Regularization and Stability

- ERM algorithm

$$w = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; L_S(w)$$

- Regularized Loss Minimization (RLM)

$$w = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; L_S(w) + R(w)$$

where $R : \mathbb{R}^d \to \mathbb{R}$.

- The function R is called a <u>regularizer</u>.

- Two interpretations:

  1) Similar to SRM or MDL

  2) Stability

- Regarding 1), assume R is non-negative. We can consider

$$H_n = \{ w \in \mathbb{R}^d : R(w) \leq n \}$$

Then

$$H_1 \subseteq H_2 \subseteq H_3 \subseteq \cdots$$

$$H = \bigcup_{n=1}^{\infty} H_n$$

- By far the most common regularizer is

$$R(w) = \lambda \|w\|_2^2$$

where $\lambda > 0$ and

$$\|w\|_2^2 = \sum_{i=1}^{d} w_i^2$$ (Euclidean norm squared)

It is called L2-regularizer or Tikhonov regularizer.

Ridge regression

- Linear least squares with L2-regularizer is called

Ridge regression:

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; \lambda \|w\|_2^2 + \frac{1}{m} \sum_{i=1}^{m} \left(w^T x_i - y_i \right)^2$$

where $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$

and $y_1, y_2, \ldots, y_m \in \mathbb{R}$.

- Ridge regression has a solution

$$\hat{w} = \left(\lambda m I + A \right)^{-1} b$$

where $A = \sum_{i=1}^{m} x_i x_i^T$

$$b = \sum_{i=1}^{m} y_i x_i$$

$i=1$

- To derive the solution, let

$$f(w) = \lambda \|w\|_2^2 + \sum_{i=1}^{m} (w^T x_i - y_i)^2.$$

- The solution satisfies

$$\nabla f(w) = 0$$

$$\Updownarrow$$

$$2\lambda w + \frac{2}{m} \sum_{i=1}^{m} (w^T x_i - y_i) x_i = 0$$

$$\Updownarrow$$

$$m\lambda w + \left(\sum_{i=1}^{m} x_i x_i^T \right) w = \sum_{i=1}^{m} y_i x_i$$

$$\Updownarrow$$

$$(m\lambda I + A)\, w = b$$

- A is symmetric positive semi-definite matrix.

  Its eigen values are non-negative:

$$A = V^T D V$$

$$V^T = V^{-1}, \quad D \text{ is non-negative diagonal matrix}$$

$$D = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ 0 & \cdots & \lambda_d \end{pmatrix}$$

$$m\lambda I + A = m\lambda V^T V + V^T D V$$

$$= V^T (m\lambda I + D) V$$

$$= V^T D' V$$

$$D' = \begin{pmatrix} \lambda_1 + \lambda m & & 0 \\ & \ddots & \\ 0 & & \lambda_d + \lambda m \end{pmatrix}$$ ← strictly positive diagonal entries

So eigenvalues of $(m\lambda I + A)$ are positive and thus $m\lambda I + A$ is invertible:

$$(m\lambda I + A)^{-1} = (V^T D' V)^{-1}$$

$$= V^{-1} (D')^{-1} (V^T)^{-1}$$

$$= V^T (D')^{-1} V$$

$= V'(D')\ V$

↗

This exists.

- L2 regularization is used also for logistic regression and also in neural networks.

Stability

- Let $Z = X \times Y$ be the space of labeled samples. (This simplifies the notation)

- Let D be a distribution over Z.

- Let $S = (z_1, z_2, \ldots, z_m)$ be an i.i.d. sample from D.
- Let H be a space of predictors.
- Let $\ell : H \times Z \to \mathbb{R}$ be a loss function
- Let $A : Z^* \to H$ be a learning algorithm.
- Overfitting happens when

$$L_D(A(S)) - L_S(A(S))$$

is large positive number.

- More specifically we look at

$$\mathbb{E}\left[L_D(A(S)) - L_S(A(S)) \right]$$

(The expectation is w.r.t. random choice of $S \sim D^m$)

- This quantity can be expressed in alternative way, using stability.

- Suppose $z' \sim D$ is a single example independent of S

- For any $i = 1, 2, \ldots, m$ we form "alternative" sample

$$S_i = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$$

z_i is replaced by z'

- Note: S_i is an i.i.d. sample from D

Definition (On-Average-Replace-One-Stability)

Let $\varepsilon : \mathbb{N} \to \mathbb{R}$ be a non-increasing function.

Let $A : Z^* \to H$ be an algorithm.

Algorithm A is called <u>on-average-replace-one-stable</u> if for any distribution D and any $m \geq 1$,

$$\mathbb{E}\left[\ell\left(A(S_I), z_I\right) - \ell\left(A(S), z_I\right)\right] \leq \varepsilon(m)$$

where $S \sim D^m$

and $I \sim \text{Uniform}\left(\{1, 2, \ldots, m\}\right)$.

The function $\varepsilon(m)$ is called a rate.

<u>Theorem</u>:

With notation from above definition

with ...

$$\mathbb{E}\left[\ell\left(A(S_I), z_I\right) - \ell\left(A(S), z_I\right)\right]$$

$$= \mathbb{E}\left[L_D(A_S) - L_S(A(S))\right]$$

Proof:

- $\mathbb{E}\left[L_D(A(S))\right] = \mathbb{E}\left[L_D(A(S_I))\right]$

$$= \mathbb{E}\left[\ell\left(A(S_I), z_I\right)\right]$$

These are independent

- $\mathbb{E}\left[L_D(A(S))\right] - \mathbb{E}\left[\frac{1}{m}\sum \ell(A(S), \ldots)\right]$

$$+\left[L_S(A(S))\right] - \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \ell(A(S), z_i)\right]$$

$$= \mathbb{E}\left[\ell(A(S), z_I)\right]$$

$I \sim \text{Uniform}(\{1, 2, \ldots, m\})$

Strongly convex functions

Definition:

Let $C \subseteq \mathbb{R}^d$ be a convex set.

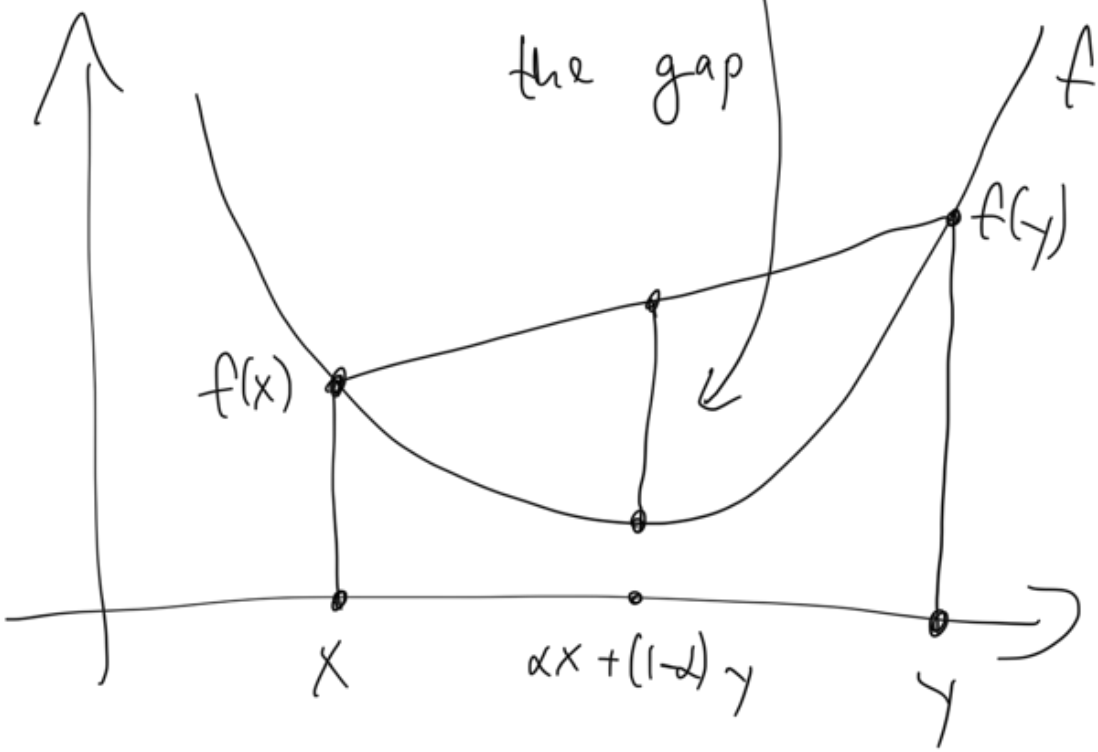
Let $\lambda \geq 0$ be a real number.

A function $f: C \to \mathbb{R}$ is called

λ-strongly convex if

for every $\alpha \in [0,1]$ and every $x, y \in C$

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\lambda}{2} \alpha(1-\alpha) \|x-y\|_2^2$$



Note: Any convex function is

0-strongly convex

Lemma:

1) The function $f(x) = \lambda \|x\|_2^2$ is (2λ)-strongly convex $(\lambda \geq 0)$.

2) If f is λ-strongly convex and g is convex, then $f+g$ is λ-strongly convex

3) If f is λ-strongly convex and x^* is a mimizer of f then for any x

$$f(x) - f(x^*) \geq \frac{\lambda}{2} \|x - x^*\|_2^2$$

□

Proof:

1) $f(x) = \|x\|_2^2$ satisfies (2λ)-strong convexity condition with equality:

$$\lambda\|\alpha x + (1-\alpha)y\|_2^2 = \alpha\lambda\|x\|_2^2 + (1-\alpha)\lambda\|y\|_2^2 - \lambda\alpha(1-\alpha)\|x-y\|_2^2$$

2) A more general fact is true: If f is λ_1-strongly convex and g is λ_2-strongly convex, then $f+g$ is $(\lambda_1+\lambda_2)$-strongly convex.

3) By definition of λ-strong convexity, for $\alpha \in (0,1)$

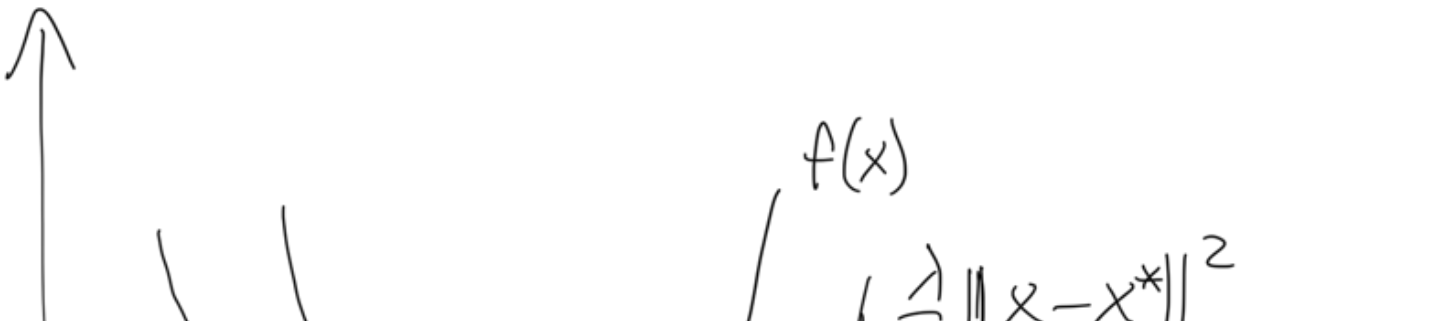
$$\underbrace{\frac{f(x^* - \alpha(x-x^*)) - f(x^*)}{\alpha}}_{\downarrow\!\downarrow\ 0} \leq f(x) - f(x^*) - \frac{\lambda}{2}(1-\alpha)\|x-x^*\|_2$$

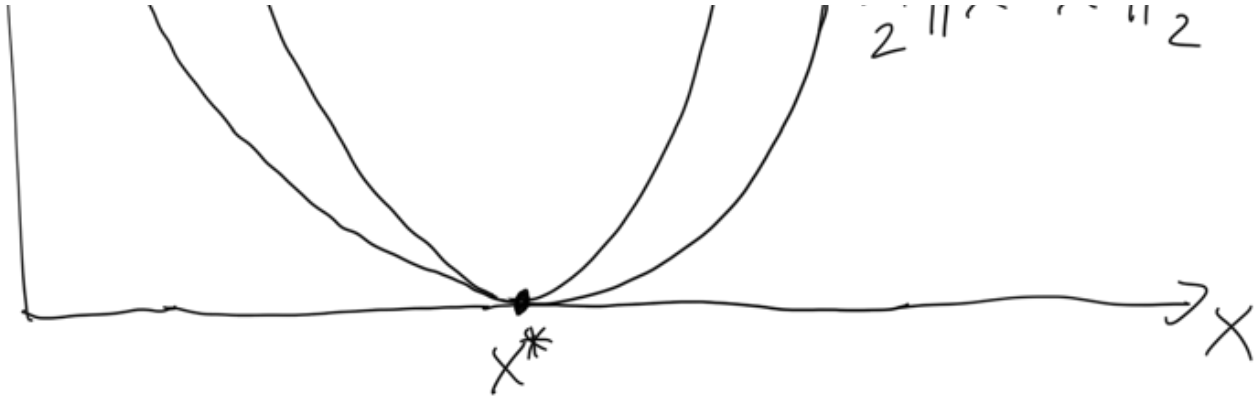
- Thus, for any $\alpha \in (0,1)$

$$0 \leq f(x) - f(x^*) - \frac{\lambda}{2}(1-\alpha)\|x-x^*\|_2^2$$

- Thus

$$0 \leq f(x) - f(x^*) - \frac{\lambda}{2}\|x-x^*\|_2^2$$





Stability of RLM

- RLM with L2-regularizer

$$A(S) = \underset{w \in C}{\operatorname{argmin}} \; L_S(w) + \lambda \|w\|_2^2$$

- $C \subseteq \mathbb{R}^d$ is a convex set (Subset of domain of L_S)

Lemma: Let A be RLM,

$$A(S) = \operatorname{argmin} \; L_S(w) + \lambda \|w\|^2 .$$

$A(S) = \underset{w \in C}{\text{argmin}} \; L_S(w)$

If A is on-average-replace-one-stable with rate $\varepsilon(m)$ then for all $w \in C$,

$$\mathbb{E}\left[L_D(A(S))\right] \leq L_D(w) + \lambda \|w\|_2^2 + \varepsilon(m)$$

where $S \sim D^m$.

Proof:

- $\mathbb{E}\left[L_D(A(S))\right] \leq \mathbb{E}\left[L_S(A(S))\right] + \varepsilon(m)$
- $L_S(A(S)) \leq L_S(A(S)) + \lambda \|A(S)\|_2^2$

$$\leq L_S(w) + \lambda \|w\|_2^2$$

since $A(S)$ minimizes

$$L_S(w) + \lambda \|w\|_2^2$$

- Note: The result is similar to non-uniform PAC learning: Instead of a "high-probability result", lemma proves an upper bound "in-expectation."

- If C is bounded, then $\lambda\|w\|^2$ is bounded and we get a "uniform" PAC-like result

  (This is NOT uniform convergence, though.)

- It remains to show that $R(\Pi$ is on-average-replace-one-stable

We will need convex analysis for that.

- Let $f_S : C \to \mathbb{R}$,

$$f_S(w) = L_S(w) + \lambda \|w\|_2^2.$$

- Note: If $L_S : C \to \mathbb{R}$ is convex then f_S is (2λ)-strongly convex.

Lemma:

Assuming $L_S : C \to \mathbb{R}$ is convex then

$$\lambda \|A(S^i) - A(S)\|_2^2 \leq \frac{\ell(A(S^i), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^i), z')}{m}$$

Proof:

- Let $v = A(S_i)$ and $u = A(S)$
- We lower and upper bound

$$f_S(v) - f_S(u) .$$

- Since f_S is (2λ)-strongly convex and u is the minimizer of f_S, for any $w \in C$:

$$f_S(v) - f_S(u) \geq \lambda \|u - v\|^2$$

- $f_S(v) - f_S(u) =$

$$= L_S(v) + \lambda \|v\|_2^2 - \left(L_S(u) + \lambda \|u\|^2 \right)$$

$$= L_{S_i}(v) + \lambda \|v\|_2^2 - \left(L_{S_i}(u) + \lambda \|u\|^2 \right)$$

$$+ \frac{\ell(v, z_i) - \ell(u, z_i)}{m} + \frac{\ell(u, z') - \ell(v, z')}{m}$$

$$\leq \frac{\ell(v, z_i) - \ell(u, z_i)}{m} + \frac{\ell(u, z') - \ell(v, z')}{m}$$

Since $v = \operatorname{argmin}_{w \in C} L_{S_i}(w) + \lambda \|w\|_2^2$

- Combine upper and lower bounds

∎

Lipschitz losses

Theorem: (First coordinate)

Suppose $\ell : H \times Z \to \mathbb{R}$ is ρ-Lischitz in the first coordinate.

Let A be RLM algorithm

$$A(S) = \operatorname*{argmin}_{w \in C} L_S(w) + \lambda \|w\|^2.$$

Then A is on-average-replace-one-stable with rate $\varepsilon(m) = \frac{2\rho^2}{\lambda m}$

Proof:

- Let $v = A(S_i)$ and $u = A(S)$

- Since ℓ is ρ-Lipschitz,

$$\ell(v, z_i) - \ell(u, z_i) \leq \rho \|u - v\|$$

and

$$\ell(u, z') - \ell(v, z') \leq \rho \|u - v\|$$

- By previous lemma,

$$\lambda \|u - v\|_2^2 \leq \frac{2\rho \|u - v\|_2}{m}$$

- Therefore

$$\|u - v\|_2 \leq \frac{2\rho}{\lambda m}$$

- Thus

$$\ell\left(A(S_i), z_i\right) - \ell\left(A(S), z_i\right)$$

$$\leq \ell\left(v, z_i\right) - \ell\left(u, z_i\right)$$

$$\leq \rho \|u - v\|_2$$

$$\leq \frac{2\rho^2}{\lambda m} .$$

- Take expectation over S and I:

$$\mathbb{E}\left[\ell\left(A(S_I), z_I\right) - \ell\left(A(S), z_I\right)\right] \leq \frac{2\rho^2}{\lambda m} .$$

Corollary: (First coordinate)

Let $l : C \times Z \to \mathbb{R}$ be convex and ρ-Lipchitz in the first coordinate.

Let $A : Z^* \to \mathbb{R}$ be the RLM algorithm,

$$A(S) = \underset{w \in C}{\operatorname{argmin}} \; L_S(w) + \lambda \|w\|_2^2.$$

Then, for any $w \in C$

$$E\left[L_D(A(S))\right] \leq L_D(w) + \lambda \|w\|_2^2 + \frac{2\rho^2}{\lambda m}$$

Corollary:

Let $C \subseteq \mathbb{R}^d$ be a convex set such that for every $w \in C$,

$$\|w\|_2 \leq B.$$

Let $\ell : C \times Z \to \mathbb{R}$ be convex and ρ-Lipschitz in the first coordinate.

Let $A : Z^* \to \mathbb{R}$ be the RLM algorithm

$$A(S) = \operatorname*{argmin}_{w \in C} L_S(w) + \lambda \|w\|_2^2$$

where $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ and $m = |S|$.

Then,

$$\mathbb{E}\left[L_D(A(S))\right] \leq \inf_{w \in C} L_D(w) + \rho B \sqrt{\frac{8}{m}}.$$

Note: The $\rho B \sqrt{\frac{8}{m}}$ decreases

Note: the term $\sqrt{\frac{\ln(|\mathcal{H}|/\delta)}{m}}$ decreases with sample size.

So if $m \geq \frac{8\rho^2 B^2}{\varepsilon^2}$ then

$$\mathbb{E}\left[L_D(A(S))\right] \leq \inf_{w \in C} L_D(w) + \varepsilon.$$

Smooth Non-negative Loss

Lemma:
Let $f: \mathbb{R}^d \to \mathbb{R}$ be β-smooth.
Suppose f is also non-negative.
Then, for all $x \in \mathbb{R}^d$

$$\|\nabla f(x)\|^2 \leq 2\beta f(x)$$

$\|\nabla f(x)\|^2 \leq 2\beta f(x)$

Proof:

- Since f is β-smooth, for any $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|x - y\|^2$$

- Substitute

$$y = x - \frac{1}{\beta} \nabla f(x)$$

and use

$$f(y) \geq 0$$

- We get

$$0 \le f(x) + \nabla f(x)^T\left(-\frac{1}{\beta}\nabla f(x)\right) + \frac{\beta}{2}\left\|\frac{1}{\beta}\nabla f(x)\right\|^2$$

$$\Updownarrow$$

$$0 \le f(x) - \frac{1}{\beta}\|\nabla f(x)\|^2 + \frac{1}{2\beta}\|\nabla f(x)\|^2$$

$$\Updownarrow$$

$$\|\nabla f(x)\|^2 \le f(x)$$