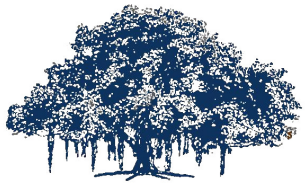
Bias Variance Trade-Off, Regularization, Early Stopping, Dropout

Naresh Manwani

Machine Learning Lab, IIIT Hyderabad



INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

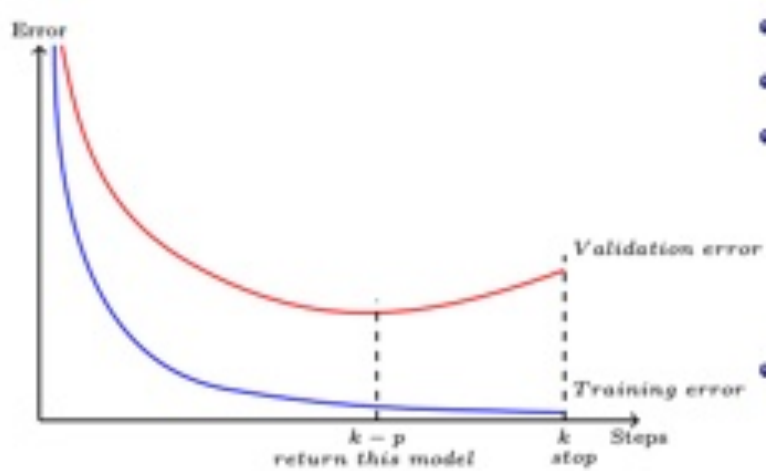
Early Stopping

Early Stopping

1. When training models with sufficient representational capacity to overfit the task, we often observe that training error decreases steadily over time, while the error on the validation set begins to rise again.
2. The occurrence of this behaviour in the scope of our applications is almost certain.
3. This means we can obtain a model with better validation set error (and thus,hopefully better test set error) by returning to the parameter setting at the point in time with the lowest validation set error.
4. This is termed **Early Stopping**.

Early Stopping: Implementation

1. Track the validation error. Have a patience parameter p
2. If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$
3. Basically, stop the training early before it drives the training error to 0 and blows up the validation error



Early Stopping: Pros

1. Early Stopping is probably one of the most used regularization strategies in deep learning.
2. Early stopping can be thought of as a hyperparameter selection method, where training time is the hyperparameter to be chosen.
3. Choosing the training time automatically can be done through a single run through the training phase, the only addition being the evaluation of the validation set error at every n iterations. This is usually done on a second GPU.
4. Overhead for writing parameters to disk is negligible.

Early Stopping As Regularization

1. How does it act as a regularizer ?
2. We will first see an intuitive explanation and then a mathematical analysis

Early Stopping As Regularization

1. Recall that the update rule in SGD is

$$\begin{aligned}\mathbf{w}^{t+1} &= \mathbf{w}^t - \eta \Delta \mathbf{w}^t \\ &= \mathbf{w}^0 - \eta \sum_{i=0}^{t} \Delta \mathbf{w}^i\end{aligned}$$

2. Let τ be the maximum value of $\|\Delta \mathbf{w}^t\|_2$ for any t, then $\|\mathbf{w}^t - \mathbf{w}^0\|_2 = \| - \eta \sum_{i=0}^{t-1} \Delta \mathbf{w}^i\|_2 \leq \eta \sum_{i=0}^{t-1} \|\Delta \mathbf{w}^i\|_2 \leq \eta t \tau$.
3. Thus, τ controls how far $\mathbf{w}^t$ can go from the initial $\mathbf{w}^0$. In other words, it controls the space of exploration.

Early Stopping

1. Let $\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$.
2. 2nd order Taylor series approximation of $\mathcal{L}(\mathbf{w})$ around $\mathbf{w}^*$ is as follows.

$$\begin{aligned}\hat{\mathcal{L}}(\mathbf{w}) &= \mathcal{L}(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^T\nabla\mathcal{L}(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^*)^T H(\mathbf{w}-\mathbf{w}^*) \\ &= \mathcal{L}(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^*)^T H(\mathbf{w}-\mathbf{w}^*), \qquad (\because \nabla\mathcal{L}(\mathbf{w}^*) = \mathbf{0}) \\ \nabla\hat{\mathcal{L}}(\mathbf{w}) &= H(\mathbf{w}-\mathbf{w}^*)\end{aligned}$$

   where H is hessian of $\mathcal{L}(\mathbf{w})$ evaluated at $\mathbf{w}^*$.
3. Now the SGD on $\hat{\mathcal{L}}$ gives us the following update rule.

$$\begin{aligned}\mathbf{w}^t &= \mathbf{w}^{t-1} - \eta\nabla\hat{\mathcal{L}}(\mathbf{w}^{t-1}) \\ &= \mathbf{w}^{t-1} - H(\mathbf{w}^{t-1} - \mathbf{w}^*) \\ &= (I - \eta H)\mathbf{w}^{t-1} + \eta H\mathbf{w}^*\end{aligned}$$

Early Stopping

1. $\mathbf{w}^t = (I - \eta H)\mathbf{w}^{t-1} + \eta H \mathbf{w}^*$
2. Using EVD of H as $H = Q\Lambda Q^T$, we get:

$$\mathbf{w}^t = (I - \eta Q\Lambda Q^T)\mathbf{w}^{t-1} + \eta Q\Lambda Q^T \mathbf{w}^*$$

3. If we start with $\mathbf{w}^0 = \mathbf{0}$ then we can show that $\mathbf{w}^t = Q[I - (I - \eta\Lambda)^t]Q^T\mathbf{w}^*$

Theorem

The below two equations are equivalent

$$\mathbf{w}^t = (I - \eta Q\Lambda Q^T)\mathbf{w}^{t-1} + \eta Q\Lambda Q^T \mathbf{w}^*$$
$$\mathbf{w}^t = Q[I - (I - \eta\Lambda)^t]Q^T\mathbf{w}^*$$

Proof of the Theorem

1. Proof by induction:
2. Base Case: $t = 1$ and $\mathbf{w}^0 = \mathbf{0}$
   - $\mathbf{w}^1$ according to the first equation:
     $\mathbf{w}^1 = (I - \eta Q \Lambda Q^T)\mathbf{w}^0 + \eta Q \Lambda Q^T \mathbf{w}^* = \eta Q \Lambda Q^T \mathbf{w}^*$
   - $\mathbf{w}^1$ according to the second equation:
     $\mathbf{w}^1 = Q[I - (I - \eta\Lambda)^1]Q^T \mathbf{w}^* = \eta Q \Lambda Q^T \mathbf{w}^*$
3. Induction Step: Let the two equations be equivalent for t^{th} step

$$\mathbf{w}^t = (I - \eta Q \Lambda Q^T)\mathbf{w}^{t-1} + \eta Q \Lambda Q^T \mathbf{w}^* = Q[I - (I - \eta\Lambda)^t]Q^T \mathbf{w}^*$$

Proof Continue

- Proof that the statement holds for $t+1$ also. Using $\mathbf{w}^t = Q[I-(I-\eta\Lambda)^t]Q^T\mathbf{w}^*$, we get

$$
\begin{aligned}
\mathbf{w}^{t+1} &= (I-\eta Q\Lambda Q^T)\mathbf{w}^t + \eta Q\Lambda Q^T\mathbf{w}^* \\
&= (I-\eta Q\Lambda Q^T)Q[I-(I-\eta\Lambda)^t]Q^T\mathbf{w}^* + \eta Q\Lambda Q^T\mathbf{w}^* \\
&= IQ[I-(I-\eta\Lambda)^t]Q^T\mathbf{w}^* - \eta Q\Lambda Q^T Q[I-(I-\eta\Lambda)^t]Q^T\mathbf{w}^* \\
&\quad + \eta Q\Lambda Q^T\mathbf{w}^* \\
&= Q[I-(I-\eta\Lambda)^t]Q^T\mathbf{w}^* - \eta Q\Lambda[I-(I-\eta\Lambda)^t]Q^T\mathbf{w}^* \\
&\quad + \eta Q\Lambda Q^T\mathbf{w}^* \\
&= Q\left[I-(I-\eta\Lambda)^t - \eta\Lambda[I-(I-\eta\Lambda)^t] + \eta\Lambda\right]Q^T\mathbf{w}^* \\
&= Q\left[I-(I-\eta\Lambda)^t + \eta\Lambda(I-\eta\Lambda)^t\right]Q^T\mathbf{w}^* \\
&= Q\left[I-(I-\eta\Lambda)^t(I-\eta\Lambda)\right]Q^T\mathbf{w}^* \\
&= Q\left[I-(I-\eta\Lambda)^{t+1}\right]Q^T\mathbf{w}^*
\end{aligned}
$$

Early Stopping

1. Compare this with the expression we had for optimum $\tilde{\mathbf{w}}$ with l_2 regularization $\tilde{\mathbf{w}} = Q(\Lambda + \alpha I)^{-1}\Lambda Q^T \mathbf{w}^*$.
2. We observe that $\mathbf{w}^t = \tilde{\mathbf{w}}$, if we choose η, t and α such that

$$I - (I - \eta\Lambda)^t = (\Lambda + \alpha I)^{-1}\Lambda$$

3. Early stopping only allows t updates to the parameters.
4. If a parameter w corresponds to a dimension which is important for the loss $\mathcal{L}$ then $\frac{\partial \mathcal{L}}{\partial w}$ will be large.
5. However if a parameter is not important ($\frac{\partial \mathcal{L}}{\partial w}$ is small) then its updates will be small and the parameter will not be able to grow large in 't steps.
6. Early stopping will thus effectively shrink the parameters corresponding to less important directions (same as weight decay).