# SVM
# (Chap2. Loss Functions and Their Risks)

Ingo et al. 2017

Presented by Jiin Seo

February 19, 2018

# Outline

# Outline

# 1. Loss Functions

Def 2.1 *(Loss Function)* Let $(\mathbf{X}, \mathcal{A})$ be a m'able space and $\mathbf{Y} \subset \mathbb{R}$ be a closed subset. Then a function $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is called a ***loss function*** , or simply a ***loss***, if it is m'able.

$\mathrm{L}(x, y, f(x))$ is the cost of predicting $y$ by $f(x)$ if $x$ is observed.

Our goal is to have a small average loss for future unseen obs. $(x, y)$.

# 1. Loss Functions

Def 2.2 - 2.3 *(L-Risk and Bayes risk)*
$\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ : loss ftn
$\mathbf{P}$ : p.m. on $\mathbf{X} \times \mathbf{Y}$.
Then, for a m'able ftn $f : \mathbf{X} \to \mathbb{R}$ , the ***L-Risk*** is defined by

$$\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) := \int_{\mathbf{X}\times\mathbf{Y}} \mathrm{L}(x, y, f(x))d\mathbf{P}(x, y) = \int_{\mathbf{X}} \int_{\mathbf{Y}} \mathrm{L}(x, y, f(x))d\mathbf{P}(y \mid x)d\mathbf{P}_{\mathbf{X}}(x)$$

And, the minimal L-risk

$$\mathcal{R}^*_{\mathrm{L},\mathbf{P}} := inf\{\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) \mid f : \mathbf{X} \to \mathbb{R}\ m'able\}$$

is called the ***Bayes risk*** w.r.t. $\mathbf{P}$ and $\mathrm{L}$ .
In addition, a m'able $f^*_{\mathrm{L},\mathbf{P}} : \mathbf{X} \to \mathbb{R}$ with $\mathcal{R}_{\mathrm{L},\mathbf{P}}(f^*_{\mathrm{L},\mathbf{P}}) = \mathcal{R}^*_{\mathrm{L},\mathbf{P}}$ is called a Bayes decision function.

# 1. Loss Functions

Example *(Empirical L-Risk)*
For a given sequence $\mathcal{D} := ((x_1, y_1), \cdots, (x_n, y_n)) \in (\mathbf{X} \times \mathbf{Y})_n$ , we write $\mathbf{D} := \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, y_i, f(x_i))$ . ($\mathbf{D}$ is the empirical measure).
The risk of a function $f : \mathbf{X} \to \mathbb{R}$ w.r.t this measure is called the ***empirical L-risk***

$$\mathcal{R}_{\mathrm{L},\mathbf{D}}(f) := \frac{1}{n} \sum_{i=1}^{n} \mathrm{L}(x_i, y_i, f(x_i))$$

- We assume that $\mathcal{D}$ is a seq. of i.i.d. obs. generated by $\mathbf{P}$ and $f$ satisfies $\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) < \infty$.
  By L.L.N. , we see that $\mathcal{R}_{\mathrm{L},\mathbf{D}}(f) \to \mathcal{R}_{\mathrm{L},\mathbf{P}}(f)$ with high prob.

# 1. Loss Functions

Example2.4 *(Standard binary classification)*
The goal is to predict the label $y$ by $t$ if $x$ is observed.
Let $\mathbf{Y} := \{-1, 1\}$ and $\mathbf{P}$ be an unknown distn on $\mathbf{X} \times \mathbf{Y}$.
The ***classification loss*** $\mathrm{L}_{class} : \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is defined by

$$\mathrm{L}_{class} := \mathbf{I}_{(-\infty,0]}(y \text{ sign } t), \quad y \in \mathbf{Y}, t \in \mathbb{R}.$$

$$\begin{aligned}\mathcal{R}_{\mathrm{L}_{class},\mathbf{P}}(f) &= \int_{\mathbf{X}} \{\eta(x)\mathbf{I}_{(-\infty,0)}(f(x)) + (1-\eta(x))\mathbf{I}_{[0,\infty)}(f(x))\} d\mathbf{P}_{\mathbf{X}}(x) \\ &= \mathbf{P}(\{(x,y) \in \mathbf{X} \times \mathbf{Y} : \text{sign } f(x) \neq y\}), \\ &( \ \eta(x) := \mathbf{P}(y = 1|x) \ )\end{aligned}$$

$$\mathcal{R}^*_{\mathrm{L}_{class},\mathbf{P}} = \int_{\mathbf{X}} min\{\eta, 1-\eta\} d\mathbf{P}_{\mathbf{X}}.$$

# 1. Loss Functions

Example2.5 *(Weighted binary classification)*
The goal is to predict the label $y$ by $t$ if $x$ is observed.
Let $\mathbf{Y} := \{-1, 1\}$ and $\alpha \in (0, 1)$..
The $\alpha$-weighted classification loss $\mathrm{L}_{\alpha-class} : \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is defined by

$$\mathrm{L}_{\alpha-class}(y, t) := \begin{cases} 1-\alpha & \text{if } y = 1 \text{ and } t < 0 \\ \alpha & \text{if } y = -1 \text{ and } t \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

$$\mathcal{R}_{\mathrm{L}_{\alpha-class}, \mathbf{P}}(f) = (1-\alpha)\int_{f<0} \eta \, d\mathbf{P}_{\mathbf{X}} + \alpha \int_{f \geq 0} (1-\eta) \, d\mathbf{P}_{\mathbf{X}},$$

$$(\ \eta(x) := \mathbf{P}(y = 1|x)\ )$$

$$\mathcal{R}^{*}_{\mathrm{L}_{\alpha-class}, \mathbf{P}} = \int_{\mathbf{X}} min\{(1-\alpha)\eta, \alpha(1-\eta)\} \, d\mathbf{P}_{\mathbf{X}}.$$

# 1. Loss Functions

Example2.6 *(Least squares regression)*
The goal is to predict the label $y \in \mathbb{R}$ by $t$ if $x$ is observed.
The least squares loss $\mathrm{L}_{LS} : \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is defined by

$$\mathrm{L}_{LS}(y, t) := (y - t)^2, \quad y \in \mathbf{Y}, t \in \mathbb{R}$$

# 1. Loss Functions

Def 2.7 - 2.8 *(supervised/unsupervised Loss Function)*
A function $\mathrm{L} : \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is called a ***supervised loss function*** ,
if it is m'able.
$\mathrm{L}$ can be canonically identified with the loss ftn $\bar{\mathrm{L}} : (x, y, t) \to \mathrm{L}(y, t)$.

A function $\mathrm{L} : \mathbf{X} \times \mathbb{R} \to [0, \infty)$ is called a ***unsupervised loss function*** ,
if it is m'able.
$\mathrm{L}$ can be canonically identified with the loss ftn $\bar{\mathrm{L}} : (x, y, t) \to \mathrm{L}(x, t)$.

$$\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) = \mathcal{R}_{\bar{\mathrm{L}},\mathbf{P}}(f) = \int_{\mathbf{X}} \mathrm{L}(x, f(x)) d\mathbf{P}_{\mathbf{X}}(x)$$

$$\mathcal{R}^*_{\mathrm{L},\mathbf{P}} := \mathcal{R}^*_{\bar{\mathrm{L}},\mathbf{P}}$$

# 1. Loss Functions

Example2.9 *(Density level detection Loss).*
$\mathcal{D} := (x_1, \cdots, x_n) \sim \ i.i.d.\ \mathbf{Q}$ (unkown)
The goal is to find the region where $\mathbf{Q}$ has relatively high concentration.
We assume that $\mathbf{Q}$ is abs. conti. w.r.t. some known reference measure $\mu$.
Let $g : \mathbf{X} \to [0, \infty)$ be the corresponding unknown density w.r.t. $\mu$.
$(\mathbf{Q} = g\mu)$
(Find the density level sets $\{g > \rho\}$ or $\{g \geq \rho\}$.)

$$\mathrm{L}_{LDL}(x, t) := \mathbf{1}_{(-\infty,0)}((g(x) - \rho)\text{sign } t)$$

$$\mathcal{R}_{\mathrm{L}_{LDL},\mu}(f) := \mathcal{R}_{\mathrm{L}_{LDL},\mathbf{P}}(f) = \int_{\mathbf{X}} \mathrm{L}_{DLD}(x, f(x))d\mu(x), \quad \mathbf{P_X} = \mu$$

# 1. Loss Functions

Example2.10 *(Density estimation - Unsupervised Loss).*
$\mu$ : known p.m. on $\mathbf{X}$
$g : \mathbf{X} \to [0, \infty)$ : unknown density w.r.t $\mu$
The goal is to estimate the density $g$. The unsupervised loss $\mathrm{L}_q : \mathbf{X} \times \mathbb{R} \to [0, \infty), q > 0$, defined by

$$\mathrm{L}_q(x, t) := |g(x) - t|^q, \quad x \in \mathbf{X}, t \in \mathbb{R}$$

$$\mathcal{R}_{\mathrm{L}_q, \mathbf{P}}(f) = \int_{\mathbf{X}} |g(x) - f(x)|^q d\mu(x), \quad \forall f : \mathbf{X} \to \mathbb{R} \text{ ( m'able)} \quad \mathbf{P}_{\mathbf{X}} = \mu.$$

# Outline

# 2. Basic Properties of Loss Functions and Their Risks

Lemma2.11 shows that under some circumstances risk functionals ($\mathcal{R}_{\mathrm{L},\mathbf{P}}$) are m'able.

Lemma 2.11 *(Measurability of risks)* Let $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ be a loss and $\mathcal{F} \subset \mathcal{L}_0(\mathbf{X})$ be a subset that is equipped with a complete and separable metric $d$ and its corresponding Borel $\sigma$-algebra. Assume that the metric $d$ ***dominates the pointwise convergence***, i.e.,

$$\lim_{n\to\infty} d(f, f_n) = 0 \qquad \lim_{n\to\infty} f_n(x) = f(x), x \in \mathbf{X} \forall f, f_n \in \mathcal{F}.$$

Then the evaluation map $(f, x) \to f(x)$ defined on $\mathcal{F} \times \mathbf{X}$ is measurable, and consequently the map $(x, y, f) \to \mathrm{L}(x, y, f(x))$ defined on $\mathbf{X} \times \mathbf{Y} \times \mathcal{F}$ is also measurable. Finally, given a distribution $\mathbf{P}$ on $\mathbf{X} \times \mathbf{Y}$, the risk function $\mathcal{R}_{\mathrm{L},\mathbf{P}} : \mathcal{F} \to [0, \infty)$ is measurable.

# 2. Basic Properties of Loss Functions and Their Risks

Def 2.12 *(Convexity of Loss functions)* A loss $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is called ***(strictly) convex*** if $\mathrm{L}(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is (strictly) convex $\forall x \in \mathbf{X}$ and $y \in \mathbf{Y}$.

Lemma 2.13 *(Convexity of risks)* Let $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ be a (strictly) convex loss and $\mathbf{P}$ be a distribution on $\mathbf{X} \times \mathbf{Y}$. Then $\mathcal{R}_{\mathrm{L},\mathbf{P}} : \mathcal{L}_0(\mathbf{X}) \to [0, \infty]$ is (strictly) convex.

# 2. Basic Properties of Loss Functions and Their Risks

Def 2.14 *(Continuity of Loss functions)* A loss $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is called ***(strictly) continuous*** if $\mathrm{L}(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is continuous $\forall x \in \mathbf{X}$ and $y \in \mathbf{Y}$.

- In general, $\mathrm{L}(x, y, f_n(x)) \to \mathrm{L}(x, y, f(x)), \forall (x, y)$ does not imply $\mathcal{R}_{\mathrm{L},\mathbf{P}}(f_n) \to \mathcal{R}_{\mathrm{L},\mathbf{P}}(f)$

Lemma 2.15 *(Lower semi-continuity of risks)* Let $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ be a continuous loss, $\mathbf{P}$ be a distribution on $\mathbf{X} \times \mathbf{Y}$, and $(f_n) \subset \mathcal{L}_0(\mathbf{P_X})$ be a seq. that converges to an $f \in \mathcal{L}_0(\mathbf{P_X})$ in prob. w.r.t. $\mathbf{P_X}$. Then we have

$$\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) \leq \liminf_{n\to\infty} \mathcal{R}_{\mathrm{L},\mathbf{P}}(f_n)$$

## 2. Basic Properties of Loss Functions and Their Risks

Def 2.16 *(Nemitski loss )*
We call a loss $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ a ***Nemitski loss*** if $\exists$ a m'able ftn $b : \mathbf{X} \times \mathbf{Y} \to [0, \infty)$ and an increasing ftn $h : [0, \infty) \to [0, \infty)$ s.t.

$$\mathrm{L}(x, y, t) \leq b(x, y) + h(|t|), \quad (x, y, t) \in \mathbf{X} \times \mathbf{Y} \times \mathbb{R}$$

We say that $\mathrm{L}$ is a ***Nemitski loss of order*** $p \in (0, \infty)$ if $\exists$ a constant $c > 0$ s.t

$$\mathrm{L}(x, y, t) \leq b(x, y) + c|t|^p, \quad (x, y, t) \in \mathbf{X} \times \mathbf{Y} \times \mathbb{R}$$

If $\mathbf{P}$ is a dist.n on $\mathbf{X} \times \mathbf{Y}$ with $b \in \mathcal{L}_1(\mathbf{P})$, we say that $\mathrm{L}$ is a ***P-integrable Nemitski loss.***

- The notion of Nemitski losses will become of particular interest when dealing with unbounded $\mathbf{Y}$.(reg. problem)

Lemma 2.17 *(Continuity of risks)*
Let $\mathbf{P}$ be a distribution on $\mathbf{X} \times \mathbf{Y}$ and $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ be a continuous, P-integrable Nemitski loss. Then the following statements hold:

i) Let $f_n : \mathbf{X} \to \mathbb{R}, n \geq 1$, be bdd m'able ftns for which $\exists$ a constant $\mathbf{B} > 0$ with $||f_n||_\infty \leq \mathbf{B}\ \forall n \geq 1$. If the seq. $(f_n) \to f\ \mathbf{P_X} - a.s.$, then we have

$$\lim_{n \to \infty} \mathcal{R}_{\mathrm{L},\mathbf{P}}(f_n) = \mathcal{R}_{\mathrm{L},\mathbf{P}}(f)$$

ii) The map $\mathcal{R}_{\mathrm{L},\mathbf{P}} : \mathrm{L}_\infty(\mathbf{P_X}) \to [0, \infty)$ is well-defined and continuous.
iii) If $\mathrm{L}$ is of order $p \in [1, \infty)$, then $\mathcal{R}_{\mathrm{L},\mathbf{P}} : \mathrm{L}_p(\mathbf{P_X}) \to [0, \infty)$ is well-defined and continuous.

# 2. Basic Properties of Loss Functions and Their Risks

Def 2.18 ( *Locally Lipschitz continuous* )
A loss $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is called ***locally Lipschitz continuous*** if $\forall a \geq 0 \; \exists$ a constant $c_a \geq 0$ s.t.

$$\sup_{x \in \mathbf{X}, y \in \mathbf{Y}} |\mathrm{L}(x, y, t) - \mathrm{L}(x, y, t')| \leq c_a |t - t'|, \quad t, t' \in [-a, a].$$

For $a \geq 0$, the smallest $c_a$ is denoted by $|\mathrm{L}|_{a,1}$.
If we have $|\mathrm{L}|_1 := sup_{a \geq 0} |\mathrm{L}|_{a,1} < \infty$, we call $\mathrm{L}$ ***Lipschitz continuous*** .

- Every convex function is locally Lipschitz continuous.
- Locally Lipschitz continuous loss $\mathrm{L}$ is a Nemitski loss.

Lemma 2.19 *(Lipschitz continuity of risks).* Let $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ be a locally Lipschitz continuous loss and $\mathbf{P}$ be a distn on $\mathbf{X} \times \mathbf{Y}$. Then $\forall \mathbf{B} \geq 0$ and all $f, g, \in \mathrm{L}_\infty(\mathbf{P_X})$ with $||f||_\infty \leq \mathbf{B}$ and $||g||_\infty \leq \mathbf{B}$, we have

$$|\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) - \mathcal{R}_{\mathrm{L},\mathbf{P}}(g)| \leq |\mathrm{L}|_{|b,1}| |f - g||_{\mathrm{L}_1(\mathbf{P_X})}.$$

# 2. Basic Properties of Loss Functions and Their Risks

Def 2.20 *( Differentiability )*
A loss $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is called ***differentiable*** if
$\mathrm{L}(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is differentiable $\forall x \in \mathbf{X}, y \in \mathbf{Y}$.
$\mathrm{L}'(x, y, t)$ denotes the derivative of $\mathrm{L}(x, y, \cdot)$ at $t \in \mathbb{R}$

- For certain integrable Nemitski losses, we can actually establish the differentiability of the associated risk.

Lemma 2.21 *(Differentiability of risks).*
Let $\mathbf{P}$ be a dist. on $\mathbf{X} \times \mathbf{Y}$ and $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ be a diff'able loss s.t. both $\mathrm{L}$ and $|\mathrm{L}'| : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ are P-integrable Nemitski losses. Then the risk functional $\mathcal{R}_{\mathrm{L},\mathbf{P}} : \mathrm{L}_\infty(\mathbf{P_X}) \to [0, \infty)$ is Frechet differentiable and its derivative at $f \in \mathrm{L}_\infty(\mathbf{P_X})$ is the bdd linear operator $\mathcal{R}'_{\mathrm{L},\mathbf{P}}(f) : \mathrm{L}_\infty(\mathbf{P_X}) \to \mathbb{R}$ given by

$$\mathcal{R}'_{\mathrm{L},\mathbf{P}}(f)g = \int_{\mathbf{X}\times\mathbf{Y}} g(x)\mathrm{L}'(x, y, f(x))d\mathbf{P}(x, y), \quad g \in \mathrm{L}_\infty(\mathbf{P_X}).$$

# 2. Basic Properties of Loss Functions and Their Risks

Def 2.22 *( Clipped loss : Restriction to domains of the form* $\mathbf{X} \times \mathbf{Y} \times [\,-\mathbf{M}, \mathbf{M}]$ *)*
We say that a loss $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ can be ***clipped*** at $M > 0$ if, $\forall (x, y, t) \in \mathbf{X} \times \mathbf{Y} \times \mathbb{R}$, we have

$$\mathrm{L}(x, y, \hat{t}) \le \mathrm{L}(x, y, t),$$

where $\hat{t}$ denotes the ***clipped value*** of $t$ at $\pm\mathbf{M}$ , that is

$$\hat{t} := \begin{cases} -\mathbf{M} & \text{if } t < -\mathbf{M} \\ t & \text{if } t \in [-\mathbf{M}, \mathbf{M}] \\ \mathbf{M} & \text{if } t > \mathbf{M} \end{cases}$$

We say that $\mathrm{L}$ can be clipped if it can be clipped at some $\mathbf{M} > 0$

Lemma 2.23 *(Clipped convex losses).*
Let $\mathrm{L} : \mathbf{X} \times \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss and $\mathbf{M} > 0$. Then the following statements are equivalent:
i) $\mathrm{L}$ can be clipped at $\mathbf{M}$.
ii) $\forall (x, y) \in \mathbf{X} \times \mathbf{Y}$, the function $\mathrm{L}(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ has at least one global minimizer in $[-\mathbf{M}, \mathbf{M}]$

# Outline

# 3. Margin-Based Losses for Classification Problems

- Both $\mathrm{L}_{class}$ and $\mathrm{L}_{\alpha-calss}$ are not convex, which may lead to computational problems to minimize an empirical risk $\mathcal{R}_{\mathrm{L}_{class},\mathbf{D}}(\cdot)$ over some set $\mathcal{F}$.
- The empirical risk $\mathcal{R}_{\mathrm{L},\mathbf{D}}(\cdot)$ of a surrogate loss function $\mathrm{L}$ is used in SVMs. (Hinge loss).

# 3. Margin-Based Losses for Classification Problems

Def 2.24 *( Margin-based Loss )*
A supervised loss $\mathrm{L} : \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is called ***margin-based*** if there exists a ***representing function*** $\varphi : \mathbb{R} \to [0, \infty)$ s.t.

$$\mathrm{L}(y, t) = \varphi(yt), \quad y \in \mathbf{Y}, t \in \mathbb{R}$$

Lemma 2.25 *(Properties of margin-based losses).*
Let $\mathrm{L}$ be a margin-based loss represented by $\varphi$
i) $\mathrm{L}$ is (strictly) convex. $\iff$ $\varphi$ is (strictly) convex.
ii) $\mathrm{L}$ is continuous. $\iff$ $\varphi$ is.
iii) $\mathrm{L}$ is (locally) Lipschitz continuous. $\iff$ $\varphi$ is.
iv) $\mathrm{L}$ is convex. $\implies$ It is locally Lipschitz continuous.
v) $\mathrm{L}$ is a P-integrable Nemitski loss for all m'able spaces $\mathbf{X}$ and all dist. $\mathbf{P}$ on $\mathbf{X} \times \mathbf{Y}$ .

# 3. Margin-Based Losses for Classification Problems

## Margin-Based Losses

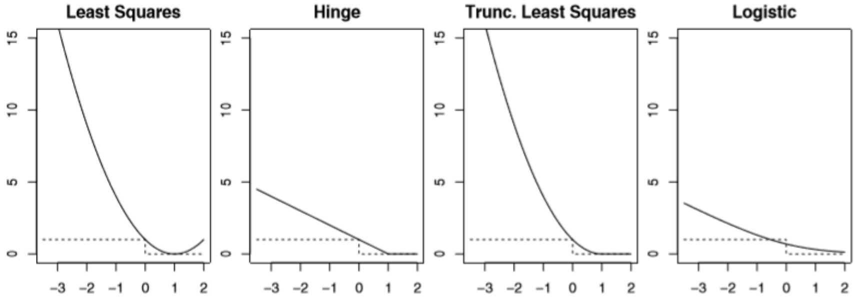



Figure: The shape of the representing function $\varphi$ for some margin-based loss functions.

# 3. Margin-Based Losses for Classification Problems

Example 2.27 *( Hinge loss )*
The ***hinge loss*** $\mathrm{L}_{hinge} : \mathbf{Y} \times \mathbb{R} \to [0, \infty)$ is defined by

$$\mathrm{L}_{hinge}(y, t) := \max\{0, 1 - yt\}, \quad y = \pm 1, t \in \mathbb{R}$$

$\Rightarrow \mathrm{L}_{hinge}$ is margin-based loss. It is convex and Lipschitz conti. with $|\mathrm{L}_{hinge}|_1 = 1$. Finally, $\mathrm{L}_{hinge}$ can be clipped at $\mathbf{M} = 1$.

Example 2.28 *( Truncated least squares loss = Squared hinge loss )*
The ***truncated least squares loss*** $\mathrm{L}_{trunc-ls}$ is defined by

$$\mathrm{L}_{trunc-ls}(y,t) := (max\{0, 1-yt\})^2, \quad y = \pm 1, t \in \mathbb{R}$$

$\Rightarrow \mathrm{L}_{trunc-ls}$ is margin-based loss. It is convex and Lipschitz constants are $|\mathrm{L}_{trunc-ls}|_{a,1} = 2a+2$, $a > 0$. Finally, $Loss_{trunc-ls}$ can be clipped at $\mathbf{M} = 1$.

# 3. Margin-Based Losses for Classification Problems

Example 2.28 *( Logistic loss for classification )*
The ***logistic loss for classification*** $\mathrm{L}_{c-logit}$ is defined by

$$\mathrm{L}_{c-logit}(y,t) := ln(1 + exp(-yt)), \quad y = \pm 1, t \in \mathbb{R}$$

$\Rightarrow \mathrm{L}_{c-logit}$ is margin-based loss. It is infinitely many times differentiable, convex and Lipschitz conti. with $|\mathrm{L}_{c-logit}|_1 = 1$. Finally, $Loss_{trunc-ls}$ cannot be clipped .

# 3. Margin-Based Losses for Classification Problems

Thm 2.31 *(Zhang's inequality )*
Given a dist. $\mathbf{P}$ on $\mathbf{X} \times \mathbf{Y}$, we write $\eta(x) := \mathbf{P}(y = 1|x), x \in \mathbf{X}$.
Let $f^*_{\mathrm{L}_{class},\mathbf{P}}$ be the Bayes classification ftn given by
$f^*_{\mathrm{L}_{class},\mathbf{P}}(x) := sign(2\eta(x) - 1), x \in \mathbf{X}$.
Then, $\forall$ m'able $f : \mathbf{X} \to [-1, 1]$, we have

$$\mathcal{R}_{\mathrm{L}_{hinge},\mathbf{P}}(f) - \mathcal{R}^*_{\mathrm{L}_{hinge},\mathbf{P}} = \int_{\mathbf{X}} |f(x) - f^*_{\mathrm{L}_{class},\mathbf{P}}(x)|$$

Moreover, for every measurable $f : \mathbf{X} \to \mathbb{R}$, we have

$$\mathcal{R}_{\mathrm{L}_{class},\mathbf{P}}(f) - \mathcal{R}^*_{\mathrm{L}_{class},\mathbf{P}} \leq \mathcal{R}_{\mathrm{L}_{hinge},\mathbf{P}}(f) - \mathcal{R}^*_{\mathrm{L}_{hinge},\mathbf{P}}$$

# Outline

# 4. Distance-Based Losses for Regression Problems

Def 2.32 *(Distance-based loss )*
We say that a supervised loss $\mathrm{L} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is :
i) ***distance-based*** if there exists a ***representing function*** $\psi : \mathbb{R} \to [0, \infty)$ satisfying $\psi(0) = 0$ and

$$\mathrm{L}(y, t) = \psi(y - t), \quad y \in \mathbf{Y}, t \in \mathbb{R};$$

ii) ***symmetric*** if *Loss* is distance-based and its representing function $\psi$ satisfies

$$\psi(r) = \psi(-r), \quad r \in \mathbb{R}$$

Lemma 2.33 *(Properties of distance-based losses).*
Let $\mathrm{L}$ be a distance-based loss with representing function $\psi : \mathbb{R} \to [0, \infty)$.
i) $\mathrm{L}$ is (strictly) convex.$\Longleftrightarrow$ $\psi$ is (strictly) convex.
ii) $\mathrm{L}$ is conti $\Longleftrightarrow$ $\psi$ is conti.
iii) $\mathrm{L}$ is Lipschitz conti.$\Longleftrightarrow$ $\psi$ is Lipschitz conti.

# 4. Distance-Based Losses for Regression Problems

- Our goal is to investigate under which conditions on the dist. **P** a distance-based loss ftn is a P-integrable Nemitski loss.

i) the analysis of the integrals of the form

$$\mathcal{C}_{\mathrm{L},\mathbf{Q}}(t) := \int_{\mathbb{R}} \mathrm{L}(y,t)d\mathbf{Q}(y), \quad \mathbf{Q} := \mathbf{P}(\mathbf{Y}|x)$$

ii) analysis of the averaging w.r.t. $\mathbf{P_X}$

Def 2.34 *(p-th moment )*
For a distribution $\mathbf{Q}$ on $\mathbb{R}$, the p-th moment, p $\in (0,\infty)$, is defined by

$$|\mathbf{Q}|_p := (\int_{\mathbb{R}} |y|^p d\mathbf{Q}(y))^{1/p}.$$

Its $\infty$-moment is defined by $|\mathbf{Q}|_\infty := sup|supp\mathbf{Q}|$.

# 4. Distance-Based Losses for Regression Problems

Def 2.35 *(growth behavior )*
Let $p \in (0, \infty) and \mathrm{L} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a distance-based loss with representing function $\psi$. We say that *Loss* is of:
i) ***upper growth*** $p$ if there is a constant $c > 0$ s.t.

$$\psi(r) \leq (|r|^p + 1), \quad r \in \mathbb{R};$$

ii) ***lower growth*** $p$ if there is a constant $c > 0$ s.t.

$$\psi(r) \geq (|r|^p - 1), \quad r \in \mathbb{R};$$

iii) ***growth type*** $p$ if $\mathrm{L}$ is of both upper and lower growth type $p$.

# 4. Distance-Based Losses for Regression Problems

- For convex distance-based loss ftns $\mathrm{L}$ , the representing $\psi$ is locally Lipschitz conti. on every interval $[-r, r]$.
- $r \to |\psi_{|[-r,r]}|_1, r \geq 0$ defines an increasing, non-negative function

Lemma 2.36 *(Growth type and moments)*
Let $\mathrm{L}$ be a distance-based loss with representing function $\psi$ and $\mathbf{Q}$ be a distribution on $\mathbb{R}$. For $p \in (0, \infty)$, we then have:
i) If $\psi$ is convex and $\lim_{|r|\to\infty} \psi(r) = \infty$, then $\mathrm{L}$ is of lower growth type 1.
ii) If $\psi$ is Lipschitz conti., then $\mathrm{L}$ is of upper growth type 1.
iii) If $\psi$ is convex, then $\forall r > 0$ we have

$$|\psi_{|[-r,r]}|_1 \leq \frac{2}{r}||\psi_{|[-2r,2r]}||_\infty \leq 4|\psi_{|[-2r,2r]}|_1.$$

iv) If $\mathrm{L}$ is convex and of upper growth type 1, then it is Lipschitz continuous.

# 4. Distance-Based Losses for Regression Problems

Lemma 2.36 *(Properties of distance-based losses)*
v) If $\mathrm{L}$ is of upper growth type $p$, then there exists a constant $c_{\mathrm{L},p} > 0$ independent of $\mathbf{Q}$ s.t

$$\mathcal{C}_{\mathrm{L},\mathbf{Q}}(t) \leq c_{\mathrm{L},p}(|\mathbf{Q}|_p^p + |t|^p + 1), \quad t \in \mathbb{R}.$$

$\mathrm{L}$ is a Nemitski loss of order $p$.
vi) If $\mathrm{L}$ is of lower growth type $p$, then there exists a constant $c_{\mathrm{L},p} > 0$ independent of $\mathbf{Q}$ s.t

$$|\mathbf{Q}|_p^p \leq c_{\mathrm{L},p}(\mathcal{C}_{\mathrm{L},\mathbf{Q}}(t) + |t|^p + 1), \quad t \in \mathbb{R}.\text{and}$$

$$|t|^p \leq c_{\mathrm{L},p}(\mathcal{C}_{\mathrm{L},\mathbf{Q}}(t) + |\mathbf{Q}|_p^p + 1), \quad t \in \mathbb{R}.$$

vii) If $\mathrm{L}$ is of growth type $p$, then we have $\mathcal{C}^*_{\mathrm{L},\mathbf{Q}} < \infty$ if and only if $|Q\|_p < \infty$.

# 4. Distance-Based Losses for Regression Problems

Def 2.37 *( average p-th moment )*
For a distribution $\mathbf{P}$ on $\mathbf{X} \times \mathbb{R}$, the ***average p-th moment***, $p \in (0, \infty)$, is defined by

$$|\mathbf{P}|_p := \left(\int_{\mathbf{X}} \int_{\mathbb{R}} |y|^p d\mathbf{P}(x, y)\right)^{1/p} = \left(\int_{\mathbf{X}} |\mathbf{P}(\cdot|x)|_p^p d\mathbf{P}_{\mathbf{X}}(x)\right)^{1/p}.$$

Its average 0-moment is defined by $|\mathbf{P}|_0 := 1$ and its average $\infty$-moment is defined by $|\mathbf{P}|_\infty := \text{ess-sup}_{x \in \mathbf{X}} |\mathbf{P}(\cdot|x)|_\infty$.

# 4. Distance-Based Losses for Regression Problems

Lemma 2.38 *(Average moments and risks).*
Let $\mathrm{L}$ be a distance-based loss and **P** be a distribution on$\mathbf{X} \times \mathbf{Y}$. For $p > 0$, we then have:
i) If $\mathrm{L}$ is of upper growth type $p$, there exists a constant $c_{\mathrm{L},p} > 0$ indep. of **P** s.t., $\forall$ m'able $f : \mathbf{X} \to \mathbb{R}$, we have

$$\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) \leq c_{\mathrm{L},p}(\mathbf{P}|_p^p + ||f||_{\mathrm{L}_p(\mathbf{P_X})}^p + 1).$$

If, $|\mathbf{P}|_p < \infty$ ,then $\mathrm{L}$ is a **P**-integrable Nemitski loss of order $p$, and $\mathcal{R}_{\mathrm{L},\mathbf{P}}$ is well-defined and conti.

# 4. Distance-Based Losses for Regression Problems

Lemma 2.38 *(Average moments and risks).*
ii) If $\mathrm{L}$ is convex and of upper growth type $p$ with $p \geq 1$, then $\forall q \in [p-1, \infty]$ with $q > 0$ $\exists$ a constant $c_{\mathrm{L},p,q} > 0$ indep. of $\mathbf{P}$ s.t., $\forall$ m'able $f : \mathbf{X} \to \mathbb{R}$ *and* $\mathbf{g} : \mathbf{X} \to \mathbb{R}$, we have

$$|\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) - \mathcal{R}_{\mathrm{L},\mathbf{P}}(g)| \\ \leq c_{\mathrm{L},p,q}(|\mathbf{P}|_q^{p-1} + ||f||_{\mathrm{L}_q(\mathbf{P_X})}^{p-1} + ||g||_{\mathrm{L}_q(\mathbf{P_X})}^{p-1} + 1)||f-g||_{\mathrm{L}_{\frac{q}{q-p+1}}\mathbf{P_X}}.$$

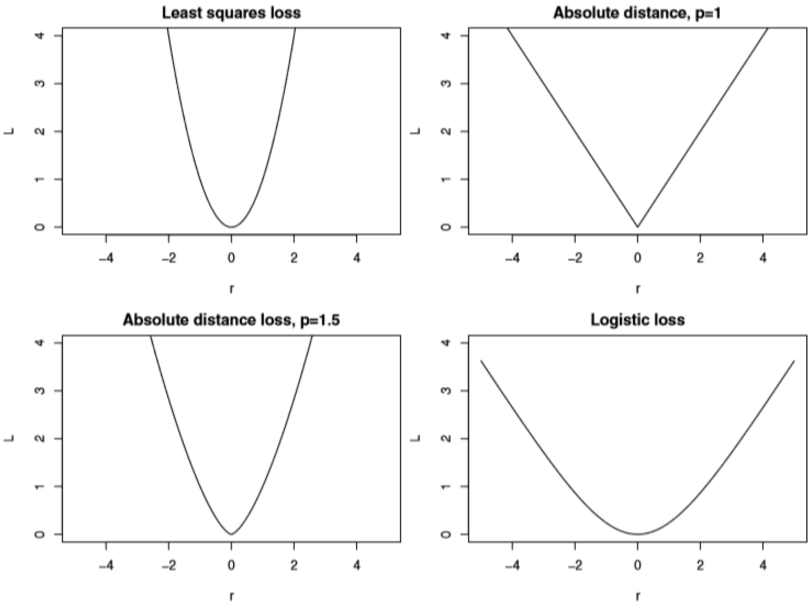
# 4. Distance-Based Losses for Regression Problems

Lemma 2.38 *(Average moments and risks).*
iii) If $\mathrm{L}$ is lower growth type, $\exists$ a constant $c_{\mathrm{L},p} > 0$ indep. of $\mathbf{P}$ s.t., $\forall$ m'able $f : \mathbf{X} \to \mathbb{R}$, we have

$$|\mathbf{P}|_p^p \leq c_{\mathrm{L},p}(\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) + ||f||^p_{\mathrm{L}_{\mathbf{P}}(\mathbf{P_X})} + 1) \quad \textit{and}$$
$$||f||^p_{\mathrm{L}_{\mathbf{P}}(\mathbf{P_X})} \leq c_{\mathrm{L},p}(\mathcal{R}_{\mathrm{L},\mathbf{P}}(f) + |\mathbf{P}|_p^p + 1).$$

# 4. Distance-Based Losses for Regression Problems

## Margin-Based Losses



# 4. Distance-Based Losses for Regression Problems

## Margin-Based Losses

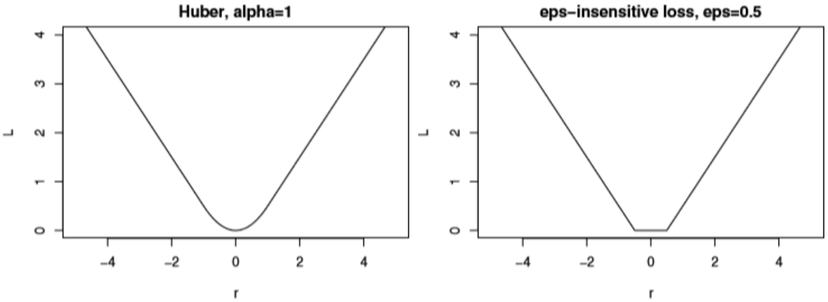



Figure: The shape of the representing function $\psi$ for some distance-based loss functions.

# 4. Distance-Based Losses for Regression Problems

Example 2.39 ( *p-th power absolute distance loss* )
For $p > 0$, the ***p-th power absolute distance loss*** $\mathrm{L}_{p-dist}$ is the distance-based loss function represented by

$$\psi(r) := |r|^p, \quad r \in \mathbb{R}.$$

$\Rightarrow p = 2 : \mathrm{L}_{p-dist}$ is the least squares loss.
$\Rightarrow p = 1 : \mathrm{L}_{p-dist}$ is the absolute distance loss.
$\Rightarrow p \geq 1 : \mathrm{L}_{p-dist}$ is growth type $p$ and $\mathrm{L}_{p-dist}$ is convex.
$\Rightarrow p > 1 \iff \mathrm{L}_{p-dist}$ is strictly convex .
$\Rightarrow p = 1 \iff \mathrm{L}_{p-dist}$ is Lipschitz conti.

Example 2.40 ( *logistic loss for regression* )
The distance-based ***logistic loss for regression*** $\mathrm{L}_{r-logist}$ is represented by

$$\psi(r) : - = -ln\frac{4e^r}{(1+e^r)^2}, \quad r \in \mathbb{R}.$$

$\Rightarrow \mathrm{L}_{r-logist}$ is strictly convex and Lipschitz continuous, and consequently $\mathrm{L}_{r-logist}$ is of growth type 1.

# 4. Distance-Based Losses for Regression Problems

Example 2.41 *( Huber's loss )*
For $\alpha > 0$, ***Huber's loss*** $\mathrm{L}_{\alpha-\text{Hubor}}$ is the distance-based loss represented by

$$\psi(r) := \begin{cases} \frac{r^2}{2} & \text{if } |r| \leq \alpha \\ \alpha|r| - \frac{\alpha^2}{2} & \text{o.w.} \end{cases}$$

$\Rightarrow \mathrm{L}_{\alpha-\text{Hubor}}$ is convex but not strictly convex. Furthermore, it is Lipschitz continuous, and thus $\mathrm{L}_{\alpha-\text{Hubor}}$is of growth type 1. The derivative of $\psi$ equals the clipping operation for $\mathbf{M} = \alpha$.

# 4. Distance-Based Losses for Regression Problems

Example 2.42 ( *$\epsilon$-insensitive loss* )
The ***$\epsilon$-insensitive loss*** $\mathrm{L}_{\epsilon\text{-insens}}$ is represented by

$$\psi(r) := max\{0, |r| - \epsilon\}, \quad r \in \mathbb{R}.$$

$\Rightarrow$ $\mathrm{L}_{\epsilon\text{-insens}}$ ignores deviances smaller than $\epsilon$.
$\Rightarrow$ $\mathrm{L}_{\epsilon\text{-insens}}$ is Lipschitz conti. and convex but not strictly convex. It is of growth type 1.
$\Rightarrow$ $\mathrm{L}_{\epsilon\text{-insens}}$ can be used to estimate the conditional median.

Example 2.42 ( *Pinball loss* )
For $\tau \in (0, 1)$, the ***pinball loss*** $\mathrm{L}_{\tau\text{-pin}}$ is represented by

$$\psi(r) := \begin{cases} -(1-\tau)r, & \text{if } r < 0 \\ \tau r & \text{if } r \geq 0 \end{cases}$$

$\Rightarrow$ $\mathrm{L}_{\tau\text{-pin}}$ is Lipschitz conti. and convex. (But for $\tau \neq 1/2$ it is not symm.)
$\Rightarrow$ $\mathrm{L}_{\tau\text{-pin}}$ can be used to estimate condi. $\tau$-quantiles .