Data-dependent Generalization Bounds for Multi-class Classification

Yunwen Lei

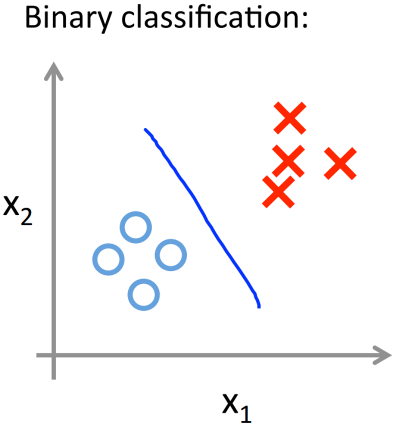
University of Kaiserslautern

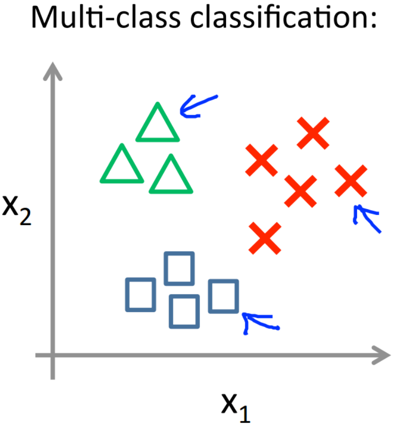
yunwen.lei@hotmail.com

Joint work with Ürün Dogan, Alexander Binder, Ding-Xuan Zhou and Marius Kloft

Outline

Multi-class Classification (MCC): Classic Problem in ML

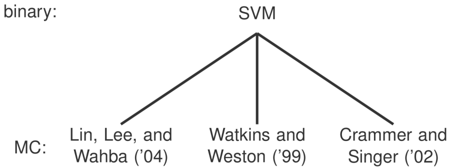


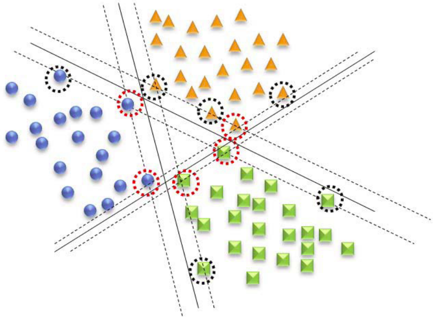


Many MCC **Algorithms** out there...

E.g.:

- Multinomial logistic regression
- Multi-class SVMs





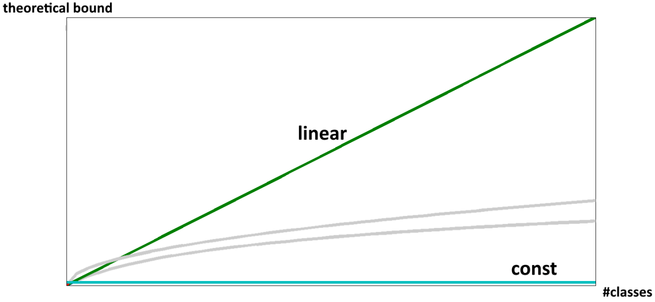
In this talk: **Theory** for MCC

Koltchinskii and Panchenko (2002); Mohri et al. (2012); Kuznetsov et al. (2014)

In this talk: **Theory** for MCC

Especially interesting in XC:

What is the **scaling** of generalization bounds for MCC in the **number of classes**?

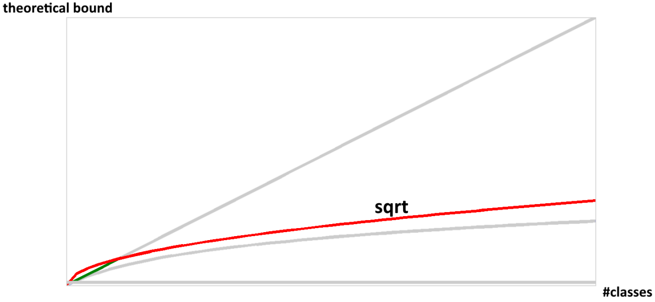


Koltchinskii and Panchenko (2002); Mohri et al. (2012); Kuznetsov et al. (2014)

In this talk: **Theory** for MCC

Especially interesting in XC:

What is the **scaling** of generalization bounds for MCC in the **number of classes**?

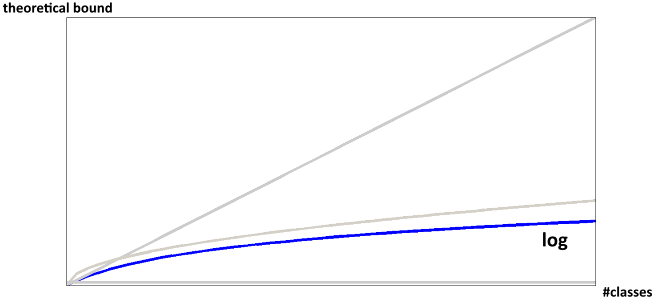


Lei et al. (2015); Maurer (2016); Cortes et al. (2016)

In this talk: **Theory** for MCC

Especially interesting in XC:

What is the **scaling** of generalization bounds for MCC in the **number of classes**?



Lei et al. (2019)

Outline

Problem Setting

Multi-class Classification

Given training data:

- (,`dog`), (,`car`), (,`airplane`), ...

Multi-class Classification

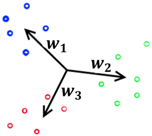
Given training data:

- (,`dog`), (,`car`), (,`airplane`), ...
- Formally $\underbrace{z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)}_{\in \mathcal{X} \times \mathcal{Y}} \overset{\text{i.i.d.}}{\sim} P$
  - $\mathcal{Y} := \{1, 2, \ldots, \mathbf{c}\}$
  - $\mathbf{c}$ = number of classes

Multi-class Classification

Aim:

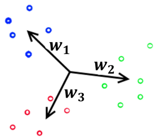
- Define a hypothesis class H of functions $h = (h_1, \ldots, h_c)$
  - e.g., $h_y(x) = \langle \mathbf{w}_y, \phi(x) \rangle \in H_K$



Multi-class Classification

Aim:

- Define a hypothesis class H of functions $h = (h_1, \ldots, h_c)$
  - e.g., $h_y(x) = \langle \mathbf{w}_y, \phi(x) \rangle \in H_K$



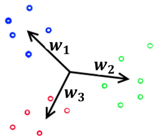
- Find an $h \in H$ that "predicts well" via

$$\hat{y} := \boxed{\arg\max}_{y \in \mathcal{Y}} h_y(x)$$

Multi-class Classification

Aim:

- Define a hypothesis class H of functions $h = (h_1, \ldots, h_c)$
  - e.g., $h_y(x) = \langle \mathbf{w}_y, \phi(x) \rangle \in H_K$



- Find an $h \in H$ that "predicts well" via

$$\hat{y} := \boxed{\arg\max}_{y \in \mathcal{Y}} h_y(x)$$

- Want $h_{y_i}(x_i)$ being larger than all other $h_y(x_i)$
  - otherwise loss incurred through loss function $\Psi_y : \mathbb{R}^c \to \mathbb{R}_+$

Want: small generalization error $\mathbb{E}_{\mathbf{X},\mathbf{Y}} \Psi_Y(h_Y(X))$.

Types of Generalization bounds for MCC

Data-independent bounds

- based on covering numbers
  (Guermeur, 2002; Zhang, 2004a,b; Hill and Doucet, 2007)
- unable to adapt to data

Data-dependent bounds

- based on Rademacher complexity
  (Koltchinskii and Panchenko, 2002; Mohri et al., 2012; Cortes et al., 2013; Kuznetsov et al., 2014)
- computable from the data

In this talk: **data-dependent bounds**

Generalization Error Bounds

Data-dependent bounds based on **Rademacher Complexity** (RC)

Definition (RC)

$$\mathfrak{R}_S(H) := \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H} \frac{1}{n}\sum_{i=1}^{n} \boxed{\epsilon_i}\, h(z_i)\Big]$$

where $\epsilon_1, \ldots, \epsilon_n$ are random signs ("Rademacher variables")

Interpretation: RC measures how much **the hypothesis class can correlate with random noise**.

Data-dependent bounds based on **Rademacher Complexity** (RC)

Definition (RC)

$$\mathfrak{R}_S(H) := \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}\boxed{\epsilon_i}\,h(z_i)\Big]$$

where $\epsilon_1, \ldots, \epsilon_n$ are random signs ("Rademacher variables")

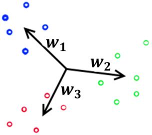
Interpretation: RC measures how much **the hypothesis class can correlate with random noise**.

$$\forall h \in H_K^c:\ \underbrace{\mathbb{E}_Y \Psi_Y(h(X))}_{\text{expectation}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Psi_{y_i}(h(x_i))}_{\text{empirical}} \le 2\mathfrak{R}_S\Big(\Psi_y(h(x)) : h \in H_K^c\Big)$$

Data-dependent bounds based on **RC**

Example (Crammer & Singer):

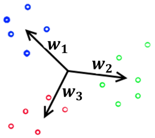
$$H = \{ h^{\mathbf{w}} = (\langle \mathbf{w}_1, \phi(x) \rangle, \ldots, \langle \mathbf{w}_c, \phi(x) \rangle) : \mathbf{w} = (\mathbf{w}_j)_{j=1}^{c}, \boxed{\sum_{j=1}^{c} \|\mathbf{w}_j\|_2^2} \leq 1 \}$$



Data-dependent bounds based on **RC**

Example (Crammer & Singer):

$$H = \{h^{\mathbf{w}} = (\langle \mathbf{w}_1, \phi(x)\rangle, \ldots, \langle \mathbf{w}_c, \phi(x)\rangle) : \mathbf{w} = (\mathbf{w}_j)_{j=1}^c, \boxed{\sum_{j=1}^{c} \|\mathbf{w}_j\|_2^2} \le 1\}$$



Multi-class margin: for any $h : \mathcal{X} \mapsto \mathbb{R}^c$, we denote by

$$\rho_h(\mathbf{x}, y) := h_y(\mathbf{x}) - \max_{y' : y' \neq y} h_{y'}(\mathbf{x}) \tag{1}$$

Multi-class margin loss:

$$\Psi_y(h(\mathbf{x})) = \max\left(1 - \rho_h(\mathbf{x}, y), 0\right)$$

Key step is to estimate

$$\mathfrak{R}_S\Big(\Psi_y(h) : h \in H\Big) \Leftarrow \mathfrak{R}_S\Big(\rho_h(\mathbf{x}, y) : h \in H\Big) \Leftarrow \mathfrak{R}_S\Big(\max_{j=1,\ldots,c}(h(\mathbf{x})) : h \in H\Big)$$

linear dependency on #classes

Classic analysis based on:

$$\mathfrak{R}_S(\max\{h_1, \ldots, h_c\} : h_j \in H_j, j = 1, \ldots, c) \leq \boxed{\sum_{j=1}^{c} \mathfrak{R}_S(H_j)} \tag{2}$$

Koltchinskii and Panchenko (2002); Mohri et al. (2012); Cortes et al. (2013); Kuznetsov et al. (2014)

linear dependency on #classes

Classic analysis based on:

$$\mathfrak{R}_S(\max\{h_1, \ldots, h_c\} : h_j \in H_j, j = 1, \ldots, c) \leq \boxed{\sum_{j=1}^{c} \mathfrak{R}_S(H_j)} \tag{2}$$

Implies linear dependence on number of classes

Koltchinskii and Panchenko (2002); Mohri et al. (2012); Cortes et al. (2013); Kuznetsov et al. (2014)

From **linear** to **sqrt** dependency on #classes

Key is to use the Lipschitz continuity of loss function:

A function $f : \mathbb{R}^c \to \mathbb{R}$ is L-Lips. cont. w.r.t. a norm $\|\cdot\|$ in $\mathbb{R}^c$ if

$$|f(\mathbf{t}) - f(\mathbf{t}')| \leq L\|(t_1 - t'_1, \ldots, t_c - t'_c)\|, \quad \forall \mathbf{t}, \mathbf{t}' \in \mathbb{R}^c.$$

- e.g., ℓ_∞-norm: $\|\mathbf{t}\|_\infty = \max_{j=1,\ldots,c} |t_j|$ (Crammer & Singer)

Y. Lei, U. Dogan, A. Binder, and M. Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms.
In **Advances in Neural Information Processing Systems**, pages 2026–2034, 2015.

From **linear** to **sqrt** dependency on #classes

Key is to use the **Lipschitz** continuity of loss function:

A function $f : \mathbb{R}^c \to \mathbb{R}$ is L-Lips. cont. w.r.t. a norm $\|\cdot\|$ in $\mathbb{R}^c$ if

$$|f(\mathbf{t}) - f(\mathbf{t}')| \leq L\|(t_1 - t_1', \ldots, t_c - t_c')\|, \quad \forall \mathbf{t}, \mathbf{t}' \in \mathbb{R}^c.$$

- e.g., ℓ_∞-norm: $\|\mathbf{t}\|_\infty = \max_{j=1,\ldots,c} |t_j|$ (Crammer & Singer)

Key result

If $f_1, \ldots, f_n$ are L-Lips. cont. w.r.t. $\|\cdot\|_2$, then

$$\mathbb{E}_\epsilon \sup_{h=(h_1,\ldots,h_c)\in H} \sum_{i=1}^n \epsilon_i f_i(h(x_i)) \leq \sqrt{2} L \mathbb{E}_\epsilon \sup_{h=(h_1,\ldots,h_c)\in H} \sum_{i=1}^n \sum_{j=1}^c \epsilon_{ij} h_j(x_i) \tag{3}$$

Y. Lei, U. Dogan, A. Binder, and M. Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms.
In **Advances in Neural Information Processing Systems**, pages 2026–2034, 2015.

Crammer & Singer

The function $f_i(\mathbf{t}) = \max_{j=1,\ldots,c} t_j$ is 1-Lipschitz continuous w.r.t. ℓ_2-norm:

$$\Big|\max_{j=1,\ldots,c} t_j - \max_{j=1,\ldots,c} \tilde{t}_j\Big| \leq \|\mathbf{t} - \tilde{\mathbf{t}}\|_2 = \Big(\sum_{j=1}^{c} |t_j - \tilde{t}_j|^2\Big)^{1/2}$$

We have the **constraint**: $\sum_{j=1}^{c} \|\mathbf{w}_j\|_2^2 \leq 1$

- by (2),

$$\mathfrak{R}_S\Big(\max_{j=1,\ldots,c}(h(\mathbf{x})) : h \in H\Big) \leq \sum_{j=1}^{c} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\|\mathbf{w}_j\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} g_i \langle \mathbf{w}_j, \mathbf{x}_i \rangle$$

- by (3),

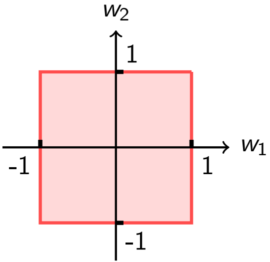
$$\mathfrak{R}_S\Big(\max_{j=1,\ldots,c}(h(\mathbf{x})) : h \in H\Big) \leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\|(\mathbf{w}_1,\ldots,\mathbf{w}_c)\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} g_i \langle \mathbf{w}_j, \mathbf{x}_i \rangle$$

Result Preserves Correlation

classic result (2):

$$\sup_{\|w_1\|_2 \le 1} \sum_{i=1}^{n} \epsilon_i \langle w_1, x_i \rangle + \sup_{\|w_2\|_2 \le 1} \sum_{i=1}^{n} \epsilon_i \langle w_2, x_i \rangle$$

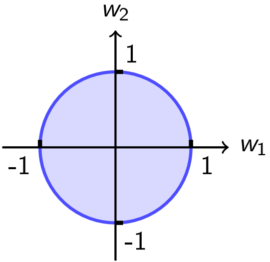
supremum taken **separately**



Lipschitz result (3):

$$\sup_{\|(w_1, w_2)\|_2 \le 1} \sum_{i=1}^{n} [\epsilon_{i1} \langle w_1, x_i \rangle + \epsilon_{i2} \langle w_2, x_i \rangle]$$

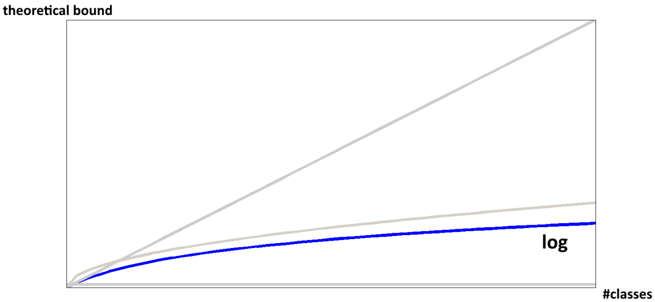
supremum taken **jointly**



Preserving the coupling means supremum in a smaller space!

Overview

Overview



Key observation

- Structural result (3) uses **lipschitz continuity** of maximum w.r.t. $\|\cdot\|_2$

$$\Big|\max_{j=1,\dots,c} t_j - \max_{j=1,\dots,c} \tilde{t}_j\Big| \le \|\mathbf{t}-\tilde{\mathbf{t}}\|_2 = \Big(\sum_{j=1}^{c} |t_j - \tilde{t}_j|^2\Big)^{1/2}$$

- However, maximum is 1-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$

$$\Big|\max_{j=1,\dots,c} t_j - \max_{j=1,\dots,c} \tilde{t}_j\Big| \le \|\mathbf{t}-\tilde{\mathbf{t}}\|_\infty = \max_{j=1,\dots,c} |t_j - \tilde{t}_j|$$

- the same Lipschitz constant but ℓ_∞-norm is much milder:

$$\|\mathbf{t}\|_2 = \sqrt{c}\|\mathbf{t}\|_\infty \quad \text{if elements of } \mathbf{t} \text{ are the same}$$

Key observation

- Structural result (3) uses **lipschitz continuity** of maximum w.r.t. $\|\cdot\|_2$

$$\Big|\max_{j=1,\ldots,c} t_j - \max_{j=1,\ldots,c} \tilde{t}_j\Big| \leq \|\mathbf{t} - \tilde{\mathbf{t}}\|_2 = \Big(\sum_{j=1}^{c} |t_j - \tilde{t}_j|^2\Big)^{1/2}$$

- However, maximum is 1-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$

$$\Big|\max_{j=1,\ldots,c} t_j - \max_{j=1,\ldots,c} \tilde{t}_j\Big| \leq \|\mathbf{t} - \tilde{\mathbf{t}}\|_\infty = \max_{j=1,\ldots,c} |t_j - \tilde{t}_j|$$

- the same Lipschitz constant but ℓ_∞-norm is much milder:

$$\|\mathbf{t}\|_2 = \sqrt{c}\|\mathbf{t}\|_\infty \quad \text{if elements of } \mathbf{t} \text{ are the same}$$

Can we directly use ℓ_∞ Lipschitz continuity?

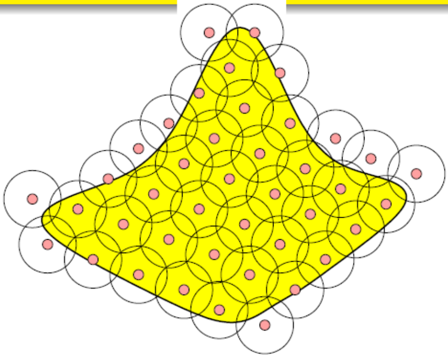
Background: Covering numbers

- F is a class of scalar-valued functions defined over a space $\tilde{\mathcal{Z}}$
- $S := \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \subset \tilde{\mathcal{Z}}$ is a set of cardinality n

$\{\mathbf{v}^1, \ldots, \mathbf{v}^m\} \subset \mathbb{R}^n$ is an (ϵ, ℓ_∞)-cover of F w.r.t. S if

$$\sup_{f \in F} \min_{j=1,\ldots,m} \max_{i=1,\ldots,n} |f(\mathbf{z}_i) - \mathbf{v}_i^j| \leq \epsilon.$$

$\mathcal{N}_\infty(\epsilon, F, n)$: the smallest cardinality m of such an (ϵ, ℓ_∞)-cover



Core Idea

Introduce the **linear** and scalar-valued function class

$$\widetilde{H} := \{\mathbf{v} \to \langle \mathbf{w}, \mathbf{v} \rangle : \|\mathbf{w}\| \leq 1, \mathbf{v} \in \widetilde{S}\},$$

$$\widetilde{S} := \{\underbrace{\tilde{\phi}_1(x_1), \tilde{\phi}_2(x_1), \ldots, \tilde{\phi}_c(x_1)}_{\text{induced by } \mathbf{x}_1}, \ldots, \underbrace{\tilde{\phi}_1(x_n), \tilde{\phi}_2(x_n), \ldots, \tilde{\phi}_c(x_n)}_{\text{induced by } \mathbf{x}_n}\},$$

$$\tilde{\phi}_j(x) := \big(\underbrace{0, \ldots, 0}_{j-1}, \phi(x), \underbrace{0, \ldots, 0}_{c-j}\big) \in H_K^c, \quad j \in \mathbb{N}_c.$$

Core Idea

Introduce the **linear** and scalar-valued function class

$$\widetilde{H} := \{\mathbf{v} \to \langle \mathbf{w}, \mathbf{v} \rangle : \|\mathbf{w}\| \le 1, \mathbf{v} \in \widetilde{S}\},$$

$$\widetilde{S} := \{\underbrace{\tilde{\phi}_1(x_1), \tilde{\phi}_2(x_1), \ldots, \tilde{\phi}_c(x_1)}_{\text{induced by } \mathbf{x}_1}, \ldots, \underbrace{\tilde{\phi}_1(x_n), \tilde{\phi}_2(x_n), \ldots, \tilde{\phi}_c(x_n)}_{\text{induced by } \mathbf{x}_n}\},$$

$$\tilde{\phi}_j(x) := \big(\underbrace{0, \ldots, 0}_{j-1}, \phi(x), \underbrace{0, \ldots, 0}_{c-j}\big) \in H_K^c, \quad j \in \mathbb{N}_c.$$

Key identity:

$$\langle \mathbf{w}, \tilde{\phi}_j(\mathbf{x}_i) \rangle = \Big\langle (\mathbf{w}_1, \ldots, \mathbf{w}_c), \big(\underbrace{0, \ldots, 0}_{j-1}, \phi(\mathbf{x}_i), \underbrace{0, \ldots, 0}_{c-j}\big) \Big\rangle = \langle \mathbf{w}_j, \phi(\mathbf{x}_i) \rangle.$$

Core Idea

Introduce the **linear** and scalar-valued function class

$$\widetilde{H} := \{\mathbf{v} \to \langle \mathbf{w}, \mathbf{v} \rangle : \|\mathbf{w}\| \leq 1, \mathbf{v} \in \widetilde{S}\},$$

$$\widetilde{S} := \{\underbrace{\tilde{\phi}_1(x_1), \tilde{\phi}_2(x_1), \ldots, \tilde{\phi}_c(x_1)}_{\text{induced by } \mathbf{x}_1}, \ldots, \underbrace{\tilde{\phi}_1(x_n), \tilde{\phi}_2(x_n), \ldots, \tilde{\phi}_c(x_n)}_{\text{induced by } \mathbf{x}_n}\},$$

$$\tilde{\phi}_j(x) := \big(\underbrace{0, \ldots, 0}_{j-1}, \phi(x), \underbrace{0, \ldots, 0}_{c-j}\big) \in H_K^c, \quad j \in \mathbb{N}_c.$$

Key identity:

$$\langle \mathbf{w}, \tilde{\phi}_j(\mathbf{x}_i) \rangle = \Big\langle (\mathbf{w}_1, \ldots, \mathbf{w}_c), \big(\underbrace{0, \ldots, 0}_{j-1}, \phi(\mathbf{x}_i), \underbrace{0, \ldots, 0}_{c-j}\big) \Big\rangle = \langle \mathbf{w}_j, \phi(\mathbf{x}_i) \rangle.$$

Traversing all i, j means
extracting all components $\mathbf{w}_j$ over all examples $\mathbf{x}_i$

New Structural Result based on Covering Numbers

$$\mathcal{N}_\infty(\epsilon, \{\Psi_y(h(\mathbf{x})) : h \in H\}, n) \leq \mathcal{N}_\infty(\epsilon/L, \widetilde{H}, \boxed{nc}). \quad (4)$$

Y. Lei, U. Dogan, D.-X. Zhou, and M. Kloft. Data-dependent generalization error bounds for multi-class classification. **IEEE Transactions on Information Theory**, to appear.

New Structural Result based on Covering Numbers

$$\mathcal{N}_\infty(\epsilon, \{\Psi_y(h(\mathbf{x})) : h \in H\}, n) \le \mathcal{N}_\infty(\epsilon/L, \widetilde{H}, \boxed{nc}). \qquad (4)$$

- Complexity of $\widetilde{H}$ is readily tackled (Zhang, 2002; Srebro et al., 2010)

Y. Lei, U. Dogan, D.-X. Zhou, and M. Kloft. Data-dependent generalization error bounds for multi-class classification. **IEEE Transactions on Information Theory**, to appear.

Main result

Theorem (Lei, Dogan, Zhou, and Kloft, 2019)

If Ψ_y is L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$, then

$$\mathfrak{R}_S(F) \leq 27L\sqrt{c}\,\boxed{\mathfrak{R}_{nc}(\widetilde{H})}.$$

Main result

Theorem (Lei, Dogan, Zhou, and Kloft, 2019)

If Ψ_y is L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$, then

$$\mathfrak{R}_S(F) \leq 27L\sqrt{c}\,\boxed{\mathfrak{R}_{nc}(\widetilde{H})}.$$

Proof?

$\mathfrak{R}_S(F)$ $\mathcal{N}_\infty(\epsilon, F, n)$ $\mathcal{N}_\infty(\epsilon, \widetilde{H}, nc)$ $\mathfrak{R}_{nc}(\widetilde{H})$

Main result

Theorem (Lei, Dogan, Zhou, and Kloft, 2019)

If Ψ_y is L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$, then

$$\mathfrak{R}_S(F) \leq 27L\sqrt{c}\,\boxed{\mathfrak{R}_{nc}(\widetilde{H})}.$$

Proof?

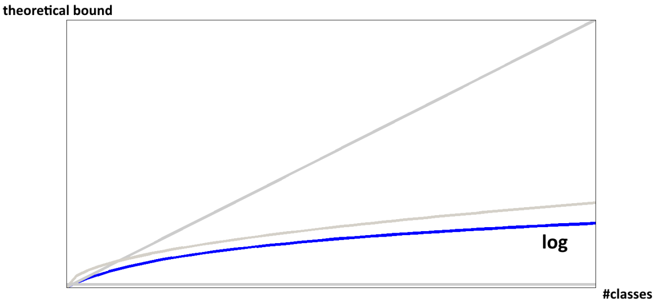
$\mathfrak{R}_S(F)$ $\quad\quad$ $\mathcal{N}_\infty(\epsilon, F, n)$ $\quad\quad$ $\mathcal{N}_\infty(\epsilon, \widetilde{H}, nc)$ $\quad\quad$ $\mathfrak{R}_{nc}(\widetilde{H})$

Example

If $\|\mathbf{w}\| = \|\mathbf{w}\|_2$, then

$$\max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2 (2nc)^{-\frac{1}{2}} \leq \mathfrak{R}_{nc}(\widetilde{H}) \leq \max_{i\in\mathbb{N}_n} \|\phi(\mathbf{x}_i)\|_2 (nc)^{-\frac{1}{2}}.$$

Overview



Applications & Discussions

Applications–classic MC-SVMs

MC-SVM in Cramer & Singer (2002): Crammer and Singer (2002)

$$\min_{\mathbf{w}} \ \frac{1}{2}\Big[\sum_{j=1}^{c} \|\mathbf{w}_j\|_2^2\Big] + C\sum_{i=1}^{n} \max_{y':y'\neq y_i} \big(1 - \langle \mathbf{w}_{y_i} - \mathbf{w}_{y'}, \phi(x_i)\rangle\big)_+$$

Multinomial logistic regression: Bishop (2006)

$$\min_{\mathbf{w}} \ \frac{1}{2}\Big[\sum_{j=1}^{c} \|\mathbf{w}_j\|_2^2\Big] + C\sum_{i=1}^{n} \log\Big(\sum_{y'=1}^{c} \exp\big(\langle \mathbf{w}_{y'} - \mathbf{w}_{y_i}, \phi(x_i)\rangle\big)\Big)$$

Applications–**classic MC-SVMs**

MC-SVM in Cramer & Singer (2002): Crammer and Singer (2002)

$$\min_{\mathbf{w}} \frac{1}{2}\left[\sum_{j=1}^{c} \|\mathbf{w}_j\|_2^2\right] + C\sum_{i=1}^{n} \max_{y': y' \neq y_i} \left(1 - \langle \mathbf{w}_{y_i} - \mathbf{w}_{y'}, \phi(x_i)\rangle\right)_+$$

Multinomial logistic regression: Bishop (2006)

$$\min_{\mathbf{w}} \frac{1}{2}\left[\sum_{j=1}^{c} \|\mathbf{w}_j\|_2^2\right] + C\sum_{i=1}^{n} \log\left(\sum_{y'=1}^{c} \exp\left(\langle \mathbf{w}_{y'} - \mathbf{w}_{y_i}, \phi(x_i)\rangle\right)\right)$$

Classic bound by (2)	$O\left(n^{-1}\boxed{c}\sqrt{\sum_{i=1}^{n}\langle\phi(x_i),\phi(x_i)\rangle}\right)$
Lipschitz bound by (3)	$O\left(n^{-1}\sqrt{\boxed{c}\sum_{i=1}^{n}\langle\phi(x_i),\phi(x_i)\rangle}\right)$
Covering no. bound by (4)	$O\left(n^{-\frac{1}{2}}\boxed{\log c}\max_{i\in\mathbb{N}_n}\|\phi(x_i)\|_2\right)$

Applications–ℓ_p-norm MC-SVM

ℓ_p-norm MC-SVM (Lei, Dogan, Binder, and Kloft, 2015)

$$\min_{\mathbf{w}} \ \frac{1}{2}\Big[\sum_{j=1}^{c}\|\mathbf{w}_j\|_2^p\Big]^{\frac{2}{p}} + C\sum_{i=1}^{n}\max_{y':y'\neq y_i}\big(1-\langle \mathbf{w}_{y_i}-\mathbf{w}_{y'},\phi(x_i)\rangle\big)_+$$

Applications–ℓ_p-norm MC-SVM

ℓ_p-norm MC-SVM (Lei, Dogan, Binder, and Kloft, 2015)

$$\min_{\mathbf{w}} \ \frac{1}{2}\Big[\sum_{j=1}^{c}\|\mathbf{w}_j\|_2^p\Big]^{\frac{2}{p}} + C\sum_{i=1}^{n}\max_{y':y'\neq y_i}\big(1-\langle \mathbf{w}_{y_i}-\mathbf{w}_{y'},\phi(x_i)\rangle\big)_+$$

Classic bound by (2)	$O\Big(n^{-1}\boxed{c}\sqrt{\sum_{i=1}^n\langle\phi(x_i),\phi(x_i)\rangle}\Big)$
Lipschitz bound by (3)	$O\Big(n^{-1}\boxed{c^{1-\frac{1}{p}}}\sqrt{\sum_{i=1}^n\langle\phi(x_i),\phi(x_i)\rangle}\Big)$
Covering no. bound by (4)	$O\Big(n^{-\frac{1}{2}}\boxed{c^{\frac{1}{2}-\frac{1}{\max(2,p)}}\log c}\max_{i\in\mathbb{N}_n}\|\phi(x_i)\|_2\Big)$

- Bound by (3) enjoys **logarithmic** dependency if $p \approx 1$ and sublinear dependency $c^{1-\frac{1}{p}}$ otherwise — Lei et al. (2015)
- Bound by (4) enjoys **logarithmic** dependency if $p \leq 2$ and sublinear dependency $c^{\frac{1}{2}-\frac{1}{p}}$ otherwise — Lei et al. (2019)

Empirical Verification

- We consider two datasets **ALOI** and **Sector**
- Vary the number of classes by grouping class labels
- **Approximation of the Empirical Rademacher Complexity** (AERC) defined by

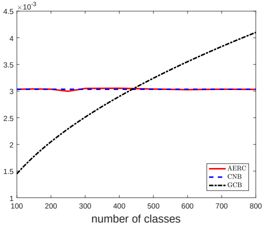
$$\mathrm{AERC}(F) := \frac{1}{50}\sum_{t=1}^{50}\widetilde{\mathfrak{R}}_S(\boldsymbol{\epsilon}^{(t)}, F),$$

where (Monte Carlo approximation)

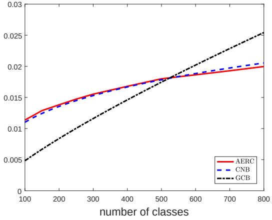
$$\widetilde{\mathfrak{R}}_S(\boldsymbol{\epsilon}, F) := \frac{1}{n}\sup_{\substack{\mathbf{w}\in\mathbb{R}^{d\times c}\\ \|\mathbf{w}\|_{2,p}\leq\Lambda}}\sum_{i=1}^{n}\epsilon_i\Psi_{y_i}\big(\langle\mathbf{w}_1,\mathbf{x}_i\rangle,\ldots,\langle\mathbf{w}_c,\mathbf{x}_i\rangle\big). \tag{5}$$

AERC w.r.t. #classes

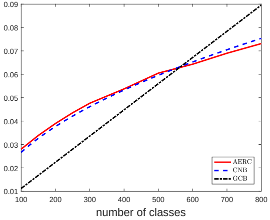
ALOI



$p = 2$



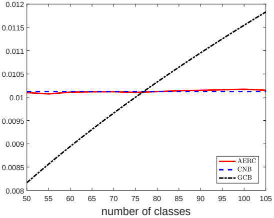
$p = 5$



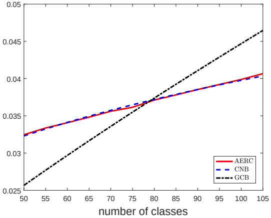
$p = \infty$

AERC w.r.t. #classes

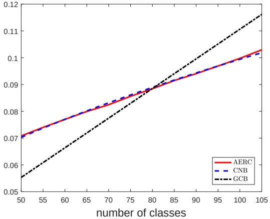
Sector



$p = 2$



$p = 5$



$p = \infty$

Conclusions & Future Directions

Conclusions:

- New data-dependent bound with **mild** dependency on c
  - logarithmic for Cramer & Singer MC-SVM
  - logarithmic for Multinomial logistic regression
  - sublinear for ℓ_p-norm MC-SVM
- Key is **structural result** (4) using lips. cont. w.r.t. $\|\cdot\|_\infty$

Conclusions & Future Directions

Conclusions:

- New data-dependent bound with **mild** dependency on c
  - logarithmic for Cramer & Singer MC-SVM
  - logarithmic for Multinomial logistic regression
  - sublinear for ℓ_p-norm MC-SVM
- Key is **structural result** (4) using lips. cont. w.r.t. $\|\cdot\|_\infty$

Directions:

- Extension to multi-label
- A data-dependent bound **independent** of the class size?

References I

C. M. Bishop. Pattern recognition and machine learning. Springer, 2006.

C. Cortes, M. Mohri, and A. Rostamizadeh. Multi-class classification with maximum margin multiple kernel. In ICML-13, pages 46–54, 2013.

C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang. Structured prediction theory based on factor graph complexity. In Advances in Neural Information Processing Systems, pages 2514–2522, 2016.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2:265–292, 2002.

Y. Guermeur. Combining discriminant models with new multi-class svms. Pattern Analysis & Applications, 5(2):168–179, 2002.

S. I. Hill and A. Doucet. A framework for kernel-based multi-category classification. J. Artif. Intell. Res.(JAIR), 30:525–564, 2007.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics, pages 1–50, 2002.

V. Kuznetsov, M. Mohri, and U. Syed. Multi-class deep boosting. In Advances in Neural Information Processing Systems, pages 2501–2509, 2014.

Y. Lei, U. Dogan, A. Binder, and M. Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In Advances in Neural Information Processing Systems, pages 2026–2034, 2015.

Y. Lei, Ü. Dogan, D.-X. Zhou, and M. Kloft. Data-dependent generalization bounds for multi-class classification. IEEE Transactions on Information Theory, 65(5):2995–3021, 2019.

A. Maurer. A vector-contraction inequality for rademacher complexities. In International Conference on Algorithmic Learning Theory, pages 3–17. Springer, 2016.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. Foundations of machine learning. MIT press, 2012.

A. Rakhlin, K. Sridharan, and A. Tewari. Sequential complexities and uniform martingale laws of large numbers. Probability Theory and Related Fields, 161(1-2):111–153, 2015.

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In Advances in neural information processing systems, pages 2199–2207, 2010.

T. Zhang. Covering number bounds of certain regularized linear function classes. Journal of Machine Learning Research, 2: 527–550, 2002.

T. Zhang. Class-size independent generalization analsysis of some discriminative multi-category classification. In Advances in Neural Information Processing Systems, pages 1625–1632, 2004a.

T. Zhang. Statistical analysis of some multi-category large margin classification methods. The Journal of Machine Learning Research, 5:1225–1251, 2004b.