THE UNIVERSITY OF SYDNEY

Advanced Machine Learning

(COMP 5328)

Loss Functions and Convex Optimisation

Tongliang Liu

THE UNIVERSITY OF SYDNEY

Review

THE UNIVERSITY OF SYDNEY

assumed function to predict

Hypothesis

hypothesis class

this class contains 2 hypothesis

$$H_1 = \{h_1(x), h_2(x) : x \in \mathbb{R}\}.$$

infinitely many hypothesis

$$H_2 = \{h(x) = w_0 + w_1 x + w_2 x^2 : x, w_0, w_1, w_2 \in \mathbb{R}\}.$$

$$H_3 = \{h(x) = w^\top x : x, w \in \mathbb{R}^d\}.$$

$$H_4 = \{h(x) = \text{sgn}(p(y = 1|x, \theta) - 0.5) : x \in \mathbb{R}^d, \theta \in \Theta\}.$$

What is Machine Learning?

THE UNIVERSITY OF SYDNEY

Informally: Making predictions from data.

Formally: The construction of a statistical model that is an underlying distribution from which the data is drawn from.

Mathematically: A machine learning algorithm is a mapping to find a hypothesis to fit the data

$$\mathcal{A} : S \in (\mathcal{X} \times \mathcal{Y})^n \mapsto h_S \in H.$$

mapping — map to — each sample — n training samples

Elements of Machine Learning Algorithms

- I. Input training data
- II. Predefined hypothesis class
- III. Objective function
- IV. Optimisation method
- V. Output hypothesis

Objective function

- Given a classification task, we should firstly defined which hypothesis or classifier is the best.
- One intuitive way to defined the best classifier: the classifier that has the minimum classification error on the all possible data generated from the task.

Best classifier

THE UNIVERSITY OF SYDNEY

- For a given data point (X, Y), the classification error for a hypothesis h is measured by the 0-1 loss function:

$$1_{\{Y \neq \text{sign}(h(X))\}} = \begin{cases} 0 & Y = \text{sign}(h(X)) \\ 1 & Y \neq \text{sign}(h(X)) \end{cases}$$

right prediction

wrong prediction

- The best classifier can be mathematically defined as:

$$\arg\min_h \frac{1}{|D|} \sum_{i \in D} 1_{\{Y_i \neq \text{sign}(h(X_i))\}}$$

where D is the set of indices of **all possible data** points of the task, and $|D|$ denotes the size of the set D.

cardinality

The law of large numbers

THE UNIVERSITY OF SYDNEY

LLN describes the result of performing the same experiment a large number of times.

Let be $x_1, x_2, \ldots, x_n$ iid examples drawn from distribution D. Then

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \xrightarrow{n\to\infty} \mathbb{E}_{X\sim D}[f(X)].$$

The law of large numbers

THE UNIVERSITY OF SYDNEY

LLN describes the result of performing the same experiment a large number of times.

Toss a coin

$$E(X) = \int P(x) \cdot x \, dx$$

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i = \text{“}head\text{”}\}} \xrightarrow{n\to\infty} \mathbb{E}[1_{\{X = \text{“}head\text{”}\}}]$$

$$P(X = \text{“}head\text{”})\, 1_{\{\text{“}head\text{”}\}} + P(X = \text{“}tail\text{”})\, 1_{\{\text{“}tail\text{”}\}}$$

↑ binary function

$$= \int P(X = \text{“}head\text{”}) 1_{\{x = \text{“}head\text{”}\}} dx = P(X = \text{“}head\text{”}).$$

The law of large numbers

THE UNIVERSITY OF SYDNEY

LLN describes the result of performing the same experiment a large number of times.

The average of the results obtained from a large number of independent trials should converge to the expected value.

$$\frac{1}{|D|}\sum_{i\in D} 1_{\{Y_i \neq \text{sign}(h(X_i))\}} \xrightarrow{|D|\to\infty} \mathbb{E}\left[1_{\{Y \neq \text{sign}(h(X))\}}\right]$$

Best classifier

THE UNIVERSITY OF SYDNEY

- The best classifier (accuracy) can be mathematically defined as:

$$\arg\min_{h} \mathbb{E}[1_{\{Y \neq \text{sign}(h(X))\}}]$$

- The distribution of data is unknown. We cannot calculate the expectation.

III. Objective function

- Given a classification task, we want to find a classifier such that the following is minimised:

$$\mathbb{E}[1_{\{Y \neq \text{sign}(h(X))\}}]$$

- We don't have the distribution of data. Fortunately, we have some examples (or a training sample) draw from the distribution:

$$S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

- Because of the law of large numbers, we can use

$$\frac{1}{n} \sum_{i=1}^{n} 1_{\{Y \neq \text{sign}(h(X))\}}$$

(unbiased estimator)

to estimate $\mathbb{E}[1_{\{Y \neq \text{sign}(h(X))\}}]$

Objective function

- The empirical estimator is unbiased because

$$\arg\min_h \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \neq h(X_i)\}} \overset{n\to\infty}{\to} \arg\min_h \mathbb{E}\left[1_{\{Y \neq h(X)\}}\right]$$

- This also explains why big data is very helpful.

- The objective function is not convex or smooth, hard to optimise.

0-1 loss function

THE UNIVERSITY OF SYDNEY

Loss functions

Best classifier

THE UNIVERSITY OF SYDNEY

- The best classifier can be mathematically defined as:

$$\arg\min_{h} \mathbb{E}[1_{\{Y \neq \text{sign}(h(X))\}}]$$

- Some problems: 1, the distribution of data is unknown. We cannot calculate the expectation. 2, the objective function is not convex or smooth, hard to optimise. 3, what kind of hypothesis h should we employ to fit the data?

Surrogate loss functions

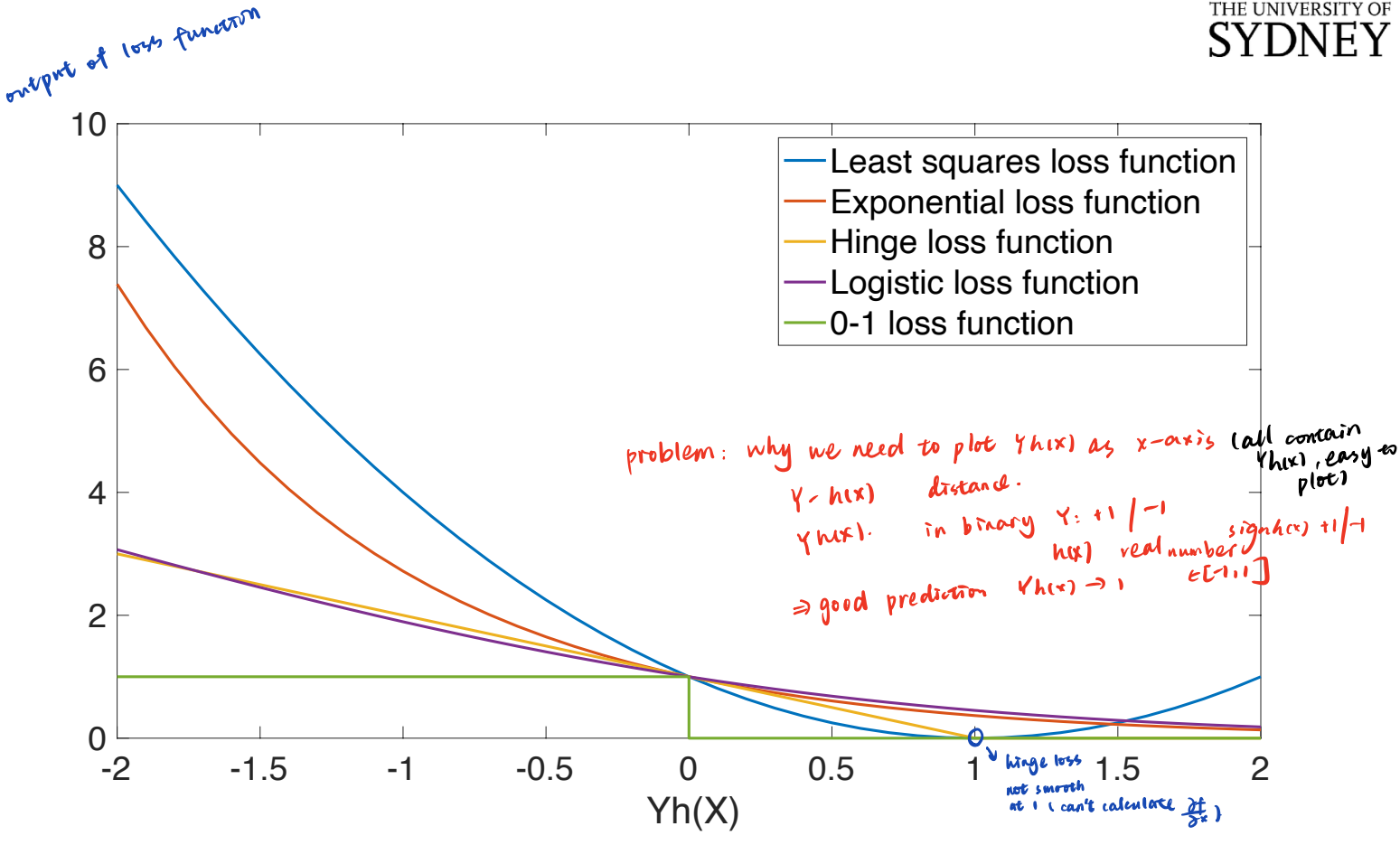
- Most optimisation methods exploit the derivative information. However, the 0-1 loss function is non-smooth and thus is non-differentiable.
- Convex objective has only one minimum. The convexity makes optimisation easier than the general case since local minimum must be a global minimum.
- Can we find some surrogate loss functions to approximate the 0-1 loss function, which are both smooth and convex?

Surrogate loss functions

THE UNIVERSITY OF SYDNEY

- Popular surrogate loss functions:
- Hinge loss: $\ell(X, Y, h) = \max\{0, 1 - Yh(X)\}$
- Logistic loss: $\ell(X, Y, h) = \log_2(1 + \exp(-Yh(X)))$
- Least square loss: $\ell(X, Y, h) = (Y - h(X))^2$
- Exponential loss: $\ell(X, Y, h) = \exp(-Yh(X))$

Surrogate loss functions



Surrogate loss functions

THE UNIVERSITY OF SYDNEY

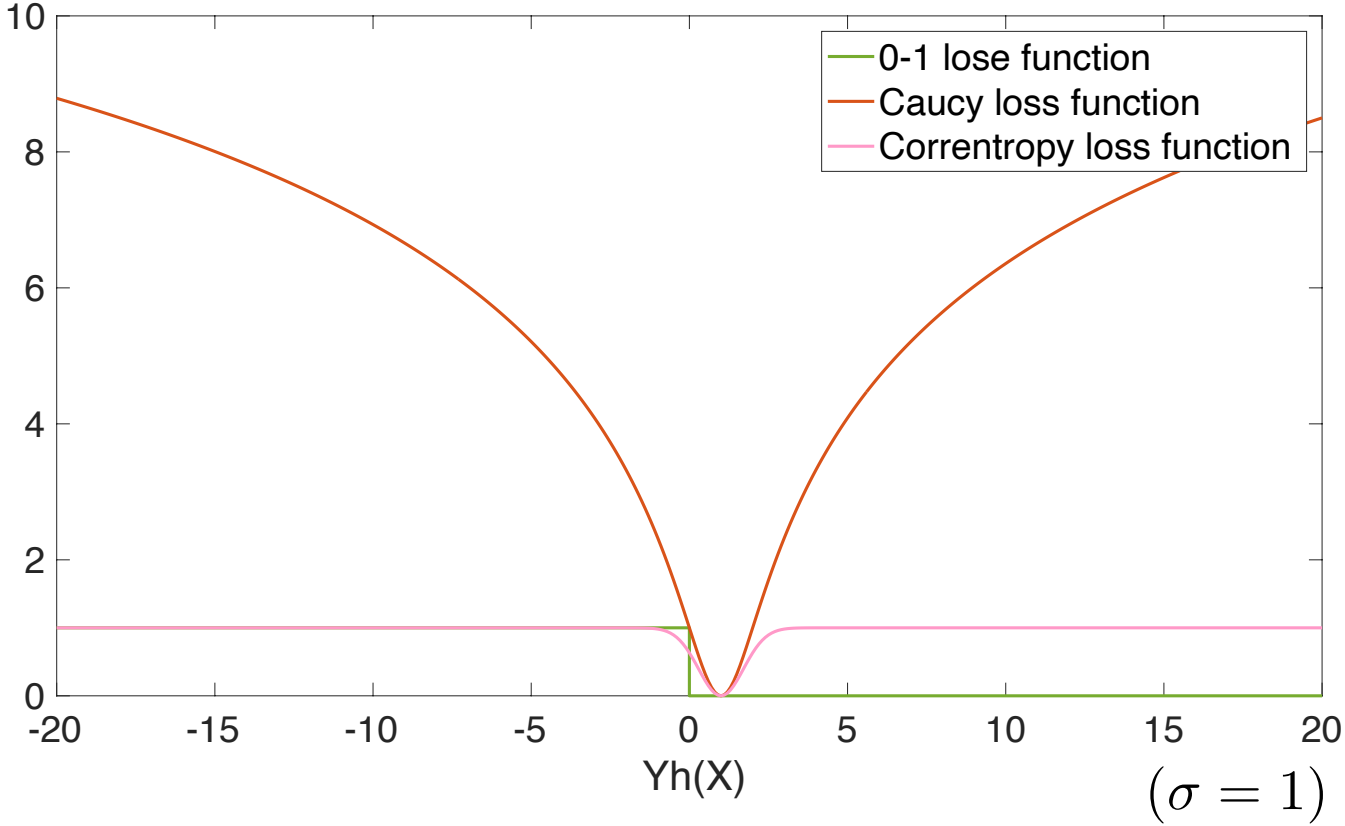
- Not all surrogate loss functions are convex
- Cauchy loss:

$$\ell(X, Y, h) = \log_2 \left(1 + \left(\frac{1 - Yh(X)}{\sigma}\right)^2\right)$$

- Correntropy loss (Welsch loss):

$$\ell(X, Y, h) = \left(1 - \exp\left(-\left(\frac{1 - Yh(X)}{\sigma}\right)^2\right)\right)$$

Surrogate loss functions



Surrogate loss functions

THE UNIVERSITY OF SYDNEY

- We have two natural questions:
- What are the differences between the 0-1 loss function and the surrogate loss functions?
- What are the differences among those different surrogate loss functions? (We will provide an answer to this question in Week 7.)

similarity between 0-1 loss & surrogate
when predictions are correct, close to 0
incorrect, output large numbers

Surrogate loss functions

THE UNIVERSITY OF SYDNEY

- What are the differences between the 0-1 loss function and the surrogate loss functions?
- Classification-calibrated surrogate loss functions: which will result in the same classifier (same accuracy) as the 0-1 loss function if the training data is sufficiently large (an asymptotical property).
- Most of the popularly used surrogate loss functions are all classification-calibrated surrogate loss functions.

Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. "Convexity, classification, and risk bounds." Journal of the American Statistical Association 101.473 (2006): 138-156.

Zhang, Jingwei, Tongliang Liu, and Dacheng Tao. "On the Rates of Convergence from Surrogate Risk Minimizers to the Bayes Optimal Classifier." arXiv preprint arXiv:1802.03688 (2018).

Surrogate loss functions

- How to check if a given surrogate loss function is a classification-calibrated surrogate loss functions?

Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. "Convexity, classification, and risk bounds." Journal of the American Statistical Association 101.473 (2006): 138-156.

Zhang, Jingwei, Tongliang Liu, and Dacheng Tao. "On the Rates of Convergence from Surrogate Risk Minimizers to the Bayes Optimal Classifier." arXiv preprint arXiv:1802.03688 (2018).

Surrogate loss functions

THE UNIVERSITY OF SYDNEY

- Popular surrogate loss functions:
- Hinge loss: $\ell(X, Y, h) = \max\{0, 1 - Yh(X)\}$
- Logistic loss: $\ell(X, Y, h) = \log_2(1 + \exp(-Yh(X)))$
- Least square loss: $\ell(X, Y, h) = (Y - h(X))^2 = (1 - Yh(X))^2$
- Exponential loss: $\ell(X, Y, h) = \exp(-Yh(X))$

$Y = \pm 1$ $(Y - h(x))^2 = (Y - h(x))^2 = (Y - h(x))^2 / Y^2$

Let $\phi(Yh(X)) = \ell(X, Y, h)$.

Surrogate loss functions

THE UNIVERSITY OF SYDNEY

- How to check if a given surrogate loss function is a classification-calibrated surrogate loss functions?

condition = ϕ is convex,
ϕ is differentiable at 0
$\phi'(0) < 0$

Let $\phi(Yh(X)) = \ell(X, Y, h)$.

sufficient and nessary

Given ϕ is convex, the loss function is classification-calibrated
if and only if ϕ is differentiable at 0, and

$$\phi'(0) < 0.$$

Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. "Convexity, classification, and risk bounds." Journal of the American Statistical Association 101.473 (2006): 138-156.

Zhang, Jingwei, Tongliang Liu, and Dacheng Tao. "On the Rates of Convergence from Surrogate Risk Minimizers to the Bayes Optimal Classifier." arXiv preprint arXiv:1802.03688 (2018).

Objective function

When employing classification-calibrated surrogate loss function, the empirical estimator is unbiased:

$$h_n = \arg\min_h \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, h)$$

$$h_c = \arg\min_h \mathbb{E}[1_{Y \neq \text{sign}(h(X))}]$$

0-1 loss function

$$\mathbb{E}[1_{Y \neq \text{sign}(h_n(X))}] \xrightarrow{n \to \infty} \mathbb{E}[1_{Y \neq \text{sign}(h_c(X))}]$$

Objective function

Recall that a machine learning algorithm is a mapping to find a hypothesis to fit the data

$$\mathcal{A} : S \in (\mathcal{X} \times \mathcal{Y})^n \mapsto h_S \in H.$$

The mapping is an optimisation procedure that picks a hypothesis from the predefined hypothesis class to minimise or maximise the objective.

$$\arg\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, h).$$

picked hypothesis — hypothesis class

THE UNIVERSITY OF
SYDNEY

Convex optimisation

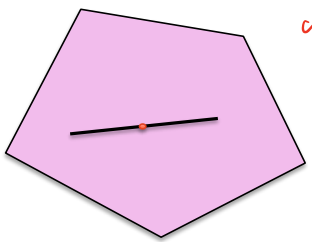
Basics I: Convex set

A set $C \in \mathbb{R}^d$ is convex if $x, y \in C$ and any $\theta \in [0, 1]$

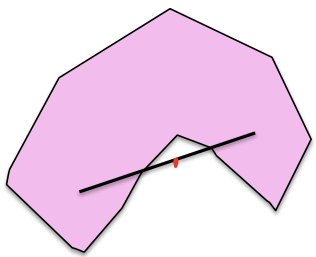
$$\theta x + (1 - \theta) y \in C \, .$$

convex combination of x, y.

Examples: convex and non-convex sets, i.e,



convex



not-convex

Basics II: Convex functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if its domain ($\text{domain } f$) is a convex set and

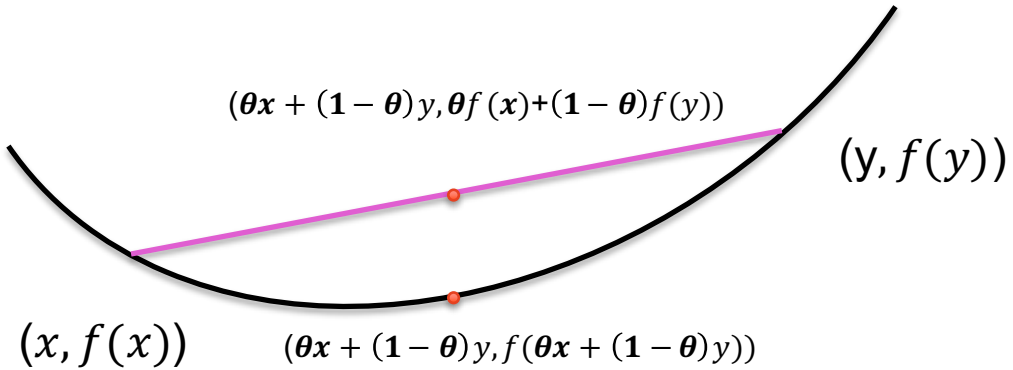
domain : convex set

convex combination of input

convex combination of function

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

for all $x, y \in \text{domain } f$, and $0 \leq \theta \leq 1$.



Basics II: Convex functions

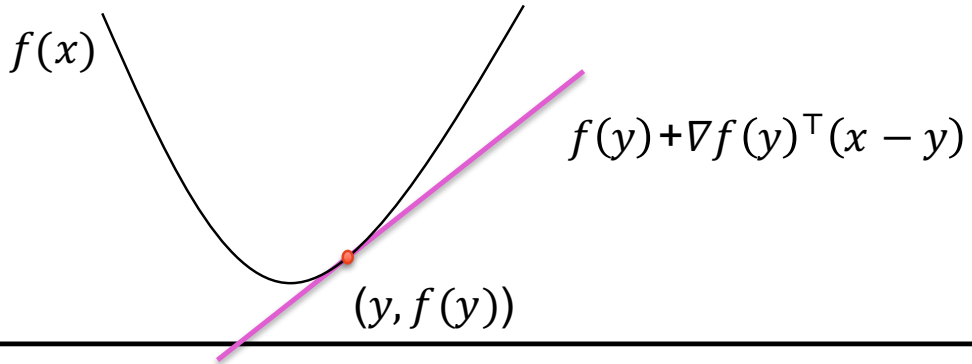
Function f is differentiable if the gradient

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_d} \right), \forall x \in \text{domain } f \subseteq \mathbb{R}^d$$

exists.

Note that differentiable f, with a convex domain, is convex if and only if

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y), \quad \forall x, y \in \text{domain } f$$



Basics II: Convex functions

Function f is twice differentiable if the Hessian matrix

$$H_{ij} = \frac{\partial f(x)}{\partial x_i \partial x_j}, \forall x \in \text{domain } f \subseteq \mathbb{R}^d$$

exists.

We now assume that f is twice differentiable, that is, its Hessian matrix exists at each point in the domain of f. Then f is convex if and only if the Hessian matrix is positive semidefinite for all point in the domain.

Basics II: Convex functions

We now assume that f is twice differentiable, that is, its Hessian matrix exists at each point in the domain of f. Then f is convex if and only if the Hessian matrix is positive semidefinite for all point in the domain.

A square matrix $H \in \mathbb{R}^{d \times d}$ is positive semidefinite if and only if

$$\forall x \in \mathbb{R}^d, x^\top H x \geq 0.$$

Or all its eigenvalues are non-negative.

Basics III: Convex functions

If f_1 and f_2 are convex functions then their pointwise maximum f, defined by

$$f(x) = \max\{f_1(x), f_2(x)\}.$$

is also convex. Note that

$$\text{domain } f = \text{domain } f_1 \cap \text{domain } f_2.$$

Basics III: Convex functions

If f_1 and f_2 are convex functions then their pointwise maximum f, defined by

$$f(x) = \max\{f_1(x), f_2(x)\}$$

is also convex. Note that $\text{domain } f = \text{domain } f_1 \cap \text{domain } f_2$.

Proof: if $0 \leq \theta \leq 1$, $x, y \in$ domain f, then

$$\begin{aligned}
& f(\theta x + (1-\theta)y) \\
&= max\{f_1(\theta x + (1-\theta)y), f_2(\theta x + (1-\theta)y)\} \\
&\leq max\{\theta f_1(x) + (1-\theta)f_1(y)\ , \theta f_2(x) + (1-\theta)f_2(y)\ \} \\
&\leq max\{\theta f_1(x), \theta f_2(x)\ \} + max\{(1-\theta)f_1(y)\ , (1-\theta)f_2(y)\ \} \\
&= \theta f(x) + (1-\theta)f(y).
\end{aligned}$$

Basics III: Convex functions

Non-negative weighted sum:

$$f(x) = \theta_1 f_1(x) + \theta_2 f_2(x)$$

Composition with affine mapping:

$$g(x) = f(Ax + b)$$

Pointwise maximum:

$$f(x) = \max_i \{f_i(x)\}$$

The objective of SVM is convex:

$$f(x) = \frac{1}{2}\|x\|^2 + C\sum_{i=1}^{n} max\{0, 1 - b_i a_i^\top x\}$$

The first term has Hessian matrix are positive, the second term is the sum of convex functions.

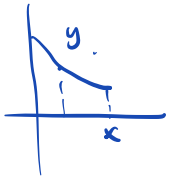
Optimality criterion for f

A point x is optimal for f if and only if it is feasible and

$$\nabla f(x)^\top (y - x) \geq 0$$

for all feasible $y \in \text{domain } f$.

eg ① $\nabla f(x) = 0$.

② 

$\nabla f(x) \leq 0$

$y - x \leq 0$.

Feasible set X: the points in the domain of f satisfying the all the constraints of the objective.

Unconstrained optimisation

- Unconstrained convex optimisation problem

$$\arg\min_{h} f(h).$$

Pick one from the predefined hypothesis class H to minimise the objective, i.e.,

$$\arg\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, h) = \arg\min_{h \in H} f(h)$$

where the loss function ℓ is a convex surrogate for the 0-1 loss function.

Taylor's Theorem

Let $k \geq 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be k times differentiable at the point $a \in \mathbb{R}$. Then there exists a function $h_k : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(a) + f'(a)(x-a) + \ldots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k$$

and $\lim_{x \to a} h_k(x) = 0.$

Taylor's Theorem

- Example 1

$f(0) + f'(0)(x-0) + \frac{f''(0)}{2!}(x-0)^2 + \dots + \frac{f^{(k)}(0)(x-0)^k}{k!} + h_k(x)(x-0)^k$

$= 0 + 0 + \dots \frac{1}{2}x^2 + 0 + \dots$

Let $f(x) = \frac{1}{2}x^2$. The Taylor series at the point 0 is as follows

$$f(x) = f(0) + f'(0)(x-0) + \frac{f''(0)}{2}(x-0)^2$$
$$+ \dots + \frac{f^{(k)}(0)}{K!}(x-0)^k + h_k(x)(x-0)^k$$
$$= \frac{1}{2}x^2$$

Taylor's Theorem

- Example 2

Let $f(x) = \frac{1}{6}x^3$, $K = 2$. The Taylor series at the point 1 is as follows

$$f(x) = f(1) + f'(1)(x-1) + \frac{f''(1)}{2}(x-1)^2$$

$$+ \ldots + \frac{f^{(k)}(1)}{K!}(x-1)^k + h_k(x)(x-1)^k$$

$$= \frac{1}{6} + \frac{1}{2}(x-1) + \frac{1}{2}(x-1)^2 + h_2(x)(x-1)^2$$

$$x \to 1$$

$$= \frac{1}{6} + \frac{1}{2}(x-1) + \frac{1}{2}(x-1)^2 + o((x-1)^2)$$

Small-o Notation

$f(x) = o(g(x)), x \to 0$ means that $\frac{f(x)}{g(x)} \xrightarrow{x \to 0} 0.$

The notation $f(x-1) = o((x-1)^2), x \to 1$ means that when x approaches 1, $f(x-1)$ converges to 0 faster than $(x-1)^2$.

Example 3 implies $\frac{\frac{1}{6}x^3 - \left(\frac{1}{6} + \frac{1}{2}(x-1) + \frac{1}{2}(x-1)^2\right)}{(x-1)^2} \xrightarrow{x \to 1} 0.$

Gradient descent method

Let $$f(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i, Y_i, h)$$

$$h_{k+1} = h_k + \eta d_k \,.$$

f(h_{k+1}) function
re-write at h_k
k=1 Taylor
$f(h_{k+1}) = f(h_k) + f'(h_{k+1})(h_{k+1} - h_k)$. $= \eta d_k$
$+ o(h_{k+1} - h_k)$.

By Taylor's theorem, we have

$$f(h_{k+1}) = f(h_k) + \eta \nabla f(h_k)^\top d_k + o(\eta)\,.$$

if η is small
$o(\eta)$ is much small
ignore

For positive but sufficiently small η,
$f(h_{k+1})$ is smaller than $f(h_k)$,
if the direction d_k is chosen so that

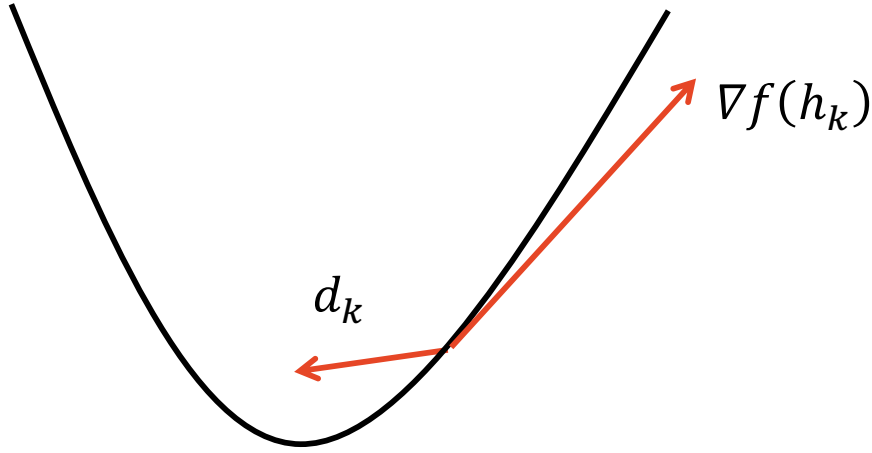
$$\nabla f(h_k)^\top d_k < 0 \quad \text{when} \quad \nabla f(h_k) \neq 0.$$

An iterative updating method

$$f(h_{k+1}) = f(h_k) + \eta \nabla f(h_k)^\top d_k + o(\eta).$$

Two problems:

- How to find d_k?
- How to choose η?



To find d_k

$$f(h_{k+1}) = f(h_k) + \eta \nabla f(h_k)^\top d_k + o(\eta).$$

Set $d_k = -D^k \nabla f(h_k)$, many gradient methods are specified in the form

$$h_{k+1} = h_k - \eta D^k \nabla f(h_k).$$

D^k is a positive definite symmetric matrix,

$$\nabla f(h_k)^\top D^k \nabla f(h_k) > 0.$$

def of positive definite

η is a positive such that

$$f(h_{k+1}) = f(h_k) - \eta \nabla f(h_k)^\top D^k \nabla f(h_k).$$

To find d_k

$$f(h_{k+1}) = f(h_k) + \eta \nabla f(h_k)^\top d_k + o(\eta)\,.$$

Set $d_k = -D^k \nabla f(h_k)$, many gradient methods are specified in the form

$$h_{k+1} = h_k - \eta D^k \nabla f(h_k).$$

- Steepest descent

$$D^k = I$$

hessian positive semi-definite

- Newton's method

$$D^k = [\nabla^2 f(h)]^{-1}$$

Gradient descent method

Basic iteration $$d_k = -\nabla f(h_k)$$
$$h_{k+1} = h_k - \eta \nabla f(h_k).$$

By Taylor's theorem, we have

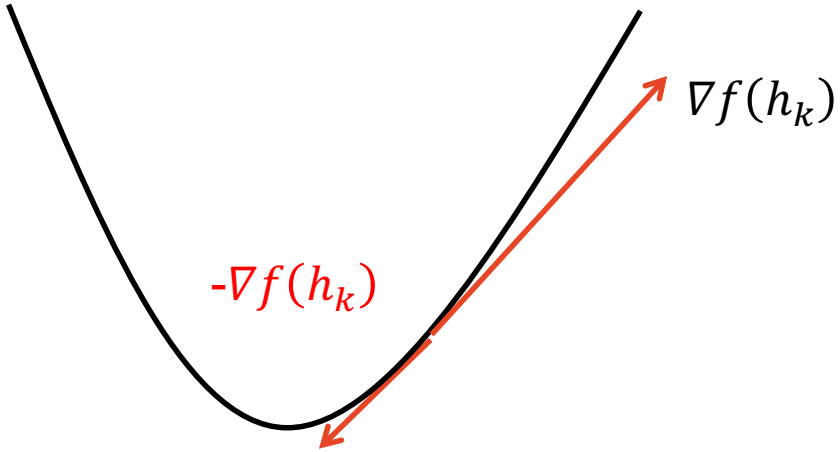
$$f(h_{k+1}) = f(h_k) - \eta \nabla f(h_k)^\top \nabla f(h_k) + o(\eta).$$

For positive but sufficiently small η,
$f(h_{k+1})$ is smaller than $f(h_k)$,
if the direction d_k is chosen so that

$$-\nabla f(h_k)^\top \nabla f(h_k) < 0, \quad \text{when} \quad \nabla f(h_k) \neq 0.$$

Gradient descent method

$$f(h_{k+1}) = f(h_k) - \eta \nabla f(h_k)^\top \nabla f(h_k) + o(\eta).$$



To find η

- Exact line search:

$$\eta = \arg\min_{\eta} f(h_k - \eta \nabla f(h_k))$$

practically expensive.

- Lipschitz smooth constant L exists for the gradient:

$$h_{k+1} = h_k - \frac{1}{L}\nabla f(h_k)$$

$$f(h_{k+1}) \leq f(h_k) - \frac{1}{2L}\|\nabla f(h_k)\|^2.$$

if L is known.

Function f is L-Lipschitz continuous if

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|, \forall x_1, x_2 \in \text{domain } f.$$

eg $f(x) = 2x^2 \quad x \in [0,1]$

$L = 4$

normally, convex function

L: largest value of derivative

Gradient convergence rate

How many iteration steps do we need to achieve the optimal solution ?

$$h_S = \arg\min_{h \in H} f(h) = \arg\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, h)$$

Gradient convergence rate

If the objective function is strongly-convex, and has Lipschitz Gradient, we have a **linear convergence rate**, i.e., defined by

$$f(h_{k+1}) - f(h_S) \leq \left(1 - \frac{\mu}{L}\right)^k (f(h_1) - f(h_S)).$$

A function is μ-strongly convex:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2, \forall x, y,$$

$$\Leftrightarrow$$

$$\mu \boldsymbol{I} \preccurlyeq \nabla^2 f(x), \forall x$$

Gradient descent method

Algorithm	Assumption	Convergence rate
Gradient	Lipshitz Gradient, Convex	$O(1/k)$
Gradient	Lipshitz Gradient, Strongly-Convex	$O(1-\mu/L)^k$
Newton	Lipshitz Gradient, Strongly-convex	$\prod_{i=1}^{k} \rho_k$, $\rho_k \to 0$

sub-linear convergence rate ← largest

linear convergence rate → base is constant

super-linear convergence rate ← smallest; base is smaller

$$f(h_{k+1}) - f(h_S) \leq \left(1 - \frac{\mu}{L}\right)^k (f(h_1) - f(h_S)).$$

Gradient descent method

$$f(h_{k+1}) = f(h_k) + \eta \nabla f(h_k)^\top d_k + o(\eta)$$

Set $d_k = -D^k \nabla f(h_k)$.

Newton's method sets

$$D^k = [\nabla^2 f(h_k)]^{-1}.$$

Obtaining D^k my be difficult. There are many practical variants of Newton's method:

- Modify the Hessian to be positive-definite
- Only compute the Hessian every m iterations
- Only use the diagonals of the Hessian
- Quasi-Newton: Update a approximate of the Hessian (BFGS, L-BFGS)

How to deal with constrained optimisation problem?

Constrained optimisation

- Constrained convex optimisation problem

$$\begin{aligned} &\min_h f_0(h) \\ &\text{s.t. } f_i(h) \leq 0, i = 1, \ldots, k \\ &\quad\; g_i(h) = 0, i = 1, \ldots, l, \end{aligned}$$

where $f_0(h), f_1(h), \ldots, f_k(h)$ are convex functions, $g_i(h)$ are affine functions, i.e., $g_i(h) = a_i^\top h - b_i$.

THE UNIVERSITY OF
SYDNEY

Thank you!

THE UNIVERSITY OF SYDNEY

Appendix Proofs (not examinable)

To find η

- Exact line search:

$$\eta = \arg\min_{\eta} f(h_k - \eta \nabla f(h_k))$$

practically expensive.

- Lipschitz smooth constant L exists for the gradient:

$$h_{k+1} = h_k - \frac{1}{L}\nabla f(h_k)$$

$$f(h_{k+1}) \leq f(h_k) - \frac{1}{2L}\|\nabla f(h_k)\|^2.$$

if L is known.

Proof I

Function f is L-Lipschitz continuous if

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|, \forall x_1, x_2 \in \text{domain } f.$$

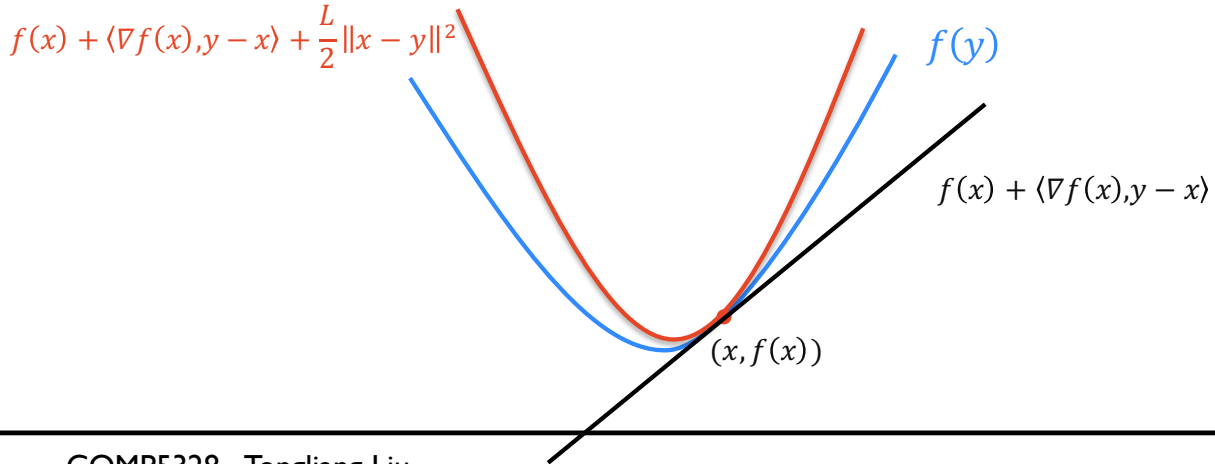
Proof I

Gradient is Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \qquad \forall x, y,$$

$$\Leftrightarrow$$

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2, \forall x, y,$$



Proof I

Gradient is Lipschitz continuous:

At x_k, we have

$$f(y) \leq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2} \|y - x_k\|^2,$$

Set x_{k+1} to minimise the upper bound in terms of y,

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$$

(gradient descent with the step-size of $\frac{1}{L}$)

Plugging into the above inequality:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2$$

(decreasing at least $\frac{1}{2L} \|\nabla f(x_k)\|^2$)

Gradient convergence rate

If the objective function is strongly-convex, and has Lipschitz Gradient, we have a **linear convergence rate**:

$$f(h_{k+1}) - f(h_S) \leq \left(1 - \frac{\mu}{L}\right)^k (f(h_1) - f(h_S)).$$

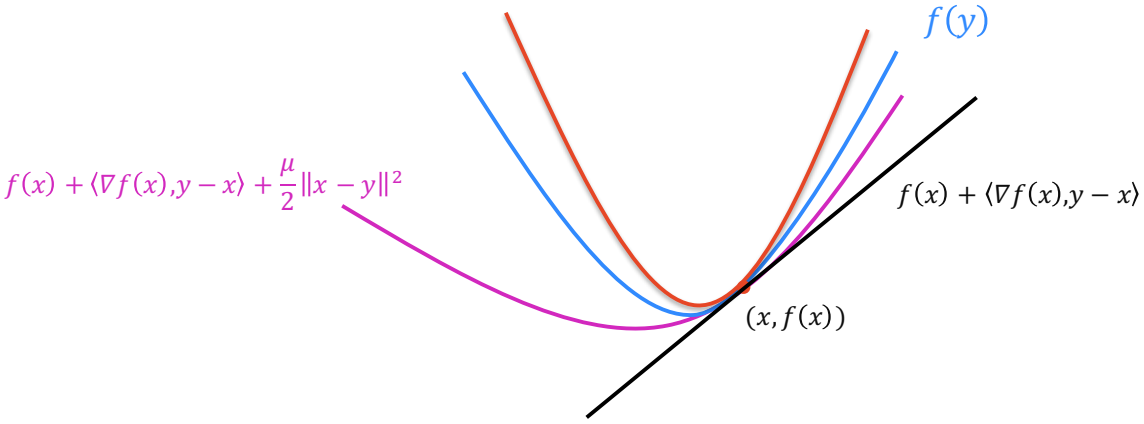
Proof II

A function is μ-strongly convex:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2, \forall x, y,$$

$$\Leftrightarrow$$

$$\mu \boldsymbol{I} \preccurlyeq \nabla^2 f(x), \forall x$$



Proof II

A function is μ-strongly convex:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2, \forall x, y.$$

Let x^* be the minimizer of $f(x)$. We have $\nabla f(x^*) = 0$, then

$$f(\mathrm{y}) \geq f(x^*) + \frac{\mu}{2} \|x^* - y\|^2.$$

Set x_k be the minimizer of $f(\mathrm{y}) - \frac{\mu}{2} \|x^* - y\|^2$, we have

$$x_k = x^* + \frac{\nabla f(x_k)}{\mu}.$$

We have

$$f(x^*) \leq f(x_k) - \frac{1}{2\mu} \|\nabla f(x_k)\|^2.$$

Proof II

Proof:

Because of

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$$

$$f(x^*) \leq f(x_k) - \frac{1}{2\mu}\|\nabla f(x_k)\|^2.$$

After some rearrangement, we have

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)\left(f(x_k) - f(x^*)\right).$$

This gives a linear convergence rate:

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_1) - f(x^*)\right).$$

Proof II

Proof:

Because of $f(x_{k+1}) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x_k)\|^2 - \frac{1}{2L}\|\nabla f(x_k)\|^2 = \left(\frac{1}{2\mu} - \frac{1}{2L}\right)\|\nabla f(x_k)\|^2$

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$$

$$f(x^*) \leq f(x_k) - \frac{1}{2\mu}\|\nabla f(x_k)\|^2.$$

$$\frac{1}{2\mu}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x^*)$$

After some rearrangement, we have Or $\|\nabla f(x_k)\|^2 \leq 2\mu\left(f(x_k) - f(x^*)\right)$

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)\left(f(x_k) - f(x^*)\right).$$

This gives a linear convergence rate:

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_1) - f(x^*)\right).$$

arXiv:1802.03688v2 [stat.ML] 25 Feb 2021

On the Rates of Convergence from Surrogate Risk Minimizers to the Bayes Optimal Classifier

Jingwei Zhang, Tongliang Liu, *Member, IEEE,* and Dacheng Tao, *Fellow, IEEE*

***Abstract*—In classification, the use of 0-1 loss is preferable since the minimizer of 0-1 risk leads to the Bayes optimal classifier. However, due to the non-convexity of 0-1 loss, this optimization problem is NP-hard. Therefore, many convex surrogate loss functions have been adopted. Previous works have shown that if a Bayes-risk consistent loss function is used as surrogate, the minimizer of the emprical surrogate risk can achieve the Bayes optimal classifier as sample size tends to infinity. Nevertheless, the comparison of convergence rates of minimizers of different empirical surrogate risks to the Bayes optimal classifier has rarely been studied. Which characterization of the surrogate loss determines its convergence rate to the Bayes optimal classifier? Can we modify the loss function to achieve a faster convergence rate? In this paper, we study the convergence rates of empirical surrogate minimizers to the Bayes optimal classifier. Specifically, we introduce the notions of *consistency intensity* and *conductivity* to characterize a surrogate loss function and exploit this notion to obtain the rate of convergence from an empirical surrogate risk minimizer to the Bayes optimal classifier, enabling fair comparisons of the excess risks of different surrogate risk minimizers. The main result of the paper has practical implications including (1) showing that hinge loss (SVM) is superior to logistic loss (Logistic regression) and exponential loss (Adaboost) in the sense that its empirical minimizer converges faster to the Bayes optimal classifier and (2) guiding the design of new loss functions to speedup the convergence rate to the Bayes optimal classifier with a data-dependent loss correction method inspired by our theorems.**

***Index Terms*—Consistency, Bayesian Optimal Classifier, Generalization, Surrogate Loss**

I. Introduction

Classification is a fundamental machine learning task used in a wide range of applications. When designing classification algorithms, the 0-1 loss function is preferred, as it helps produce the Bayes optimal classifier, which has the minimum probability of classification error. However, the 0-1 loss is difficult to optimize because it is neither convex nor smooth ([Ben-David et al., 2003], [Feldman et al., 2012]). Many different computationally-friendly surrogate loss functions have therefore been proposed as approximations for the 0-1 loss function.

However, natural questions arise of whether they are good approximations, and then what the differences are between the surrogate loss functions and the 0-1 loss. To address the first question, the Bayes-risk consistency concept has been introduced ([Lugosi and Vayatis, 2004], [Bartlett et al., 2006]). A surrogate loss function is said to be Bayes-risk consistent if its corresponding empirical minimizer converges to the Bayes optimal classifier when the predefined hypothesis class is universal. That means, with a sufficiently large sample, the minimizers of those surrogate risks are identical to the minimizer of the 0-1 risk in the sense that they achieve the same minimum probability of classification error. Existing results (e.g., [Zhang, 2004], [Bartlett et al., 2006], [Agarwal and Agarwal, 2015], [Neykov et al., 2016]) show Bayes-risk consistency under different conditions and, reassuringly, most of the frequently used surrogate loss functions are Bayes-risk consistent.

Although Bayes-risk consistency describes the interchangeable relationship between surrogate loss functions and the 0-1 loss function, it is an asymptotic concept. The non-asymptotic link between a specific surrogate loss function and the 0-1 loss function has remained elusive. In this paper, we study the rates of convergence from surrogate risk minimizers to the Bayes optimal classifier. We derive upper bounds for the difference between the probabilities of classification error w.r.t. the surrogate risk minimizer and the Bayes optimal classifier. Specifically, we introduce the notions of *consistency intensity* I and *conductivity* S, which are uniquely determined by the surrogate loss functions. We show that for any given surrogate loss function ϕ, if the convergence rate of the excess surrogate risk $R_\phi(f_n) - R_\phi^*$ is of order $\mathcal{O}(\frac{1}{n^p})$, where f_n is the empirical surrogate risk minimizer, R_ϕ is the expected surrogate risk, and R_ϕ^* is the minimal surrogate risk achievable, the corresponding convergence rate of the expected risk $R(f_n) - R^*$ is of order $\mathcal{O}(\frac{S}{n^{pI}})$, where R and R^* are the expected 0-1 risk and the minimal 0-1 risk achievable, respectively. The result is able to (1) describe the non-asymptotic differences between different surrogate loss functions and (2) fairly compare the rates of convergence from different empirical surrogate risk minimizers to the Bayes optimal classifier.

We apply our theorems to popular surrogate loss functions such as the hinge loss function in support vector machine, exponential loss function in AdaBoost, and logistic loss function in logistic regression ([Collins et al., 2002], [Vapnik, 2013]), as illustrated in Figure 1. We conclude that SVM converges faster to the Bayes optimal classifier than AdaBoost and logistic regression, while AdaBoost and logistic regression have the same convergence rate. Furthermore, we provide a general rule for fairly comparing the convergence rates of different classification algorithms.

We show that for a data-independent surrogate loss, both the consistency intensity I and conductivity S are constants, and for different surrogate loss functions, I and S vary in $(0, 1]$

J. Zhang, T. Liu, and D. Tao are with the UBTECH Sydney Artificial Intelligence Centre, and the School of Computer Science, in the Faculty of Engineering and Information Technologies, The University of Sydney, J12 Cleveland St, Darlington, NSW 2008, Australia. E-mail: zjin8228@uni.sydney.edu.au, tongliang.liu@sydney.edu.au, dacheng.tao@sydney.edu.au.

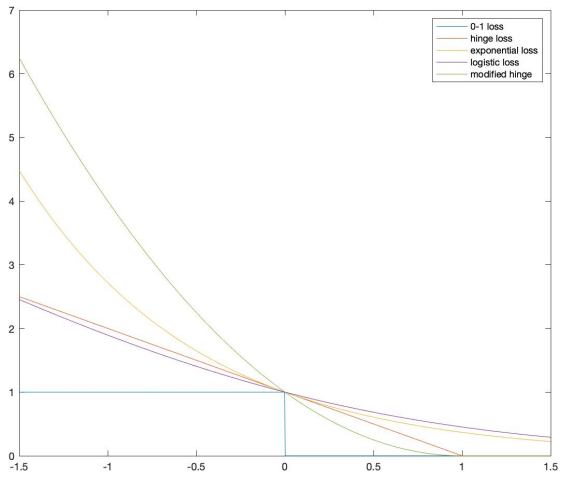


Fig. 1: Different convex surrogate loss functions studied in this paper. The modified hinge loss is studied in the second to last section where we take the modification parameter $\Delta = 1$ in the figure.

and $(0, +\infty)$, respectively. Since different minimizers converge to the Bayes optimal classifier at rate $\mathcal{O}(\frac{S}{n^{pI}})$, they do not contribute to accelerating the convergence rate. However, if we modify the surrogate loss function according to the sample size n, S can vary w.r.t. n. This finding enables us to devise a data-dependent loss modification method which can accelerate the convergence of a surrogate risk minimizer to the Bayes optimal classifier.

Advantages and Disadvantages. Despite extensive existing works that study the convergence of surrogate risk minimizer to the Bayes optimal classifier, they are focusing on the convergence of some specific loss functions satisfying certain regularity conditions, such as convex and differentiable. Besides, these works do not show which characterization of the given surrogate loss function determines the convergence rate. Furthermore, existing work does not show how to compare the convergence rate to the Bayes optimal classifier when given several different surrogate loss functions. Another advantage of our result is that we can modify the loss function to achieve a faster convergence rate to the Bayes optimal classifier, and the modification of loss is data-dependent. Hence, our theorem provides a novel way to design better surrogate loss functions such that the minimizer converges faster to the Bayes optimal classifier. In spite of such advantages in our work, there are also some critical problem that remains unsolved: what is the fastest convergence rate that is achievable by any surrogate loss function? To answer this problem, we need to derive a lower bound of the convergence rate under certain regularity conditions on the loss function, which leaves as future work.

Organization. The remainder of the paper is organized as follows. In section II, we first introduce basic mathematical notations. Then we introduce the notions of Bayes optimality and Bayes-risk consistency, along with several lemmas and related works necessary to the proofs of our main theorem. We present our main theorem in section III, namely the rate of convergence from the expected risk to the Bayes risk for empirical surrogate risk minimization. In section IV, we show several applications of our theorem. In particular, section IV-A details several specific examples of computing and comparing the rates of convergence to the Bayes risk and provides a general rule to compare the rates of convergence to the Bayes risk for any two algorithms with Bayes-risk consistent loss functions. Then, in section IV-B, we apply our theorems to modifying the hinge loss function to accelerate its convergence to the Bayes risk. In section V, we conclude the paper and briefly discuss future works.

II. Preliminaries

We present basic notations in Section II-A and briefly introduce the concept of Bayes-risk consistency and necessary lemmas in Section II-B. In Section II-C, we discuss related works about the statistical properties of Bayes-risk consistent surrogate loss functions, which play critical roles in the proof of our theorems.

A. Notation

In this paper, we focus on binary classification, where the feature space $\mathcal{X}$ is a subset of a Hilbert Space $\mathcal{H}$ and the label space is denoted by $\mathcal{Y} = \{-1, +1\}$. We assume that a pair of random variables (X, Y) is generated according to an unknown distribution D, where $P(X, Y)$ is the corresponding joint probability. Binary classification aims to find a map $f : \mathcal{X} \to \mathbb{R}$ within some particular predefined hypothesis class $\mathcal{F}$ such that the sign of $f(X)$ can be used as a prediction for $Y \in \mathcal{Y}$.

To evaluate the goodness of f, some performance measures are required. Intuitively, the 0-1 risk is employed, defined as:

$$R(f) = \mathbb{E}[\mathbb{1}[sgn(f(X)) \neq Y]] = P(sgn(f(X)) \neq Y) \quad (1)$$

where $\mathbb{1}[\cdot]$ denotes the indicator function. From the definition, we can see that minimizing the 0-1 risk is equivalent to minimizing the probability of classification error. We hope to find the function such that the probability of classification error is minimized, which is called the *Bayes optimal classifier*, defined as follows:

$$f^* = \arg\inf_f R(f) \quad (2)$$

where the infimum is over all measurable functions. The corresponding expected risk is called the Bayes risk:

$$R^* = R(f^*) \ . \quad (3)$$

In this paper, we assume that the predefined hypothesis class is universal, which means the Bayes optimal classifier is always in $\mathcal{F}$.

Since the joint distribution D is unknown, we cannot calculate $R(f)$ directly. Given a training sample $\mathbf{S}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$, the following empirical risk is widely exploited to approximate the expected risk R:

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[sgn(f(x_i)) \neq y_i] \ . \quad (4)$$

Directly minimizing the above empirical risk is NP-hard due to the non-convexity of the 0-1 loss function, which forces us to adopt convex surrogate loss functions. Similarly, for any non-negative surrogate loss function $\phi : \tilde{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}^+$, where $\tilde{\mathcal{Y}}$ is the space of the output of the classifier, we can define the ϕ-risk, optimal ϕ-risk, empirical ϕ-risk, and the empirical surrogate risk minimizer as:

$$R_\phi(f) = \mathbb{E}[\phi(f(X), Y))] \ , \tag{5}$$

$$R_\phi^* = \inf_f R_\phi(f) \ , \tag{6}$$

$$\hat{R}_\phi(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(f(x_i), y_i) \ , \tag{7}$$

$$f_n = \arg\inf_{f\in\mathcal{F}} \hat{R}_\phi(f) \ . \tag{8}$$

Moreover, we define the *excess risk* and the *excess ϕ-risk*, respectively, as follows:

$$R(f_n) - R^*, \tag{9}$$

$$R_\phi(f_n) - R_\phi^* \ . \tag{10}$$

For classification tasks, the loss function is often margin-based and we can rewrite $\phi(f(x_i), y_i)$ as $\phi(y_i f(x_i))$, where the quantity $yf(x)$ is known as the margin, which can be interpreted as the confidence in prediction ([Mohri et al., 2012]). In this paper, we study margin-based loss functions.

B. Optimality and Bayes-risk Consistency

We first introduce the concept of *Bayes optimal*. Let define

$$\eta(X) = P(Y = 1|X) \ . \tag{11}$$

Lemma 1. *([Bousquet et al., 2004]) Assume the random pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ follows a given distribution D. Then, any classifier $f : \mathcal{X} \to \mathbb{R}$, which satisfies $sgn(f(X)) = sgn(\eta(X) - 1/2)$, is Bayes optimal under D.*

Before we introduce the notion of Bayes-risk consistency for surrogate loss functions, we need to present several basic definitions ([Bousquet et al., 2004]). First, we define the *conditional ϕ-risk* as:

$$\mathbb{E}[\phi(Yf(X))|X = x] = \eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x)) \ . \tag{12}$$

We then introduce the *generic conditional ϕ-risk* by letting $z = f(x)$ and $\eta = \eta(x)$:

$$C(\eta, z) = \eta\phi(z) + (1-\eta)\phi(-z) \ . \tag{13}$$

Note that the function $C(\eta, z)$ is a convex combination of $\phi(z)$ and $\phi(-z)$. It immediately follows the definition of *optimal generic conditional ϕ-risk* :

$$C^*(\eta) = \inf_{z\in\mathbb{R}} \eta\phi(z) + (1-\eta)\phi(-z) \ . \tag{14}$$

We also define $C^{*-}(\eta)$ as:

$$C^{*-}(\eta) = \inf_{z:z(\eta-1/2)\le 0} C_\eta(z) \ . \tag{15}$$

This definition follows the *optimal generic conditional ϕ-risk* , but with the constraint that the sign of the output z differs from $sgn(\eta - 1/2)$. Under these settings, we define the Bayes-risk consistency.

Definition 1. *([Lugosi and Vayatis, 2004]) A surrogate loss function ϕ is Bayes-risk consistent if the minimizer of the conditional ϕ-risk, $f^* = \arg\inf_f \mathbb{E}[\phi(Yf(X))|X = x]$, has the same sign as the Bayes optimal classifier for any $x \in \mathcal{X}$. Or simply, $sgn(f^*(x)) = sgn(\eta(x) - 1/2)$.*

We here present a necessary and sufficient condition for the Bayes-risk consistency of surrogate loss functions.

Lemma 2. *([Bartlett et al., 2006]) A surrogate loss is Bayes-risk consistent if and only if $C^*(\eta) < C^{*-}(\eta)$, for any $\eta \neq 1/2$.*

Lemma 2 has an intuitive explanation: any Bayes-risk consistent loss requires the constraint that it always leads to strictly larger conditional ϕ-risks for any X when the signs of the output $f(X)$ differs from that of Bayes optimal classifier.

C. Asymptotic Consistency in Surrogate Risk Optimization

We briefly introduce some related works on Bayes-risk consistency and its statistical properties in classification. [Lugosi and Vayatis, 2004] proved that the Bayes-risk consistency is satisfied for empirical surrogate loss minimization under the condition that the surrogate loss ϕ is strictly convex, differentiable, and monotonic with $\phi(0) = 1$. [Bartlett et al., 2006] then offered a more general result, showing that the Bayes-risk consistency is possible if and only if the loss function is Bayes-risk consistent[1]. Other results on Bayes-risk consistency under different assumptions have been presented by [Zhang, 2004], [Steinwart, 2005], [Neykov et al., 2016].

Below, we give the main results proved in [Bartlett et al., 2006], which shows that minimizing over any surrogate risk with Bayes-risk consistent loss function is asymptotically equivalent to minimizing the 0-1 risk and thus leads to the Bayes optimal classifier.

Theorem 1. *([Bartlett et al., 2006]) For any convex loss function ϕ, it is Bayes-risk consistent if and only if it is differentiable at 0 and $\phi'(0) < 0$. Then for such a convex and Bayes-risk consistent loss function ϕ, any measurable function $f : \mathcal{X} \to \mathbb{R}$, and any distribution P over $\mathcal{X} \times \mathcal{Y}$, the following inequality holds,*

$$\psi(R(f) - R^*) \le R_\phi(f) - R_\phi^* \tag{16}$$

where $\psi(\theta) = \phi(0) - C^*(\frac{1+\theta}{2})$ is nonnegative, convex and invertible on $[0, 1]$ and has only one zero at $\theta = 0$.

The above theorem gives an upper bound on the excess risk in terms of the excess ϕ-risk and shows that minimizing over any convex Bayes-risk consistent surrogate loss is asymptotically equivalent to minimizing over 0-1 loss because the function ψ is invertible and only have a single zero at $\theta = 0$. We provide deeper insights of this theorem in the next section.

[1]In their paper, the term “classification-calibrated" is used instead of “Bayes-risk consistent".

III. The Rates of Convergence from Empirical Surrogate Risk Minimizers to the Bayes Optimal Classifier

We know that optimizing over any empirical surrogate risk with a Bayes-risk consistent loss function will lead to the Bayes optimal classifier when the training sample size n is large enough. However, a natural question arises when we optimize over different empirical surrogate risks with Bayes-risk consistent loss functions: "What are the difference between them? Do the minimizers have the same rate of convergence to the Bayes optimal classifier?" Those problems are essential because when we choose classification algorithms for a real-world problem, we may expect a fast convergence to the Bayes optimal classifier which also implies a small sample complexity.

To answer the above questions, we need to find a proper metric to measure the distance between f_n and f^*. Since we are mainly care about the probability of classification error of the proposed learning algorithm, it is reasonable to measure the rate of convergence from the expected risk $R(f_n)$ to the Bayes risk R^* instead of measuring the rate of convergence from f_n to f^* directly.

Before showing the results of convergence rates, we present the intuition of our work.

A. Intuition of the Proposed Method

For any empirical surrogate risk minimizer f_n, we can rewrite the inequality in Theorem 1 as follows:

$$\begin{aligned}&\psi(R(f_n) - R^*) \leq R_\phi(f_n) - R_\phi^* \\ &= \left(R_\phi(f_n) - \inf_{f\in\mathcal{F}} R_\phi(f)\right) + \left(\inf_{f\in\mathcal{F}} R_\phi(f) - R_\phi^*\right) \ .\end{aligned} \tag{17}$$

To achieve our goal of measuring the rate of convergence from the empirical surrogate risk minimizer to the Bayes optimal classifier, we need to bound the term $R(f_n) - R^*$. As $\psi(\theta)$ is invertible, we can bound the term on the right hand side of the inequality. We call the first term in the right hand side the *estimation error*, which depends on the learning algorithm and the training data. The second term is called the *approximation error*, depending on the choice of the hypothesis class $\mathcal{F}$. Often, the hypothesis class is predefined and universal. Thus, in this paper, we just assume that the Bayes optimal classifier is right in the hypothesis class $\mathcal{F}$. In other words, we have $\inf_{f\in\mathcal{F}} R_\phi(f) = R_\phi^*$ and (17) becomes

$$\psi(R(f_n) - R^*) \leq R_\phi(f_n) - \inf_{f\in\mathcal{F}} R_\phi(f) \ . \tag{18}$$

The right side of the above inequality can be further upper bounded by

$$R_\phi(f_n) - \inf_{f\in\mathcal{F}} R_\phi(f) \leq 2\sup_{f\in\mathcal{F}} |\hat{R}_\phi(f) - R_\phi(f)| \ , \tag{19}$$

where the defect on the right hand side is called the *generalization error*. Using the concentration of measure ([Boucheron et al., 2013]), and the uniform convergence argument, e.g., VC-dimension ([Vapnik, 2013]), covering number ([Zhang, 2002]), and Rademacher complexity ([Bartlett and Mendelson, 2002]), the generalization error can be non-asymptotically upper bounded with a high probability. Often, the upper bounds can reach the order $\mathcal{O}(\frac{1}{\sqrt{n}})$ ([Mohri et al., 2012]). We also notice that the excess ϕ-risk can achieve convergence rates faster than $\mathcal{O}(\frac{1}{\sqrt{n}})$, such as exploiting local Rademacher complexities, low noise models, and strong convexity [Tsybakov, 2004], [Bartlett et al., 2005], [Koltchinskii et al., 2006], [Sridharan et al., 2009], [Liu et al., 2017]. Here we mainly consider the ordinary case where the convergence rate of the excess ϕ-risk is of order $\mathcal{O}(\frac{1}{\sqrt{n}})$, but our results can directly generalize to other cases.

Theorem 1 also implies that if ϕ is convex and Bayes-risk consistent, then for any sequence of measurable functions $f_i : \mathcal{X} \to \mathbb{R}$ and any distribution P over $\mathcal{X} \times \mathcal{Y}$,

$$R_\phi(f_i) \to R_\phi^* \quad \Rightarrow \quad R(f_i) \to R^* \ . \tag{20}$$

This presents the dynamics of the Bayes risk consistency for any convex Bayes-risk consistent loss in an asymptotic way.

Observe that in equation (16), the asymptotic consistency property is mainly due to the uniqueness of the zero of the function $\psi(\theta)$, where the only zero of function $\psi(\theta)$ is at $\theta = 0$. Thus, when $R(f_i) - R^* \to 0$, we have $\psi(R(f_i) - R^*) \to 0$, which leads to $R(f_i) \to R^*$. However, if we want to derive the rate of convergence from the expected risk $R(f_n)$ to the Bayes risk R^*, we may need more detailed or higher order properties of the function $\psi(\theta)$ in the infinitesimal right neighborhood of $\theta = 0$, rather than just the value of $\psi(\theta)$ at $\theta = 0$.

In the next subsection, we will exploit the upper bound of $R_\phi(f_n) - R_\phi^*$ and a higher order property of the function $\psi(\theta)$ to derive upper bounds for $R(f_n) - R^*$.

B. Consistency Intensity for Bayes-risk Consistent Loss Functions

Knowing the convergence rate of the excess ϕ-risk and their relation $\psi(R(f_n) - R^*) \leq R_\phi(f_n) - R_\phi^*$, it's straightforward to consider taking an inverse of the function ψ on the both sides, yielding an upper bound, $\psi^{-1}(R_\phi(f) - R_\phi^*)$. However, in reality, for most of Bayes-risk consistent loss functions, the corresponding $\psi(\theta)$ is sometimes intractable to take an inverse analytically. Furthermore, the term $\psi^{-1}(\mathcal{O}(\frac{1}{\sqrt{n}}))$ may not reflect the order of n explicitly. Thus, we must figure out the factor that determines the convergence rate of the excess risk $R(f_n) - R^*$—that is some high order property of function $\psi(\theta)$ within the infinitesimal right neighborhood of $\theta = 0$. Before we move on to our main theorems, let's introduce some basic lemmas and propositions first.

Proposition 1. *For any two functions $f(\theta)$ and $g(\theta)$, which are differentiable at $\theta = 0$ and satisfy $f(0) = g(0) = 0$, then, the following conditions are equivalent:*

- $\lim_{\theta\to 0} \frac{f(\theta)}{g(\theta)} = A$;
- $f(\theta) = Ag(\theta) + o(g(\theta))$;
- $f(\theta) = \mathcal{O}(g(\theta))$ *when* $A \neq 0$.

Proposition 1 can be proved directly by following the definition of the o and $\mathcal{O}$ notation. We now introduce the notions of consistency intensity and conductivity for convex Bayes-risk consistent loss functions.

Lemma 3. *For any given convex Bayes-risk consistent loss function ϕ, let $\psi(\theta) = \phi(0) - C^*(\frac{1+\theta}{2})$. There exists two unique constant $\alpha \in \mathbb{R}^+$ and $M \in \mathbb{R}^+$ such that*

$$\lim_{\theta \to 0+} \frac{\psi(\theta)}{M\theta^\alpha} = 1 \ . \tag{21}$$

We call $I = \frac{1}{\alpha}$ the consistency intensity of this Bayes-risk consistent loss function and $S = M^{-\frac{1}{\alpha}}$ the conductivity of the intensity.

Proof. From Theorem 1, we know that $\psi(\theta)$ is a convex function. It's also known that any convex function on a convex open subset of $\mathbb{R}^n$ is semi-differentiable. Thus, we can denote the right derivative of $\psi(\theta)$ at $\theta = 0$ by $\partial_+\psi(0)$. Using Maclaurin expansion, we have,

$$\psi(\theta) = \psi(0) + \partial_+\psi(0) * \theta + o(\theta) = \partial_+\psi(0) * \theta + o(\theta) \ . \tag{22}$$

Followed by Proposition 1, if $\partial_+\psi(0) \neq 0$, we have,

$$M = \partial_+\psi(0) \quad \text{and} \quad \alpha = 1 \ . \tag{23}$$

If $\partial_+\psi(0) = 0$, then $\psi(\theta) = o(\theta)$, which means that $\psi(\theta)$ is the infinitesimal of higher order than θ as $\theta \to 0+$. Then, by definition, for any given ϕ, we can compute ψ. Because $\psi(\theta)$ is the higher order infinitesimal of θ as $\theta \to 0+$, there exist unique $\alpha > 1$ and $M \in \mathbb{R}^+$ such that,

$$\lim_{\theta \to 0+} \frac{\psi(\theta)}{M\theta^\alpha} = 1 \tag{24}$$

which completes the proof. □

We then introduce Lemmas 4 and 5, which are essential to the proof of our main theorems.

Lemma 4. *Given any convex Bayes-risk consistent loss function ϕ, let the inverse of its corresponding ψ-transform be denoted by ψ^{-1}. We have*

$$\lim_{\mu \to 0+} \frac{\psi^{-1}(\mu)}{S\mu^I} = 1 \ . \tag{25}$$

Proof. Let $\psi^{-1}(\mu) = \theta$, then $\mu = \psi(\theta)$. From Theorem 1, we know that ψ is monotonic increasing within $[0, 1]$, and $\psi(0) = 0$. Thus,

$$\mu \to 0+ \quad \text{implies} \quad \theta \to 0+ \ .$$

We have,

$$\lim_{\mu \to 0+} \frac{\psi^{-1}(\mu)}{S\mu^I} = \lim_{\theta \to 0+} \frac{\theta}{S(\psi(\theta))^I} \ . \tag{26}$$

By substituting the definitions of S and I, the right hand side of (26) becomes,

$$\lim_{\theta \to 0+} \frac{M^{\frac{1}{\alpha}}\theta}{(\psi(\theta))^{\frac{1}{\alpha}}} = \left[\lim_{\theta \to 0+} \frac{M\theta^\alpha}{\psi(\theta)}\right]^{\frac{1}{\alpha}} = 1 \ . \tag{27}$$

The last equality follows (21) in Lemma 3, which completes the proof. □

Lemma 4 shows that the equivalent infinitesimal of ψ^{-1} near 0 is $S\mu^I$. Then, we introduce an important property for the ψ-transform.

Lemma 5. *Given any convex Bayes-risk consistent loss function ϕ, its corresponding function ψ^{-1} is interchangeable with $\mathcal{O}(\frac{1}{n^p})$ for any $p > 0$. That is,*

$$\psi^{-1}\left(\mathcal{O}(\frac{1}{n^p})\right) = \mathcal{O}\left(\psi^{-1}(\frac{1}{n^p})\right) \ . \tag{28}$$

Proof. From Proposition 1, we have that there exists $0 < A, B < +\infty$ such that,

$$\mathcal{O}\left(\frac{1}{n^p}\right) = A\frac{1}{n^p} + o\left(\frac{1}{n^p}\right) \tag{29}$$

and

$$\mathcal{O}\left(\psi^{-1}(\frac{1}{n^p})\right) = B\psi^{-1}\left(\frac{1}{n^p}\right) + o\left(\psi^{-1}(\frac{1}{n^p})\right) \ . \tag{30}$$

Substituting (29) and (30) into (28), it's equivalent to proving that for any $0 < A < +\infty$, there exists $0 < B < +\infty$, such that,

$$\psi^{-1}\left(A\frac{1}{n^p} + o(\frac{1}{n^p})\right) = B\psi^{-1}\left(\frac{1}{n^p}\right) + o\left(\psi^{-1}(\frac{1}{n^p})\right) \ . \tag{31}$$

To prove (31), by Proposition 1, we only need to prove that, for any $0 < A < +\infty$, there exists $0 < B < +\infty$, such that,

$$\lim_{n \to +\infty} \frac{\psi^{-1}(A\frac{1}{n^p} + o(\frac{1}{n^p}))}{\psi^{-1}(\frac{1}{n^p})} = B \ . \tag{32}$$

Followed by Lemma 4 and proposition 1, we have

$$\psi^{-1}(\mu) = S\mu^I + o(\mu^I) \ . \tag{33}$$

Substituting (33) into (32), we have,

$$\begin{aligned} &\lim_{n \to +\infty} \frac{\psi^{-1}(A\frac{1}{n^p} + o(\frac{1}{n^p}))}{\psi^{-1}(\frac{1}{n^p})} \\ &= \lim_{n \to +\infty} \frac{S(A\frac{1}{n^p} + o(\frac{1}{n^p}))^I + o((A\frac{1}{n^p} + o(\frac{1}{n^p}))^I)}{S(\frac{1}{n^{p*I}}) + o((\frac{1}{n^{p*I}}))} \\ &= \lim_{n \to +\infty} \frac{S(A^I\frac{1}{n^{p*I}} + o(\frac{1}{n^{p*I}}))}{S(\frac{1}{n^{p*I}})} \\ &= A^I \ . \end{aligned} \tag{34}$$

This means that for any $0 < A < +\infty$, there exists $B = A^I \in (0, +\infty)$ such that (32) holds true, which completes the proof. □

Following the definitions of S and I in Lemma 3, we have our first main theorem.

Theorem 2. *Suppose the excess ϕ-risk satisfies $R_\phi(f_n) - R_\phi^* \leq \mathcal{O}(\frac{1}{n^p})$ with a high probability. Then, with the same high probability, we have*

$$R(f_n) - R^* \leq \mathcal{O}\left(\frac{S}{n^{pI}}\right) \ . \tag{35}$$

Proof. From Theorem 1, we have that

$$R(f_n) - R^* \leq \psi^{-1}(R_\phi(f_n) - R_\phi^*) \tag{36}$$

where f_n is the minimizer of the empirical surrogate risk $\hat{R}_\phi$ with sample size n. Under our assumption that the Bayes

optimal classifier is within the hypothesis class $\mathcal{F}$, then, with a high probability, we have ([Bousquet et al., 2004]),

$$R_\phi(f_n) - R_\phi^* \leq \mathcal{O}\left(\frac{1}{n^p}\right) \ . \tag{37}$$

Note that p is often equal to $1/2$ for the worst cases. With (37) and Lemma 5, we have,

$$\begin{aligned} &\psi^{-1}(R_\phi(f_n) - R_\phi^*) \\ &\leq \psi^{-1}\left(\mathcal{O}(\frac{1}{n^p})\right) = \mathcal{O}\left(\psi^{-1}(\frac{1}{n^p})\right) \ . \end{aligned} \tag{38}$$

Substituting (33) into (38), we have,

$$\begin{aligned} &\mathcal{O}\left(\psi^{-1}(\frac{1}{n^p})\right) = \mathcal{O}\left(S(\frac{1}{n^p})^I + o((\frac{1}{n^p})^I)\right) \\ &= \mathcal{O}\left(\frac{S}{n^{pI}}\right) \ , \end{aligned} \tag{39}$$

which completes the proof. □

From Theorem 2, we show that the consistency intensity I and conductivity S have direct influence on the convergence rate, which is of order $\mathcal{O}(\frac{S}{n^{pI}})$. When $0 < I < 1$, the convergence rate from $R(f_n)$ to R^* will be slower than the convergence rate from $R_\phi(f_n)$ to R_ϕ^*; if $I = 0$, the algorithm will never reach the Bayes optimal classifier; if $I = 1$, the convergence rates will be the same. In the next theorem, we will show that for any convex Bayes-risk consistent loss function, the range of I is $(0, 1]$.

Theorem 3. *For any convex Bayes-risk consistent loss function ϕ, it always holds true that $0 < I \leq 1$.*

Proof. We have that $0 < \alpha < +\infty$, so $I > 0$ holds trivially. To prove $I = \frac{1}{\alpha} \leq 1$, it is equivalent to proving that there exists $0 \leq C < +\infty$ such that,

$$\psi(\theta) = C\theta + o(\theta) \ , \tag{40}$$

because $I < 1$ holds true if and only if $C = 0$; and $I = 1$ holds true if and only if $0 < C < +\infty$. From proposition 1, the equation (40) implies,

$$\lim_{\theta \to 0+} \frac{\psi(\theta)}{\theta} = C \ . \tag{41}$$

Since any convex function on a convex open subset in $\mathbb{R}^n$ is semi-differentiable, $\psi(\theta)$ is at least right differentiable at $\theta = 0$. We therefore have,

$$\lim_{\theta \to 0+} \frac{\psi(\theta)}{\theta} = \lim_{\theta \to 0+} \frac{\psi(\theta) - \psi(0)}{\theta - 0} = \partial_+\psi(0) = C \tag{42}$$

where $\partial_+\psi(0)$ denotes the right derivative of $\psi(\theta)$ at $\theta = 0$. The proof ends. □

As will be shown in the later examples, for Adaboost (exponential loss) and Logistic regression (logistic loss), we have $C = 0$, then I will be strict less than one. For SVM (Hinge Loss), we have $C > 0$, then by definition we have $I = 1$. Theorem 3 shows that for data-independent surrogate loss functions, because $I \leq 1$, the convergence rate from $R(f_n)$ to R^* will not be faster than the convergence rate from $R_\phi(f_n)$ to R_ϕ^*. As $I > 0$, it also means that optimizing over any convex Bayes-risk consistent surrogate loss will finally make the excess risk $R(f_n) - R^*$ converge to 0 and thus the output is the Bayes optimal classifier as sample size n tends to infinity. This result matches our common sense: while we benefit from the computational efficiency of convex surrogate loss functions, we also suffer from a slower rate of convergence to the Bayes optimal classifier.

IV. Applications

In this section, we present several applications of our results. In Section IV-A, we use the notion of consistency intensity to measure the rates of convergence from the empirical surrogate risk minimizers to the Bayes optimal classifier for different classification algorithms, such as support vector machine, boosting, and logistic regression. We also derive a general discriminant rule for comparing the convergence rates for different leaning algorithms. In Section IV-B, we show that the notions of consistency intensity and conductivity can help to modify surrogate loss functions so as to achieve a faster convergence rate from $R(f_n)$ to R^*.

A. Consistency Measurement

In this subsection, we first apply our results to some popular classification algorithms.

Example 1 (Hinge loss in SVM). Here we have $\phi(z) = max\{0, 1 - z\}$, which is convex and $\phi'(0) = -1 < 0$. Note that we have defined $C^*(\eta) = \inf_{z \in \mathbb{R}} \eta\phi(z) + (1 - \eta)\phi(-z)$. Let

$$z^*(\eta) = \arg\inf_{z \in \mathbb{R}} \eta\phi(z) + (1 - \eta)\phi(-z) \ . \tag{43}$$

It is easy to verify that

$$z^*(\eta) = sgn(\eta - 1/2) \tag{44}$$

and

$$C^*(\eta) = \min\{2\eta, 2(1 - \eta)\} \ . \tag{45}$$

So we have,

$$\psi(\theta) = |\theta| \ . \tag{46}$$

Followed by theorem 2, we have $S = 1$ and $I = 1$. Thus for SVM, with a high probability, we have,

$$R(f_n) - R^* \leq \mathcal{O}\left(\frac{1}{n^p}\right) \ . \tag{47}$$

Example 2 (Exponential loss in Adaboost). We have $\phi(z) = e^{-z}$, which is convex and $\phi'(0) = -1 < 0$. Then, it's easy to derive that

$$z^*(\eta) = \frac{1}{2}\ln(\frac{\eta}{1 - \eta}) \tag{48}$$

and

$$C^*(\eta) = 2\sqrt{\eta(1 - \eta)} \ . \tag{49}$$

So,

$$\psi(\theta) = 1 - \sqrt{1 - \theta^2} \ . \tag{50}$$

Using Maclaurin expansion, we have,

$$\psi(\theta) = \frac{1}{2}\theta^2 + o(\theta^2) \ . \tag{51}$$

Thus, $S = \sqrt{2}$ and $I = \frac{1}{2}$. For Adaboost, with a high probability, we have

$$R(f_n) - R^* \leq \mathcal{O}\left(\frac{\sqrt{2}}{n^{\frac{p}{2}}}\right) . \tag{52}$$

Example 3 (Logistic loss in Logistic Regression). We have $\phi(z) = \log_2(1+e^{-z})$, which is convex and $\phi'(0) = -\frac{1}{2\ln 2} < 0$, we can follow similar procedures as before,

$$C^*(\eta) = -\eta \log_2 \eta - (1-\eta)\log_2(1-\eta) . \tag{53}$$

So,

$$\psi(\theta) = 1 - \frac{1+\theta}{2}\log_2\frac{2}{1+\theta} - \frac{1-\theta}{2}\log_2\frac{2}{1-\theta} . \tag{54}$$

Using Maclaurin expansion, we have,

$$\psi(\theta) = \frac{1}{2\ln 2}\theta^2 + o(\theta^2) . \tag{55}$$

Therefore $S = \sqrt{2\ln 2}$ and $I = \frac{1}{2}$. For Logistic Regression, with a high probability, we have

$$R(f_n) - R^* \leq \mathcal{O}\left(\frac{\sqrt{2\ln 2}}{n^{\frac{p}{2}}}\right) . \tag{56}$$

From the above examples, we conclude that SVM has a faster convergence rate to the Bayes optimal classifier than Adaboost and Logistic Regression. For better illustrations, we draw the leading term of the expansion of $\psi(\theta)$ around zero for these loss functions, as shown in Figure 2. It is obvious that the convergence behavior of the loss function is determined by the leading term of the expansion of $\psi(\theta)$ around zero. We note that for hinge loss, the leading term is sharper than other loss, which leads to a faster convergence, while other loss have similar asymptotic behavior around zero, which leads to the same but slower convergence rate to the Bayes optimal classifier.

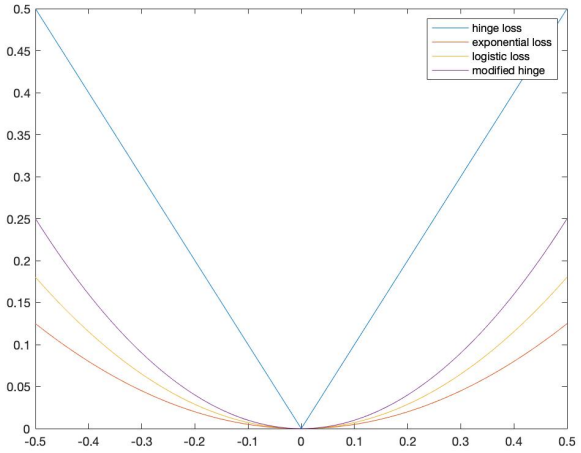


Fig. 2: The leading term of $\psi(\theta)$ for different loss functions, where we take $\Delta = 1$ for the modified hinge loss.

Note that the proposed methods also apply to many other surrogate loss functions. However, if we only need to compare the convergence rates of any two classification algorithms, we may not need to compute the consistency intensity of the two surrogate loss functions explicitly. The following theorem gives a general discriminant rule.

Theorem 4. *Given two convex Bayes-risk consistent loss functions ϕ_1 and ϕ_2, denoting their corresponding ψ-transform by ψ_1 and ψ_2, we assume that their excess ϕ-risk have the same convergence rate of order $\mathcal{O}(\frac{1}{n^p})$. Then, we define the intensity ratio λ as follows,*

$$\lambda = \lim_{\theta \to 0+} \frac{\psi_1(\theta)}{\psi_2(\theta)} . \tag{57}$$

We have the following statements:

- *if $0 < \lambda < +\infty$, then the minimizers w.r.t. ϕ_1 and ϕ_2 converge equally fast to the Bayes optimal classifier;*
- *if $\lambda = +\infty$, then the minimizer w.r.t. ϕ_1 converges faster to the Bayes optimal classifier;*
- *if $\lambda = 0$, then the minimizer w.r.t. ϕ_2 converges faster to the Bayes optimal classifier.*

Proof. Following Proposition 1 and Lemma 3, we have,

$$\psi_i(\theta) = M_i\theta^{\alpha_i} + o(\theta^{\alpha_i}) \quad \text{for} \quad i = 1, 2 \quad \text{and} \quad 0 < M_i < \infty \tag{58}$$

Then, we get,

$$\lambda = \lim_{\theta \to 0+} \frac{M_1\theta^{\alpha_1} + o(\theta^{\alpha_1})}{M_2\theta^{\alpha_2} + o(\theta^{\alpha_2})} = \frac{M_1}{M_2} \lim_{\theta \to 0+} \theta^{\alpha_1 - \alpha_2} \tag{59}$$

Thus, we can conclude:

- for $\lambda = \frac{M_1}{M_2} \in (0, +\infty)$, then we have $\alpha_1 = \alpha_2$ and $I_1 = I_2$. Therefore, the minimizers w.r.t. ϕ_1 and ϕ_2 converge equally fast to the Bayes optimal classifier;
- for $\lambda = +\infty$, we have $\alpha_1 < \alpha_2$ and $I_1 > I_2$. Thus the minimizer w.r.t. ϕ_1 converges faster to the Bayes optimal classifier;
- for $\lambda = 0$, then we have $\alpha_1 > \alpha_2$ and $I_1 < I_2$, which means that the minimizer w.r.t. ϕ_2 converges faster to the Bayes optimal classifier.

□

Using the notion of intensity ratio, we can compare the convergence rate to the Bayes optimal classifier for any two algorithms with Bayes-risk consistent loss functions without computing the consistency intensity.

We finish this section by introducing a scaling invariant property of the consistency intensity I, which is useful for comparing the convergence rate, e.g., when we scale the surrogate loss function $\phi(z)$ by $k_2\phi(k_1 z)$, we get the same I for $\phi(z)$ and $k_2\phi(k_1 z)$.

Theorem 5. *For any constants $0 < k_1, k_2 < +\infty$, the loss $\tilde{\phi}(z) = k_2\phi(k_1 z)$ have the same consistency intensity I as that of $\phi(z)$, which means that the intensity of surrogate loss function is scaling invariant in terms of ϕ.*

Proof. Notice that $\psi(R(f_n) - R^*) \leq R_\phi(f_n) - R_\phi^*$. If we scale ϕ as $\tilde{\phi}(z) = k_2\phi(z)$, the both sides of the inequality will be multiplied by k_2, which holds trivially.

We now consider $\tilde{\phi}(z) = \phi(k_1 z)$. Observe that,

$$\begin{aligned}\tilde{C}^*(\eta) &= \inf_{z\in\mathbb{R}} \eta\phi(k_1 z) + (1-\eta)\phi(-k_1 z) \\ &= \inf_{k_1 z\in\mathbb{R}} \eta\phi(k_1 z) + (1-\eta)\phi(-k_1 z) \\ &= \inf_{z'\in\mathbb{R}} \eta\phi(z') + (1-\eta)\phi(-z') = C^*(\eta)\end{aligned} \tag{60}$$

where $z' = k_1 z$. Then,

$$\tilde{\psi}(\theta) = \tilde{\phi}(0) - \tilde{C}^*(\eta) = \phi(0) - C^*(\eta) = \psi(\theta) \ , \tag{61}$$

which leads to the same I. □

Theorem 5 has many applications. For example, the exponential loss in Adaboost is $\phi(z) = e^{-z}$. Then $\phi_k(z) = e^{-kz}$ for all constants $k > 0$ must have the same consistency intensity as that of $\phi(z)$, which implies the minimizers of the corresponding empirical surrogate risks converge equally fast to the Bayes optimal classifier.

B. SVM$_\Delta$: An Example of the Data-dependent Loss Modification Method

In the previous sections, we have provided theorems that can measure the convergence rate from the expected risk $R(f_n)$ to the Bayes risk R^* for many leaning algorithms using different surrogate loss functions. In fact, the notions of consistency intensity and conductivity can achieve something beyond that. In this subsection, we propose a data-dependent loss modification method for SVM, that obtains a faster convergence rate for the bound of the excess risk $R(f_n) - R^*$ and thus makes the learning algorithm achieve a faster convergence rate to the Bayes optimal classifier.

We are familiar with the standard SVM that uses the hinge loss as a surrogate. Now, we modify the hinge loss as follows:

$$\phi(z) = max\{1-z, 0\}^{1+\Delta} \quad \text{where} \quad 0 < \Delta < +\infty \ . \tag{62}$$

We have $\phi'(0) = -(1+\Delta) < 0$, which means this modified hinge loss is also Bayes-risk consistent. Then, following the similar procedure as done for the above examples, we have,

$$C(\eta) = \eta * max\{1-z, 0\}^{1+\Delta} + (1-\eta) * max\{1+z, 0\}^{1+\Delta} \ . \tag{63}$$

It's easy to verify that

$$z^*(\eta) = \frac{\eta^{\frac{1}{\Delta}} - (1-\eta)^{\frac{1}{\Delta}}}{\eta^{\frac{1}{\Delta}} + (1-\eta)^{\frac{1}{\Delta}}} \ , \tag{64}$$

and so

$$C^*(\eta) = \frac{2^{1+\Delta}\eta(1-\eta)}{[\eta^{\frac{1}{\Delta}} + (1-\eta)^{\frac{1}{\Delta}}]^{\Delta}} \ . \tag{65}$$

Thus,

$$\begin{aligned}\psi(\theta) &= \phi(0) - C^*(\frac{1+\theta}{2}) \\ &= 1 - \frac{2^{\Delta}(1-\theta^2)}{[(1+\theta)^{\frac{1}{\Delta}} + (1-\theta)^{\frac{1}{\Delta}}]^{\Delta}} \ .\end{aligned} \tag{66}$$

Using Maclaurin expansion, we have,

$$\psi(\theta) = \frac{1+\Delta}{2\Delta}\theta^2 + o(\theta^2) \ . \tag{67}$$

Therefore, we have that intensity $I = \frac{1}{2}$ and conductivity $S = \sqrt{\frac{2\Delta}{1+\Delta}}$. According to Theorem 2, we know that, with a high probability, we have,

$$R(f_n) - R^* \leq \mathcal{O}\left(\frac{S}{n^{pI}}\right) = \mathcal{O}\left(\sqrt{\frac{2\Delta}{1+\Delta} * \frac{1}{n^p}}\right) \ . \tag{68}$$

From (68), we find that a tighter bound can be obtained when Δ converges to zero fast. Here, we introduce the notion of data-dependent loss modification. That is, the modification parameter Δ is dependent on the sample size n. For example, if $\Delta = \mathcal{O}(\frac{1}{n^2})$ and $p = \frac{1}{2}$, with a high probability, we can obtain a bound of order $\mathcal{O}\left(\frac{1}{n^{\frac{5}{4}}}\right)$, which, to our best knowledge, is the first bound faster than $\mathcal{O}\left(\frac{1}{n}\right)$ without the low noise assumption. Therefore, a tighter bound is obtained with our proposed method for SVM.

V. Conclusions and Future Work

In this paper, we defined the notions of consistency intensity and conductivity for convex Bayes-risk consistent surrogate loss functions and proposed a general framework that determines the relationship between the convergence rate of the excess risk and the convergence rate of the excess ϕ-risk. Our methods were used to compare the convergence rates to the Bayes optimal classifier for empirical minimizers of different classification algorithms. Moreover, we used the notions of consistency intensity and conductivity to guide modifying of surrogate loss functions so as to achieve a faster convergence rate.

In this work, we need the surrogate loss function to be convex and Bayes-risk consistent, which holds true for many different surrogate loss functions. However, sometimes we may encounter non-convex, but still Bayes-risk consistent loss functions. It is interesting to generalize the obtained results to the non-convex situation in the future. Besides, in the future, we will apply the modified surrogate loss function to some real-world problems. Moreover, finding some other approaches that guide modifying existing surrogate losses to achieve a faster convergence rate to the Bayes optimal classifier is also quite worth exploring.

References

[Agarwal and Agarwal, 2015] Agarwal, A. and Agarwal, S. (2015). On consistent surrogate risk minimization and property elicitation. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 4–22, Paris, France. PMLR.

[Bartlett et al., 2005] Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

[Bartlett et al., 2006] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

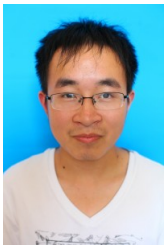
[Bartlett and Mendelson, 2002] Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

[Ben-David et al., 2003] Ben-David, S., Eiron, N., and Long, P. M. (2003). On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496 – 514.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[Bousquet et al., 2004] Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer.

[Collins et al., 2002] Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285.
[Feldman et al., 2012] Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. (2012). Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590.
[Koltchinskii et al., 2006] Koltchinskii, V. et al. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.
[Liu et al., 2017] Liu, T., Lugosi, G., Neu, G., and Tao, D. (2017). Algorithmic stability and hypothesis complexity. *arXiv preprint arXiv:1702.08712*.
[Lugosi and Vayatis, 2004] Lugosi, G. and Vayatis, N. (2004). On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, pages 30–55.
[Mohri et al., 2012] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.
[Neykov et al., 2016] Neykov, M., Liu, J. S., and Cai, T. (2016). On the characterization of a class of fisher-consistent loss functions and its application to boosting. *Journal of Machine Learning Research*, 17(70):1–32.
[Sridharan et al., 2009] Sridharan, K., Shalev-shwartz, S., and Srebro, N. (2009). Fast rates for regularized objectives. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1545–1552. Curran Associates, Inc.
[Steinwart, 2005] Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142.
[Tsybakov, 2004] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166.
[Vapnik, 2013] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.
[Zhang, 2002] Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550.
[Zhang, 2004] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85.



Jingwei Zhang received the MPhil degree in computer science from the University of Sydney in 2019 and B.E. degree in electronics engineering and information science from the University of Science and Technology of China, in 2017. He is currently pursuing the Ph.D. degree in computer science from the Hong Kong University of Science and Technology. His research interests include machine learning and theory of deep learning.

Tongliang Liu received the B.E. degree in electronic engineering and information science from the University of Science and Technology of Chi- na, Hefei, China, in 2012, and the Ph.D. degree from the University of Technology Sydney, Sydney, Australia, in 2016. He was a visiting Ph.D. student with the Barcelona Graduate School of Economics and the Department of Economics, Pompeu Fabra University, for six months. He is currently a Lecturer with the School of Computer Science and the Faculty of Engineering and Information Technologies, the University of Sydney. He has authored or co-authored over ten research papers, including the IEEE T- PAMI, T-NNLS, T-IP, NECO, ICML, KDD, IJCAI, and AAAI. His research interests include statistical learning theory, computer vision, and optimization. He received the Best Paper Award in the IEEE International Conference on Information Science and Technology 2014.

Dacheng Tao (F'15) is currently a Professor of Computer Science with the School of Computer Science and the Faculty of Engineering and Information Technologies, the University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and over 200 publications at prestigious journals and prominent conferences, such as the IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM, and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in the IEEE ICDM07, the best student paper award in the IEEE ICDM13, and the 2014 ICDM 10-year highest-impact paper award. He is a fellow of the IEEE, OSA, IAPR, and SPIE. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellors Medal for Exceptional Research.