frac{1}{n} \sum_{i=1}^n L\left(y^{(i)}, f(\mathbf{x}^{(i)} \mid \theta)\right) + \lambda \cdot J(\theta)$$

where:

- $L(f(x_i; \theta), y_i)$ is the loss function.
- $f(x_i; \theta)$ is the neural network's prediction.
- $J(\theta)$ is the regularization term (e.g., $\|\theta\|_2^2$ for L2 regularization).
- λ is the regularization parameter.

Bias: Regularization increases bias because it adds a constraint on the network parameters, preventing them from fitting the training data perfectly.

Variance: Regularization decreases variance by limiting the network parameters' magnitudes, reducing sensitivity to the training data's noise.



DERIVING THE BOUND FOR VARIANCE OF NEURAL NETWORK PREDICTIONS

To derive the bound for the variance of the parameter estimates in a neural network with L2 regularization, we follow these steps:

Neural Network with L2 Regularization: The regularized loss function is:

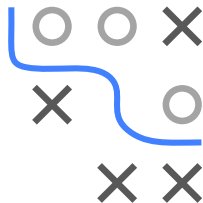
$$\mathcal{R}_{\text{reg}}(\theta) = \frac{1}{n} \sum_{i=1}^n L(y^{(i)}, f(\mathbf{x}^{(i)} | \theta)) + \lambda \|\theta\|_2^2$$

Bias-Variance Decomposition: The mean squared error (MSE) decomposition is:

$$E[(\hat{y} - y)^2] = \text{Bias}^2(\hat{y}) + \text{Var}(\hat{y}) + \sigma^2$$

Step-by-Step Derivation:

- Model the Neural Network Parameters: $\hat{\theta} = \theta^* + \epsilon$



BIAS ANALYSIS IN NEURAL NETWORKS

To analyze the bias term:

Bias Term: Regularization introduces bias by shrinking the parameter estimates towards zero:

$$\text{Bias}(f(x)) = E[f(x; \hat{\theta}_{\text{Reg}})] - f^*(x)$$

Using a linear approximation:

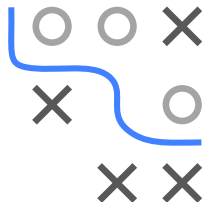
$$E[f(x; \hat{\theta}_{\text{Reg}})] \approx f(x; \theta^*) - \lambda \nabla_{\theta} f(x; \theta^*)^T H^{-1} \theta^*$$

Thus, the bias is:

$$\text{Bias}(f(x)) = -\lambda \nabla_{\theta} f(x; \theta^*)^T H^{-1} \theta^*$$

Combined Bias and Variance Analysis:

- **Bias:** $\text{Bias}^2(f(x)) = (\lambda \nabla_{\theta} f(x; \theta^*)^T H^{-1} \theta^*)^2$
- **Variance:** $\text{Var}(f(x; \hat{\theta}_{\text{Reg}})) \leq \frac{\sigma^2}{2\lambda} \|\nabla_{\theta} f(x; \hat{\theta}_{\text{Reg}})\|^2$



REDUCTION IN VARIANCE VS. INCREASE IN BIAS

To show that the reduction in variance is usually more than the increase in bias, consider:

Bias-Variance Trade-off: The MSE is decomposed as:

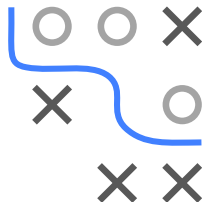
$$\text{MSE} = \text{Bias}^2(f(x)) + \text{Var}(f(x)) + \sigma^2$$

Change in Bias and Variance:

- **Change in Bias:** $\Delta \text{Bias}^2 \propto \lambda^2$
- **Change in Variance:** $\Delta \text{Var} \propto -\frac{1}{\lambda}$

For small λ , the reduction in variance is significant, while the increase in bias is relatively small. The reduction in variance usually outweighs the increase in bias, leading to an overall decrease in MSE.

Conclusion: Regularization helps in reducing the overall prediction error by balancing the bias and variance effectively.



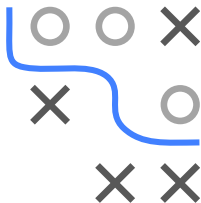
CRITIQUE: BIAS-VARIANCE TRADEOFF AND OPTIMIZATION

For linear models, it's well-established that some $\lambda > 0$ can balance the increase in bias against the reduction in variance, leading to a net decrease in MSE. For non-linear models, the situation is more complex:

- The relationship between model parameters θ , the regularization term, and the model output $f(x; \theta)$ is non-linear.
- The effects of changing λ on the bias and variance terms are not straightforward and depend heavily on the specific form of the non-linear model and the data distribution.

Proving analytically that there exists a $\lambda > 0$ such that the regularized model always outperforms the unregularized model in terms of MSE for general non-linear models involves:

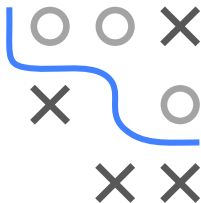
- Detailed understanding of how changes in λ affect the bias and variance for the specific type of non-linear model.



CRITIQUE: CONCLUSION

In summary, while it is conceptually feasible to argue that an appropriate $\lambda > 0$ might improve the MSE by balancing bias and variance, providing a universal, formal proof for all non-linear models would require either restrictive assumptions about the models and data or a very specific setup where the non-linearities are well understood and mathematically tractable.

For practical purposes, empirical validation through techniques such as cross-validation remains a critical method to determine the optimal λ for specific non-linear models and datasets.



COUNTEREXAMPLE

Chris: I think ChatGPT produced a lot of "almost correct" stuff that culminated in a globally useless derivation. A general proof for DNNs imo can not work by giving a simple counterexample.

- A diagonal linear network with one hidden layer and one output unit can be written as $f(x|\mathbf{u}, \mathbf{v}) = (\mathbf{u} \odot \mathbf{v})^\top \mathbf{x}$
- optimizing the network with $L2$ regularization λ and MSE loss has multiple global minima that coincide with the lasso solution for the collapsed parameter $\boldsymbol{\theta} := \mathbf{u} \odot \mathbf{v}$ using 2λ
- Since there is no existence theorem (of a λ^* that reduces the MSE over OLS) for lasso compared to ridge regression, there can not be one for $L2$ regularized DNNs in general.

