Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

Convex learning: Surrogate loss and Tikhonov reularisation

Context

- Preparation for analysing SVM

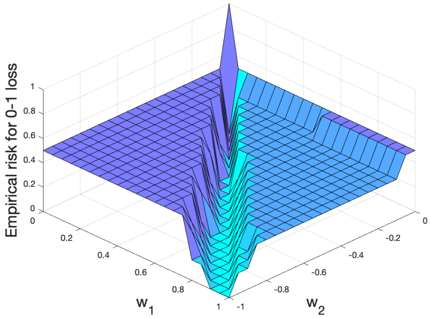
- Let each predictor parametrised by $w \in \mathbb{R}^p$

$$\mathcal{H} = \left\{ \mathrm{sign}(\langle w, x \rangle) \ : \ w \in \mathbb{R}^p \right\}.$$

- ERM with 0-1 loss: Non-convex optimisation

$$\min_{w \in \mathbb{R}^p} \frac{1}{m} \sum_{i=1}^{m} \mathbf{1} \left\{ \mathrm{sign}(\langle w, x_i \rangle) \neq y_i \right\}$$

  - Non-convex optimisation difficult to analyse
  - We may not reach global optimum



Typical landscape for 0-1 loss

Here $\langle w, x \rangle = w^\top x$
For general linear classifiers $\mathrm{sign}(w^\top x + b)$, define $w' = (w, b)$ and $x' = (x, 1)$, and write it as $\mathrm{sign}(\langle w', x' \rangle)$

Outline

- Convex losses
- Lipschitz losses (smooth)
- Surrogate loss minimisation (ERM with losses that are upper bound for 0-1 loss)
- Tikhonov regularisation (Regularisation with a convex function)

Convex set

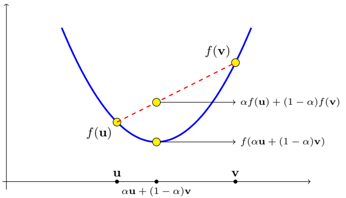
- $C =$ subset of some vector space
- C is a convex set if:
  - for every $u, v \in C$, the line segment joining u, v lies in C,
  - equivalently, for every $\alpha \in [0, 1]$, we have

$$\alpha u + (1 - \alpha) v \in C$$

Convex and strongly convex functions

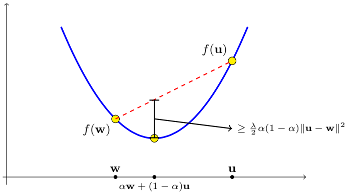
- Given C = convex set
- Function $f : C \to \mathbb{R}$ is a convex function if:
  - for every $u, v \in C$ and $\alpha \in [0,1]$

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v)$$



- Function $f : C \to \mathbb{R}$ is λ-strongly convex if:
  - for all $u, v \in C$ and $\alpha \in (0,1)$,

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v) - \frac{\lambda}{2}\alpha(1-\alpha)\|u-v\|^2$$



Some properties of convex functions

- Assume $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable function

$$f \text{ is convex} \qquad \Leftrightarrow \qquad f"(x) \geq 0 \text{ for all } x$$

- Jensen's inequality
  - $f : C \to \mathbb{R}$ is convex
  - $u_1, \ldots, u_n \in C$ and $\alpha_1, \ldots, \alpha_n \in [0,1]$ with $\sum_{i=1}^{n} \alpha_i = 1$

$$f\left(\sum_{i=1}^{n} \alpha_i u_i\right) \leq \sum_{i=1}^{n} \alpha_i f(u_i)$$

Local and global minimum

- $u \in C$ is called a global minimum for $f : C \to \mathbb{R}$ if

$$f(u) \le f(v) \qquad \text{for all } v \in C$$

- $u \in C$ is called a local minimum for f if for some $\epsilon > 0$,

$$f(u) \le f(v) \qquad \text{for all } v \text{ such that } \|v - u\| < \epsilon$$

Minimum for convex function

- f is convex:

  every local minimum of f is also a global minimum

- f is λ-strongly convex and u is a minimum:

$$f(v) \geq f(u) + \frac{\lambda}{2}\|v-u\|^2 \qquad \text{for every } v \in C$$

- Try to prove them (proof in lecture notes)

Revisiting loss functions

- Let $\mathcal{X} \subset \mathbb{R}$ and $\mathcal{Y} = \{\pm 1\}$
- Let $\mathcal{H} = \{h_w(x) = \text{sign}(wx) \ : \ w \in \mathbb{R}\}$
- We ignore $\text{sign}(\cdot)$ and view loss as function of w
    - Given (x, y) : $\ell(w)$ computed from wx and y, or sometimes $y \cdot wx$
- Examples:
    - 0-1 loss: $\ell(w) = \mathbf{1}\{y \neq \text{sign}(wx)\} = \mathbf{1}\{y \cdot wx \leq 0\}$
    - squared loss: $\ell(w) = (y - wx)^2 = (1 - y \cdot wx)^2$ assuming $y \in \{\pm 1\}$

Convexity of losses when $w \in \mathbb{R}^p$

- Linear classifier in $\mathbb{R}^p$: $\quad \mathrm{sign}(\langle w, x\rangle)$

- Let $g : \mathbb{R} \to \mathbb{R}$ is convex. For $x \in \mathbb{R}^p$, define

$$f : \mathbb{R}^p \to \mathbb{R} \qquad f(w) = g\big(\langle w, x\rangle\big)$$

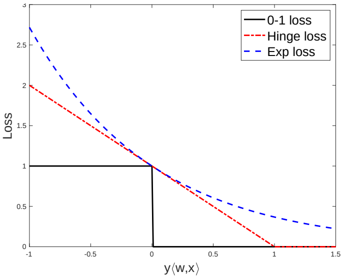
  - f is convex function with respect to w

- Proof: For $\alpha \in (0,1)$ and $w_1, w_2 \in \mathbb{R}^p$,

$$\begin{aligned} f(\alpha w_1 + (1-\alpha)w_2) &= g\big(\langle \alpha w_1 + (1-\alpha)w_2, x\rangle\big) \\ &= g\big(\alpha\langle w_1, x\rangle + (1-\alpha)\langle w_2, x\rangle\big) \\ &\leq \alpha \underbrace{g(\langle w_1, x\rangle)}_{=f(w_1)} + (1-\alpha) \underbrace{g(\langle w_2, x\rangle)}_{=f(w_2)} \end{aligned}$$

Convex and non-convex losses

- Fix $x \in \mathbb{R}^p$, $y \in \mathbb{R}$. Verify following losses are convex:
  - squared loss: $\ell(w) = (y - \langle w, x \rangle)^2$
  - hinge loss: $\ell(w) = \max\{0, 1 - y\langle w, x \rangle\}$
  - exponential loss: $\ell(w) = \exp\big(-y\langle w, x \rangle\big)$
- Verify following losses:
  - 0-1 loss: $\ell(w) = \mathbf{1}\{y\langle w, x \rangle \leq 0\}$
  - ramp loss: $\ell(w) = 1 - y\langle w, x \rangle$ for $0 \leq y\langle w, x \rangle \leq 1$, else clipped to 0 or 1



Convex learning problem

- A learning problem, characterised by $\mathcal{H}$ and loss ℓ, is **convex** if
  - $\mathcal{H}$ is a convex set (we view $\mathcal{H}$ as set of parameters w)
  - for every (x, y), the loss $\ell(h_w(x), y)$ is convex with respect to w
- ERM of convex learning is a convex optimisation problem

$$\underset{w \in \mathcal{H}}{\text{minimise}} \ \frac{1}{m} \sum_{i=1}^{m} \ell_{x_i, y_i}(w)$$

  - $\ell_{x,y}(w) =$ loss function for w computed using labelled example (x, y)
  - Objective is convex since it is sum of convex functions

Lipschitz functions

- Function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be ρ-Lipschitz if for every $u, v \in \mathbb{R}^p$,

$$|f(u) - f(v)| \leq \rho \|u - v\|$$

  where $\|\cdot\|$ is Euclidean norm

- Hinge loss $\ell(w) = \max\{0, 1 - y\langle w, x\rangle\}$ is $(|y| \cdot \|x\|)$-Lipschitz

- Proof: Consider $w_1, w_2 \in \mathbb{R}^p$
  - $y\langle w_1, x\rangle < 1$, $y\langle w_2, x\rangle < 1$: $\quad |f(w_1) - f(w_2)| = |y\langle w_2 - w_1, x\rangle| \quad \leq |y| \cdot \|x\| \cdot \|w_1 - w_2\|$
  - $y\langle w_1, x\rangle \geq 1 > y\langle w_2, x\rangle$ (same for the opposite):

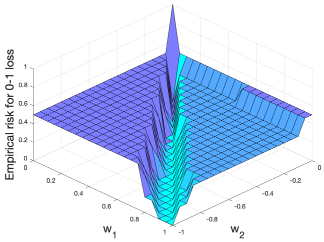
$$|f(w_1) - f(w_2)| \;=\; 1 - y\langle w_2, x\rangle \;<\; y\langle w_1 - w_2, x\rangle \quad \leq |y| \cdot \|x\| \cdot \|w_1 - w_2\|$$

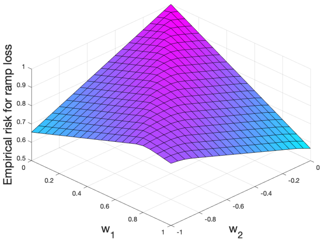
  - $y\langle w_1, x\rangle \geq 1$, $y\langle w_2, x\rangle \geq 1$: $\quad f(w_1) - f(w_2) = 0$

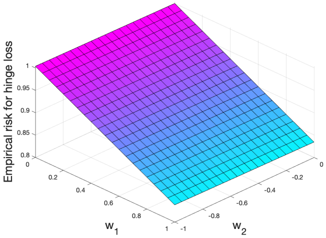
Convex Lipschitz bounded learning

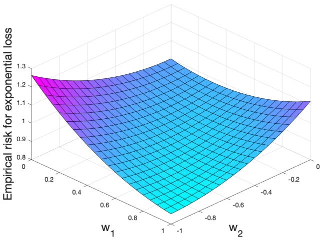
- Learning problem, characterised by $\mathcal{H}$ and loss ℓ, is convex Lipschitz bounded with parameters ρ, B if
  - $\mathcal{H}$ is a convex set
  - every $w \in \mathcal{H}$ satisfies $\|w\| \leq B$.
  - loss $\ell(w)$ is convex and ρ-Lipschitz with respect to w for any x, y
- Example: $\mathcal{X} = \{x \ : \ \|x\| \leq \rho\}$; $\quad \mathcal{H} = \{h_w(x) = \langle w, x \rangle \ : \ \|w\| \leq B\}$; $\quad$ loss = hinge

Landscape for various losses for $w \in \mathbb{R}^2$









Generalisation error w.r.t. 0-1 loss

- Convex Lipschitz losses useful for solving ERM
  - Smooth / convex cost function
  - Can useful methods from convex optimisation
- But, we are interested is expected test error (0-1 loss)
- Need losses with additional properties
  - Upper bound for 0-1 loss

Surrogate loss and convex surrogate loss

- Let ℓ, ℓ' be two loss functions notation: $\ell_{x,y}(w) = \ell(h_w(x), y)$
- ℓ' is a surrogate to ℓ if:
  - $\ell_{x,y}(w) = \leq \ell'_{x,y}(w)$ for every w, x, y
- ℓ' is a convex surrogate to ℓ if:
  - ℓ' is surrogate to ℓ
  - $\ell'_{x,y}(\cdot)$ is convex function for every x, y

Examples

Verify:

- hinge loss is convex surrogate to 0-1 loss
- exponential loss is convex surrogate to 0-1 loss
- ramp loss is surrogate to 0-1 loss, but not convex surrogate
- exponential loss is convex surrogate to ramp and hinge losses ...

Usefulness of surrogate loss minimisation

- Consider ERM w.r.t. hinge loss

- Assume we have generalisation error bound with respect to hinge loss

$$L_{\mathcal{D}}^{hinge}(\mathcal{A}_{ERM-hinge}) \leq L_{\mathcal{D}}^{hinge}(\mathcal{H}) + \epsilon$$

- Since hinge is surrogate to 0-1 loss

$$L_{\mathcal{D}}^{0-1}(\mathcal{A}_{ERM-hinge}) \leq L_{\mathcal{D}}^{hinge}(\mathcal{A}_{ERM-hinge}) \leq L_{\mathcal{D}}^{hinge}(\mathcal{H}) + \epsilon$$

- We will use this approach to derive generalisation error bound for soft SVM

Regularised loss minimisation (RLM) and Tikhonov regularisation

- View $\mathcal{H}$ as set of parameters w

- Regularised loss minimisation

$$\mathcal{A}_S = \arg\min_{w\in\mathcal{H}} \; L_S(w) + \text{penalty}(w)$$

- Tikhonov regularisation ... $\lambda > 0$

$$\mathcal{A}_S = \arg\min_{w\in\mathcal{H}} \; L_S(w) + \lambda\|w\|^2$$

  - $L_S =$ empirical risk w.r.t. some loss
  - $g(w) = \lambda\|w\|^2$ is 2λ-strongly convex
  - If loss is convex, the regularised loss is also 2λ-strongly convex

Tikhonov RLM for convex loss

- $g(w) = \lambda\|w\|^2$ is 2λ-strongly convex: Verify

$$\alpha g(w_1) + (1-\alpha)g(w_2) - \frac{2\lambda}{2}\alpha(1-\alpha)\|w_1 - w_2\|^2 = \lambda\|\alpha w_1 + (1-\alpha)w_2\|^2$$

- Recall: $\ell(w)$ is convex w.r.t $w \implies L_S(w)$ is convex

- L_S convex, g 2λ-strongly convex $\implies L_S + g$ is 2λ-strongly convex

$$(L_S + g)(\alpha w_1 + (1-\alpha)w_2) = \underbrace{L_S(\alpha w_1 + (1-\alpha)w_2)}_{\text{use convexity}} + \underbrace{g(\alpha w_1 + (1-\alpha)w_2)}_{\text{use strong convexity}}$$

$$\leq \alpha(L_S + g)(w_1) + (1-\alpha)(L_S + g)(w_2) - \frac{2\lambda}{2}\|w_1 - w_2\|^2$$

Stability of Tikhonov regularisation

Theorem Conv.1 (Tikhonov RLM is a stable learner)

- $\ell =$ *convex*, ρ*-Lipschitz loss with respect to* w
- *Tikhonov RLM* $\mathcal{A}_S$ *based on loss* ℓ *is on-average-replace-one stable with rate* $\dfrac{2\rho^2}{\lambda m}$
- *Expected generalisation error of* $\mathcal{A}_S$ *satisfies*

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

Proof

- $S \sim \mathcal{D}^m$ and $(x', y') \sim \mathcal{D}$
- $S^i =$ set where $(x_i, y_i) \in S$ is replaced by (x', y') ... used for replace one stability
- $f(w) = L_S(w) + \lambda\|w\|^2$ is 2λ-strongly convex.
- $\mathcal{A}_S =$ minimiser for $f(w)$ ... note $\mathcal{A}_S$ is optimal parameter
- Due to 2λ-strong convexity

$$f(w) - f(\mathcal{A}_S) \geq \lambda\|w - \mathcal{A}_S\|^2 \qquad \text{for all } w \in \mathcal{H}$$

Proof

We write $f(w) - f(v)$ in terms of L_{S^i}

$$
\begin{aligned}
f(w) - f(v) &= L_S(w) + \lambda\|w\|^2 - L_S(v) - \lambda\|v\|^2 \\
&= L_{S^i}(w) + \frac{\ell_{x_i,y_i}(w) - \ell_{x',y'}(w)}{m} + \lambda\|w\|^2 - L_{S^i}(v) - \lambda\|v\|^2 + \frac{\ell_{x',y'}(v) - \ell_{x_i,y_i}(v)}{m} \\
&= L_{S^i}(w) + \lambda\|w\|^2 - L_{S^i}(v) - \lambda\|v\|^2 + \frac{\ell_{x_i,y_i}(w) - \ell_{x_i,y_i}(v)}{m} + \frac{\ell_{x',y'}(v) - \ell_{x',y'}(w)}{m} \\
&\leq L_{S^i}(w) + \lambda\|w\|^2 - L_{S^i}(v) - \lambda\|v\|^2 + \frac{2\rho\|w - v\|}{m}
\end{aligned}
$$

Above use ρ-Lipschitz property of ℓ: $\quad |\ell_{x,y}(v) - \ell_{x,y}(w)| \leq \rho\|w - v\|$

Proof

- Set $w = \mathcal{A}_{S^i}$ and $v = \mathcal{A}_S$

$$L_{S^i}(w) + \lambda\|w\|^2 \leq L_{S^i}(v) + \lambda\|v\|^2 \quad \Longrightarrow \quad f(\mathcal{A}_{S^i}) - f(\mathcal{A}_S) \leq \frac{2\rho\|\mathcal{A}_{S^i} - \mathcal{A}_S\|}{m}$$

- Combining with lower bound due to strong convexity

$$\lambda\|\mathcal{A}_{S^i} - \mathcal{A}_S\|^2 \leq \frac{2\rho}{m}\|\mathcal{A}_{S^i} - \mathcal{A}_S\| \qquad \text{or} \qquad \|\mathcal{A}_{S^i} - \mathcal{A}_S\| \leq \frac{2\rho}{\lambda m}$$

- Using Lipschitz property, for every x, y

$$|\ell_{x,y}(\mathcal{A}_{S^i}) - \ell_{x,y}(\mathcal{A}_S)| \leq \rho\|\mathcal{A}_{S^i} - \mathcal{A}_S\| \leq \frac{2\rho^2}{\lambda m}$$

- Above also implies on-average-replace-one stability of learner $\mathcal{A}$ with rate $\frac{2\rho^2}{\lambda m}$
  Final statement on generalisation error discussed under generalisation from stability