Overfitting and Solutions

- Overfitting: Model performs well on training data, but poorly on test data
- Causes
- Solutions:
  - Cross-Validation
  - Proper Validation Set
  - Regularization (Bias-Variance tradeoff)
    1. Ridge: $f(\theta) = L(\theta) + \lambda\|\theta\|_2^2$
    2. Lasso: $f(\theta) = L(\theta) + \lambda\|\theta\|_1^2$
    3. Group lasso: $f(\theta) = L(\theta) + \lambda \sum_{g=1}^{G} \|\beta_g\|_2$

Proximal Gradient Descent

- Designed for optimization problems of the form:

$$\min_x F(x) = f(x) + g(x), \quad x \in \mathbb{R}^n$$

  where f and g has global minimum, f is differentiable, g is not differentiable.

- The proximal operator

$$\text{prox}_{t,g}(z) = \arg\min_x \left\{ g(x) + \frac{1}{2t}\|x - z\|_2^2 \right\}$$

- Proximal gradient descent algorithm:
  1. Regular GD update: $y_k = x_k - \eta_k \nabla f\left(x_k\right)$
  2. Proximal based update: $x_{k+1} = \text{prox}_{\eta_k,g}\left(y_k\right)$
- The proximal operator for ℓ_1 norm is the soft-thresholding function: $\text{prox}_{t,\lambda\|\cdot\|_1}(z) = S_{t,\lambda}(z) = \text{sign}(z) \cdot \max(|z| - t\lambda, 0)$

Newton's Algorithm and Variants

- Intuition: Take the curvature of the target function into consideration.
- Update rule: $x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$.
- Damped/guarded version of Newton's algorithm:
  $x_{k+1} = x_k - \eta_k \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$.
- Issues with the Newton algorithm:
  1. Hessian is not strictly positive definite or Hessian is ill-conditioned.
     Levenberg-Marquardt algorithm: $x_{k+1} = x_k - \left[\nabla^2 f(x_k) + \mu_k I\right]^{-1} \nabla f(x_k)$.
  2. Expensive computation cost to calculate the inverse of the Hessian matrix.
     - Use $H_0 = \nabla^2 f(x_0)$ or update the Hessian every ℓ iterations.
     - Use $\tilde{H}_k \equiv \operatorname{diag}\left(\frac{\partial^2 f(x_k)}{(\partial x_i)^2}\right)_{i=1}^{n}$.

Coordinate Descent Algorithm

- Coordinate Descent updates one coordinate (parameter) at a time while keeping others fixed.
- Widely used for high dimensional problems where full gradient methods can be expensive.
- Coordinate descent for linear/ridge/lasso regression.
- Advantages and Limitations