Regularization and Stability

Machine Learning 2022-23
UML book chapter 13
Regularized Loss Minimization (RLM)

Key idea: jointly minimize empirical risk and a regularization function

- **Hypothesis** $h$: defined by a vector $\mathbf{w} = (w_1, \ldots, w_d)^T \in \mathbb{R}^d$
  - e.g., coefficients of a linear model, weights in a neural network, etc..
- **Regularization function** $R: \mathbb{R}^d \rightarrow \mathbb{R}$, function of $\mathbf{w}$
- **Regularized Loss Minimization (RLM)**: select $h$ from:

  $$\text{argmin}_{\mathbf{w}} (L_s(\mathbf{w}) + R(\mathbf{w}))$$

- $L_s(\mathbf{w})$: standard loss for the considered problem
- $R(\mathbf{w})$: regularization term (measures in some way the "complexity" of the found solution)

- The regularization term balances between low empirical risk and aiming at less complex hypotheses
- It is possible to view the extra term as a "stabilizer"
Tikhonov Regularization

- Define function $R$ using the $\|w\|^2$ norm of the weights:
  \[ R(w) = \lambda \|w\|^2 = \lambda \sum_{i=1}^{d} w_i^2 \]
- Output of function $R$ is a real positive number
- Learning Rule: $A(s) = \text{argmin}_w (L_s(w) + \lambda \|w\|^2)$

- $\|w\|^2$ measures the "complexity" of the hypothesis defined by $w$
- $\lambda$: controls the amount of regularization
  - It controls the trade-off between empirical error and complexity
  - Low empirical error but risk of overfitting or higher empirical error but lower complexity
Ridge Regression:

Linear Regression with squared loss + Tikhonov regularization

Linear Regression with squared loss: find $\mathbf{w}$ that minimizes the squared loss

$$
\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2
$$

Ridge Regression: find $\mathbf{w}$ that minimizes

$$
\mathbf{w} = \arg\min_{\mathbf{w}} \left( \lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 \right)
$$

$\lambda$ balances between the 2 targets

Balancing should not depend on the size of training set
Closed Form Solution

- Find optimal $w$: minimize loss \( \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (\langle w, x_i \rangle - y_i)^2 \)

- Compute gradient w.r.t. $w$ and set to 0

\[
\frac{\partial L}{\partial w} = 2\lambda w + \frac{1}{m} \sum_{i=1}^{m} (\langle w, x_i \rangle - y_i) x_i = 0 \rightarrow 2\lambda m w + \sum_{i=1}^{m} \langle w, x_i \rangle x_i = \sum_{i=1}^{m} y_i x_i
\]

- Set (as for standard least squares)

\[
A = \left( \sum_{i=1}^{m} x_i x_i^T \right) = \begin{bmatrix} \vdots & \vdots & \vdots \\ \vdots & x_1 & \vdots \\ \vdots & x_m & \vdots \end{bmatrix} \begin{bmatrix} \vdots \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \vdots \end{bmatrix} \\
\begin{bmatrix} \vdots \\ \vdots \end{bmatrix} \begin{bmatrix} \vdots \\ \vdots \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}
\]

- The solution can be rewritten as*:

\[
2\lambda m I w + A w = b \rightarrow w = (2\lambda m I + A)^{-1} b
\]

*differently from standard least square in this case the matrix is always invertible*
Tikhonov regularization makes the learner stable w.r.t. small perturbations of the training set. This in turn leads to small bounds on generalization error.

Informally: an algorithm A is stable if a small change of the training data $S$ (i.e., its input) will lead to a small change of its output hypothesis.

- What is a “small change of the training data”? 
- What is a “small change of its output hypothesis”? 


"small change of the training data" = replace one sample!
- Given $S = (z_1, ..., z_m)$ and an additional example $z'$ (i.e., pair instance label/target) let $S^{(i)} = (z_1, ..., z_{i-1}, z', z_{i+1}, ..., z_m)$

“small change of its output hypothesis” = small change in the loss
- On-Average-Replace-One-Stable (OAROS) algorithms

**Definition:**
Let be $\epsilon : \mathbb{N} \rightarrow \mathbb{R}$ a monotonically decreasing function. We say that a learning algorithm $A$ is on-average-replace-one-stable (OAROS) with rate $\epsilon (m)$ if for every distribution $D$:

$$\mathbb{E}_{(S,z') \sim D^{m+1}, i \sim U(m)}[l(A(S^{(i)}), z_i) - l(A(S), z_i)] \leq \epsilon(m)$$

- Draw IID from D
- Select at random which to replace
- With $z'$ in place of $z_i$
- Depends on training set size
Stable Rules do not Overfit

Theorem:
If algorithm $A$ is OAROS with rate $\epsilon(m)$ then:
$\mathbb{E}_{S \sim D} [L_D(A(S)) - L_S(A(S))] \leq \epsilon(m)$

Demonstration
1. True error: expected loss on one IID sample (from $D$):
   $\forall i: \mathbb{E}_S[L_D(A(S))] = \mathbb{E}_{S,z'}[l(A(S), z')] = \mathbb{E}_{S,z'}[l(A(S^{(i)}), z_i)]$
2. Training error: average error on one sample in training set:
   $\mathbb{E}_S[L_S(A(S))] = \mathbb{E}_{S,i}[l(A(S), z_i)]$
3. Combine (1)+(2) and exploit linearity of expectation and OAROS def.
   $\mathbb{E}_S[L_D(A(S)) - L_S(A(S))] = \mathbb{E}_{S,z',i}[l(A(S^{(i)}), z_i) - l(A(S), z_i)] \leq \epsilon(m)$
Definition (Lipschitzness):

Let $C \subseteq \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is $\rho$-Lipschitz over $C$ if

$$\forall \mathbf{w}_1, \mathbf{w}_2 \in C, \text{ we have that } \|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|$$

- Intuitively: the function cannot change too fast
- For derivable functions corresponds to bound on derivative:
  - If derivative bounded by $\rho$ at any point $\Rightarrow$ function is $\rho$-Lipschitz
Tikhonov Regularization is a Stabilizer

Theorem:
Assume the loss function is convex and $\rho$-Lipschitz continuous.
Then, the RLM rule with regularizer $\lambda \|w\|^2$ is OAROS with rate $\frac{2\rho^2}{\lambda m}$.
It follows that for the RLM rule:

$$\mathbb{E}_{S \sim D^m} [L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda m}$$

- Tikhonov Regularization is a Stabilizer
- Larger $\lambda$ leads a more stable solution (→ less overfitting)
- Larger training set also leads to more stable solution
- First step: demonstration not part of the course
- Second step: consequence of previous theorem
Fitting-Stability Trade-off (1)

\[ E_S[L_D(A(S))] = E_S[L_S(A(S))] + E_S[L_D(A(S)) - L_S(A(S))] \]

- \( E_S[L_S(A(S))] \): how well A fits the training set S
- \( E_S[L_D(A(S)) - L_S(A(S))] \): measures overfitting, bounded by stability of A

In Tikhonov regularization, \( \lambda \) controls tradeoff between the 2 terms

- how do \( L_S(A(S)) \) and \( \|w\|^2 \) vary as a function of \( \lambda \)?
  - Larger \( \lambda \) leads to higher empirical risk \( L_S(A(S)) \)
- how may \( E_S[L_D(A(S)) - L_S(A(S))] \) change as a function of \( \lambda \)?
  - On the other side increasing \( \lambda \) the stability term \( E_S[L_D(A(S)) - L_S(A(S))] \) decreases
- How to set \( \lambda \)?
  - Theoretical bound in the book
Fitting-Stability Trade-off (2)

\[ E_S[L_D(A(S))] = E_S[L_S(A(S))] + E_S[L_D(A(S)) - L_S(A(S))] \]

- \( E_S[L_S(A(S))] \): how well A fits the training set S
- \( E_S[L_D(A(S)) - L_S(A(S))] \): measures overfitting, bounded by stability of A

Small \( \lambda \): focus on training error
- Training error \( L_S \): small
- Difference \( L_D - L_S \): large
- Overfitting the training data

Large \( \lambda \): focus on regularization
- Training error \( L_S \): large
- Difference \( L_D - L_S \): small
- Underfitting the training data