We will show that RLM is *stable.*

**Theorem 1.** *Let $\mathcal{D}$ be a distribution and $S = (z_1, \dots, z_m)$ an iid sample from $\mathcal{D}$. Further, let $z'$ be another iid sample from $\mathcal{D}$ and $U[m]$ denote the uniform distribution over $\{1, \dots, m\}$. Then, for any learning algorithm $A$,*

$$\mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))] = \mathbb{E}_{S,z'\sim\mathcal{D}^{m+1}, i\sim U[m]}[l(A(S^i), z_i) - l(A(S), z_i)] \tag{1}$$

*where $S^i = (z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_m)$.*

Based on the right-hand side of Eq. (1), we can define our notion of stability as follows.

**Definition 1** (On-Average-Replace-One-Stable)**.** *Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. A learning algorithm $A$ is on-average-replace-one stable with rate $\epsilon(m)$ if, for every distribution $\mathcal{D}$,*

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

Now, let $A$ be the RLM rule, i.e.,

$$A(S) = \arg\min_w (L_S(w) + \lambda\|w\|^2)$$

If we define $f_S(w) = L_S(w) + \lambda\|w\|^2$, we know that $f_S$ is $2\lambda$ strongly convex and we also know that

$$\forall v : f_S(v) - f_S(A(S)) \ge \lambda\|v - A(S)\|^2 \tag{2}$$

as $A(S)$ is a minimizer of $f_S$. If we consider arbitrary $v, u$, we can write

$$\begin{aligned}
L_S(v) - L_S(u) &= L_S(v) + \lambda\|v\|^2 - \left(L_S(u) + \lambda\|u\|^2\right) \\
&= \frac{1}{m}\sum_{x\in S} l(v, z_i) + \lambda\|v\|^2 - \left(\frac{1}{m}\sum_{x\in S} l(u, z_i) + \lambda\|u\|^2\right) \\
&= \frac{1}{m}\sum_{x\in S^i} l(v, x) + \lambda\|v\|^2 - \frac{l(v, z')}{m} + \frac{l(v, z_i)}{m} - \\
&\quad \left(\frac{1}{m}\sum_{x\in S^i} l(u, x) + \lambda\|u\|^2 - \frac{l(u, z')}{m} + \frac{l(u, z_i)}{m}\right) \\
&= \frac{1}{m}\sum_{x\in S^i} l(v, x) + \lambda\|v\|^2 - \left(\frac{1}{m}\sum_{x\in S^i} l(u, x) + \lambda\|u\|^2\right) + \\
&\quad \frac{l(v, z_i) - l(u, z_i)}{m} + \frac{l(u, z') - l(v, z')}{m}
\end{aligned}$$

Now, if we set $v = A(S^i)$, the learning algorithm $A$ is run on $S^i$ and has thus seen $z'$. On the other hand, setting $u = A(S)$ means that $A$ is run on $S$ and has thus seen $z_i$, but not $z'$. Consequently, $v = A(S^i)$ minimizes the term

$$\frac{1}{m}\sum_{x\in S^i} l(v, x) + \lambda\|v\|^2 = L_{S^i}(v) + \lambda\|v\|^2 \quad .$$

As a consequence, $L_{S^i}(v) + \lambda\|v\|^2 - (L_{S^i}(u) + \lambda\|u\|^2) \leq 0$. This means that we are in the following situation:

$$\underbrace{L_S(v) - L_S(u)}_{A} = \underbrace{\frac{l(v, z_i) - l(u, z_i)}{m} + \frac{l(u, z') - l(v, z')}{m}}_{B} + \underbrace{L_{S^i}(v) + \lambda\|v\|^2 - \left(L_{S^i}(u) + \lambda\|u\|^2\right)}_{C \leq 0}$$

So, consequently $A = B + C \leq B$ and we obtain (upon setting in the expressions for $u$ and $v$)

$$f_S(A(S^i)) - f_S(A(S)) \leq \frac{l(A(S^i), z_i) - l(A(S), z_i)}{m} + \frac{l(A(S), z') - l(A(S^i), z')}{m}$$

We can also invoke Eq. (2) now to give

$$f_S(A(S^i)) - f_S(A(S)) \geq \lambda\|A(S^i) - A(S)\|^2$$

Upon combination, we get the following inequality chain:

$$\lambda\|A(S^i) - A(S)\|^2 \leq f_S(A(S^i)) - f_S(A(S)) \leq \frac{l(A(S^i), z_i) - l(A(S), z_i)}{m} + \frac{l(A(S), z') - l(A(S^i), z')}{m} \tag{3}$$

In case the loss $l(\cdot, z_i)$ is $\rho$-Lipschitz, we additionally have

$$l(A(S^i), z_i) - l(A(S), z_i) \leq \rho\|A(S^i) - A(S)\| \tag{4}$$

$$l(A(S), z') - l(A(S^i), z') \leq \rho\|A(S^i) - A(S)\| \tag{5}$$

Using these two inequalities in Eq. (3), we obtain

$$\lambda\|A(S^i) - A(S)\|^2 \leq \frac{2\rho\|A(S^i) - A(S)\|}{\lambda m}$$

$$\Leftrightarrow \|A(S^i) - A(S)\| \leq \frac{2\rho}{m}$$

Setting this back in Eq. (4) gives

$$l(A(S^i), z_i) - l(A(S), z_i) \leq \frac{2\rho^2}{\lambda m} \tag{6}$$

As this holds for any $S, z'$ and $i$, we can use this result in Def. 1. Taking the expectation over $S, z' \sim \mathcal{D}$ and $i \sim U[m]$ does not affect $2\rho^2/\lambda m$. In combination with Theorem 1, we end up with the following corollary.

**Corollary 2.** *Under a convex and $\rho$-Lipschitz loss, the RLM rule with a Tikhonov regularizer of the form $\lambda\|w\|^2$ is on-average-replace-one stable, i.e.,*

$$\mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \epsilon(m)$$

*with rate*

$$\epsilon(m) = \frac{2\rho^2}{\lambda m}$$