Large Scale Data Analysis Using Deep Learning

Regularization for Deep Learning

U Kang
Seoul National University

In This Lecture

- Regularization
  - Motivation
  - Norm penalties and their characteristics
  - Dataset augmentation
  - Multi-task learning
  - Early stopping

Definition

- A central problem in ML is how to make an algorithm that will generalize well
- Regularization: any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error
- Most regularization strategies in deep learning are based on regularizing estimators, by trading increased bias for reduced variance
  - An effective regularizer is one that makes a profitable trade, reducing variance significantly while not overly increasing the bias

Parameter Norm Penalties

- Limit the capacity of models (e.g. neural networks, linear regression, or logistic regression) by adding a parameter norm penalty $\Omega(\theta)$ to the objective function J
  - $\tilde{J}(\Theta; X, y) = J(\Theta; X, y) + \alpha\Omega(\Theta)$ where $\alpha \in [0, \infty)$ is a hyperparameter that weights the contribution of Ω
  - Small α means less regularization; large α means more regularization
  - Most common forms: L^2 and L^1 parameter regularizations

L² Parameter Regularization

- Drives the weights closer to origin by adding a regularization term $\Omega(\Theta) = \frac{1}{2}||w||_2^2$ to the objective function
- Also known as ridge regression, Tikhonov regularization, or weight decay
- $\tilde{J}(w; X, y) = \frac{\alpha}{2} w^T w + J(w; X, y)$
- $\nabla_w \tilde{J}(w; X, y) = \alpha w + \nabla_w J(w; X, y)$
- $w \leftarrow w - \epsilon\big(\alpha w + \nabla_w J(w; X, y)\big) = (1 - \epsilon\alpha)w - \epsilon\nabla_w J(w; X, y)$
- This means adding the weight decay term shrinks the weight vector by a constant factor on each step, just before performing the gradient update

L^2 Parameter Regularization

- Further analysis by making a quadratic approximation to the objective function in the neighborhood of the optimal value $w^* = argmin_w J(w)$
  - Note that quadratic approximation of f(y) around x is given by $f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T H(y - x)$
  - $\hat{J}(w) = J(w^*) + \frac{1}{2}(w - w^*)^T H(w - w^*)$ where H is the Hessian matrix of J with respect to w evaluated at w^*
  - H is positive semidefinite since w^* is the location of a minimum of J
  - The minimum of $\hat{J}$ occurs where the gradient $\nabla_w \hat{J}(w) = H(w - w^*)$ equals 0
  - Let $\tilde{w}$ be the location of the minimum of the function $\hat{J}(w) + \frac{\alpha}{2} w^T w$
  - Then, $\alpha\tilde{w} + H(\tilde{w} - w^*) = 0 \quad \leftrightarrow \quad (H + \alpha I)\tilde{w} = Hw^*$
  - Thus, $\tilde{w} = (H + \alpha I)^{-1} H w^*$

L² Parameter Regularization

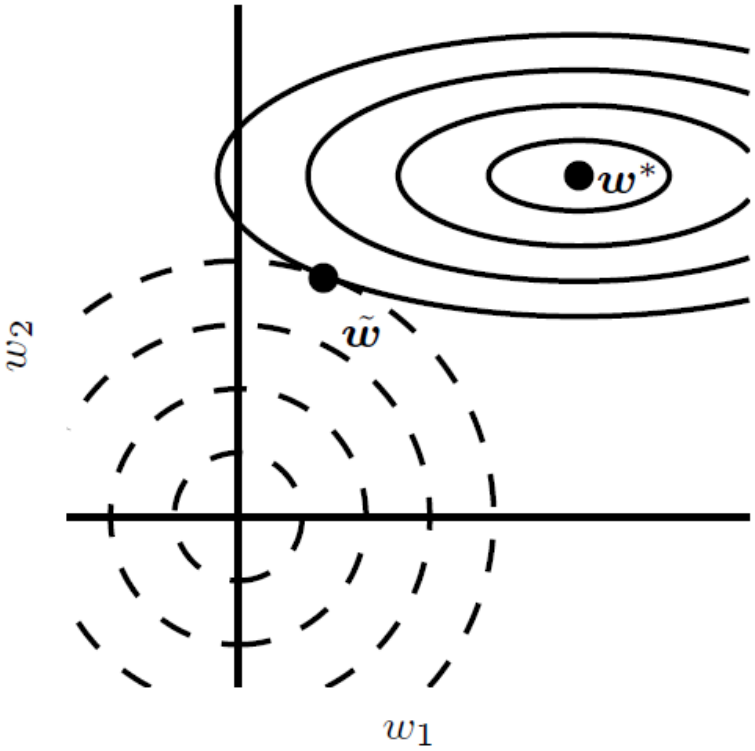
- Using eigendecomposition $H = Q\Lambda Q^T$

$$\tilde{w} = (H + \alpha I)^{-1} H w^* = (Q\Lambda Q^T + \alpha I)^{-1} Q\Lambda Q^T w^*$$
$$= [Q(\Lambda + \alpha I)Q^T]^{-1} Q\Lambda Q^T w^* = Q(\Lambda + \alpha I)^{-1}\Lambda Q^T w^*$$

- The effect of weight decay is to rescale w^* along the axes defined by the eigenvectors of H
- The component of w^* that is aligned with the i-th eigenvector of H is rescaled by a factor of $\frac{\lambda_i}{\lambda_i + \alpha}$
- Along the directions where eigenvalues of H are relatively large, e.g. $\lambda_i \gg \alpha$, the effect of regularization is relatively small
- Components with $\lambda_i \ll \alpha$ will be shrunk to have nearly zero magnitude

L^2 Parameter Regularization

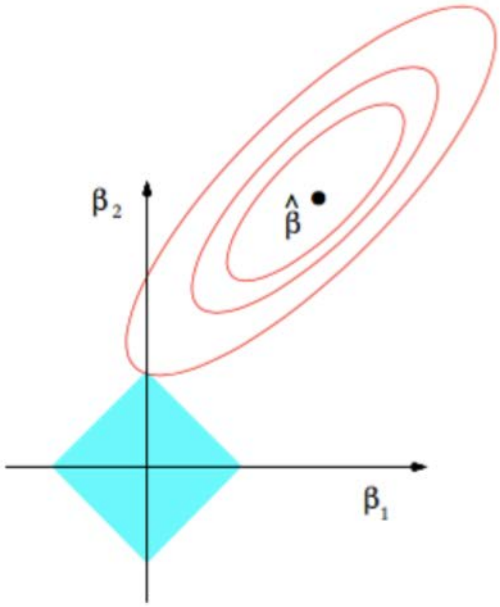
- The regularization shrinks w* more along the direction where the objective function does not decrease significantly
  - That is, along the direction which is an eigenvector of H with a small eigenvalue (the second derivative along the direction of an eigenvector q_i (with eigenvalue λ_i) is given by ${q_i}^T H q_i = \lambda_i$)

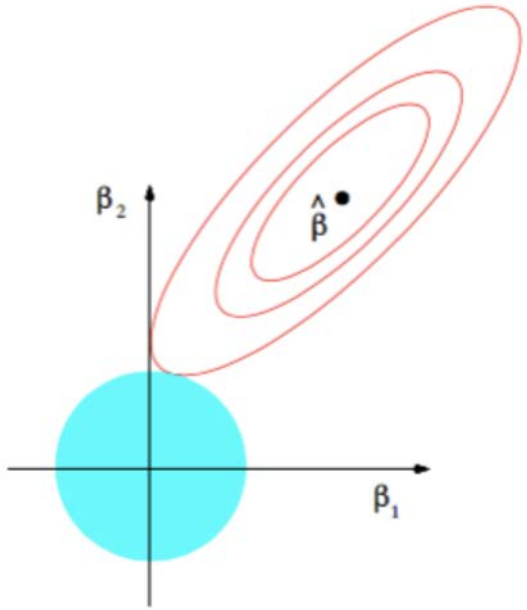


L^1 Parameter Regularization

- Use regularization term $\Omega(\Theta) = ||w||_1 = \sum_i |w_i|$
- L^1 regularization results in a solution that is more sparse
  - Some parameters have an optimal value of 0
- L^1 regularization has been used extensively as a feature selection mechanism
  - LASSO: least square with L^1 regularization term

L^1 regularization





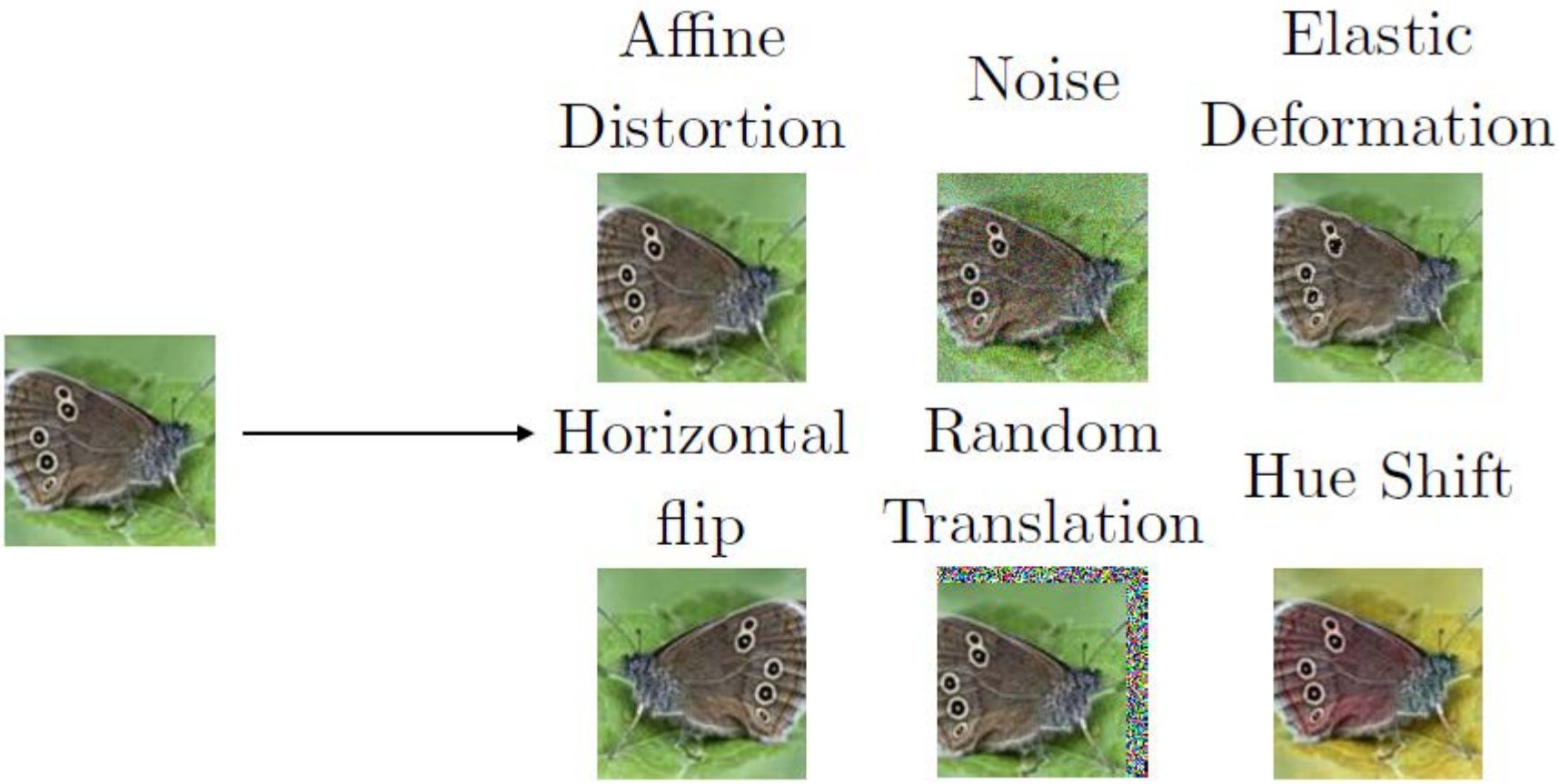
L^2 regularization

[The Elements of Stat. Learning by Hastie et al.]

Dataset Augmentation

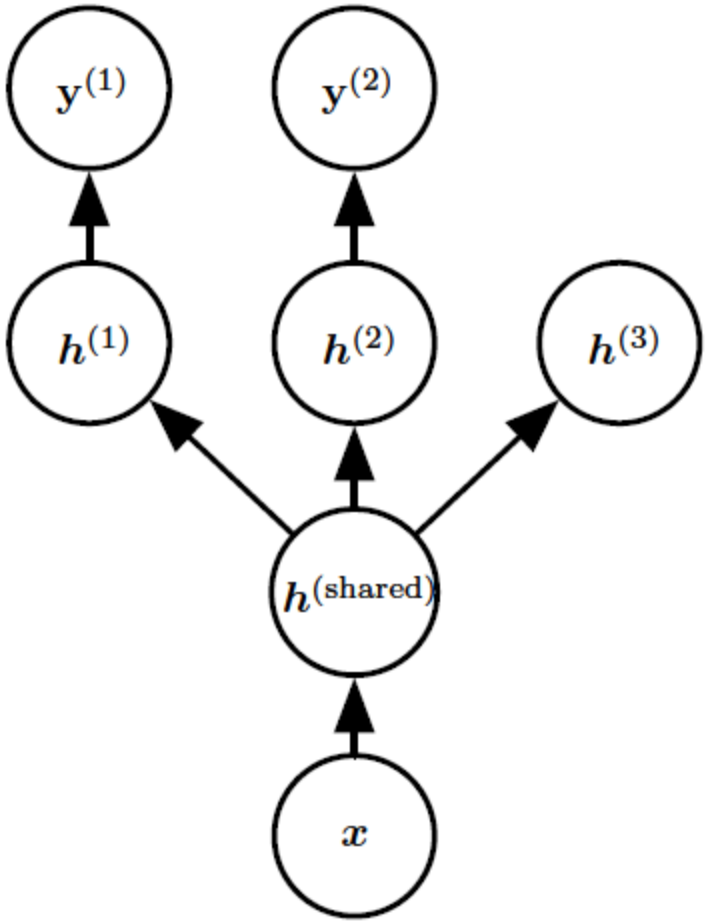
- The best way to make a machine learning model generalize better is to train it on more data
- Dataset augmentation
  - Create fake data and add it to the training set
  - Has been effective especially for object recognition
- Image augmentation
  - Translation
  - Rotation
  - Scaling
- Injecting noise: a form of data augmentation
  - Neural networks are not very robust to noise; one way to improve the robustness of neural networks is to train them with random noise applied to their inputs

Dataset Augmentation



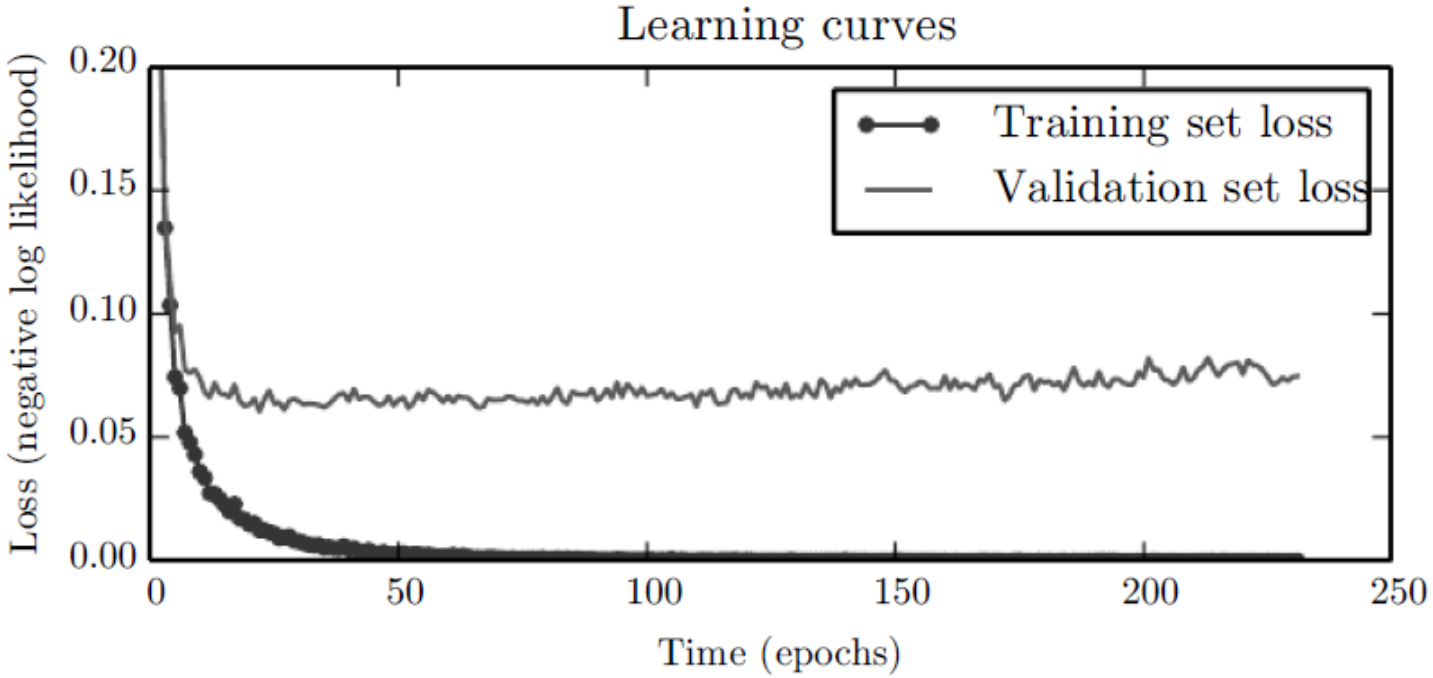
Multi-Task Learning

- Multi-task learning is a way to improve generalization by sharing the parameters arising out of several tasks
- Main idea: among the factors that explain the variations observed in the data associated with the different tasks, some are *shared* across two or more tasks
- The lower layers of a deep network can be shared across tasks, while specific parameters (connected to $\boldsymbol{h}^{(1)}$, $\boldsymbol{h}^{(2)}$, and $\boldsymbol{h}^{(3)}$) can be learned on top of those yielding a shared representation $\boldsymbol{h}^{(shared)}$



Early Stopping

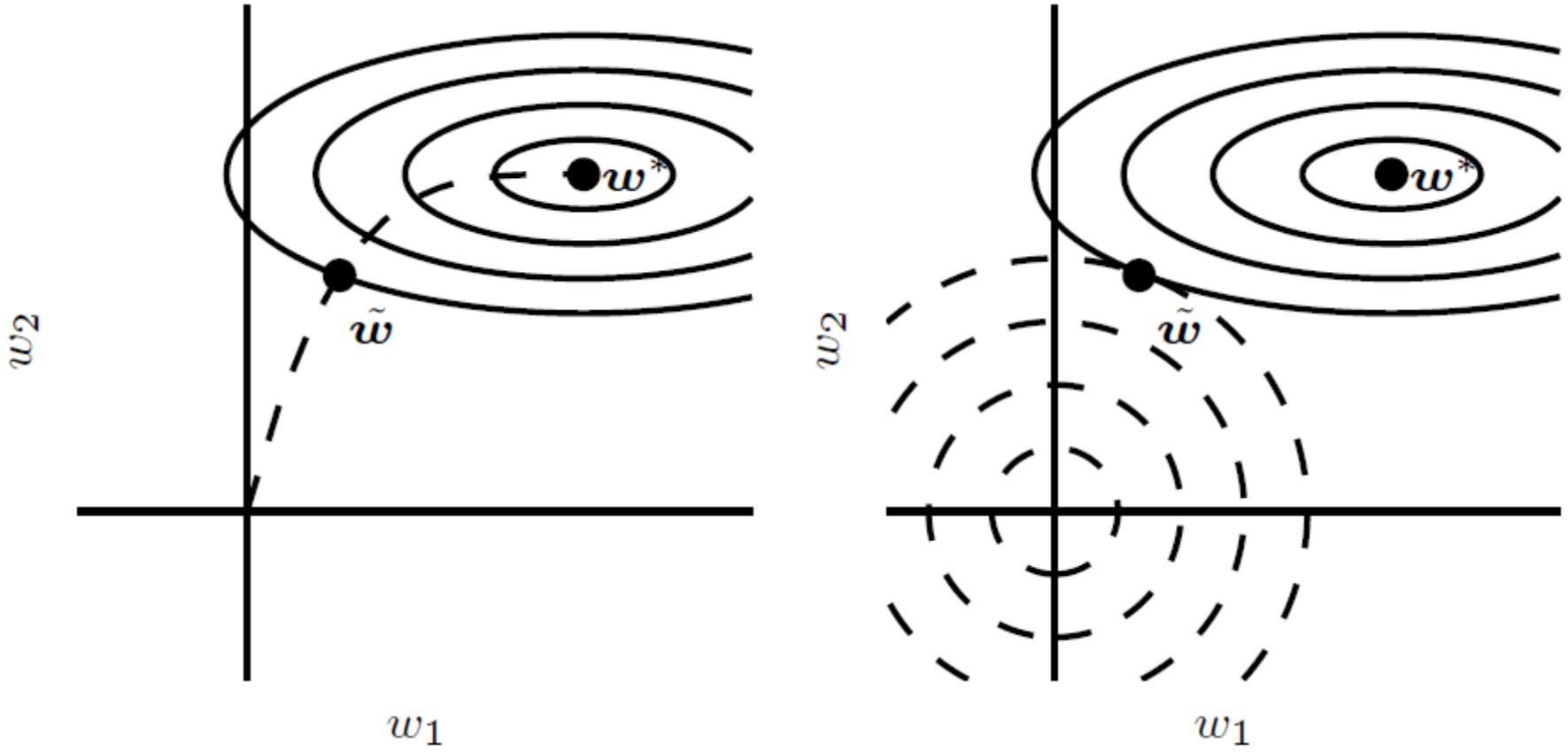
- When training large models with sufficient representational capacity to overfit the task, we often observe that training error decreases steadily over time, but validation set error begins to rise again.
- Early stopping: obtain a model with better validation set error (and thus hopefully better test error) by returning to the parameter setting at the point in time with the lowest validation set error



Early Stopping

- Early stopping is a very unobtrusive form of regularization, meaning that it is easy to apply early stopping to any ML algorithm
- Early stopping may be used either alone or in conjunction with other regularization strategies
- Early stopping also reduces the computational cost of the training procedure

Early Stopping and Weight Decay



What you need to know

- Regularization
  - Motivation
    - Make an algorithm that will generalize well
  - Norm penalties and their characteristics
    - L^2: shrinks parameters more along the direction where the objective function does not decrease significantly
    - L^1: lead to sparse solutions
  - Dataset augmentation
  - Multi-task learning
  - Early stopping
    - General and widely applicable approach

Questions?