Multi-class SVMs

From Tighter Data-Dependent Generalization Bounds to Novel Algorithms

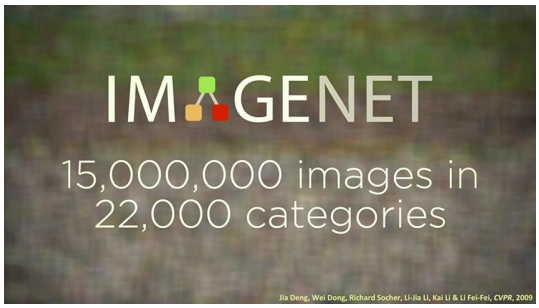
Marius Kloft

Joint work with **Yunwen Lei** (CU Hong Kong), **Urun Dogan** (Microsoft Research), and **Alexander Binder** (Singapore).

Extreme Classification

Many modern applications involve **a huge number of classes**.

- E.g., image annotation



(Deng, Dong, Socher, Li, Li, and Fei-Fei, 2009)

- Still growing datasets

Need for theory and algorithms for **extreme classification** (multi-class classification with huge amount of classes).

Discrepancy of **Theory** and Algorithms in Extreme Classification

- **Algorithms** for handling huge class sizes
  - (stochastic) dual coordinate ascent (Keerthi et al., 2008; Shalev-Shwartz and Zhang, (to appear)
- **Theory** not prepared for extreme classification
  - Data-dependent bounds scale at least **linearly** with the number of classes
    (Koltchinskii and Panchenko, 2002; Mohri et al., 2012; Kuznetsov et al., 2014)

Questions

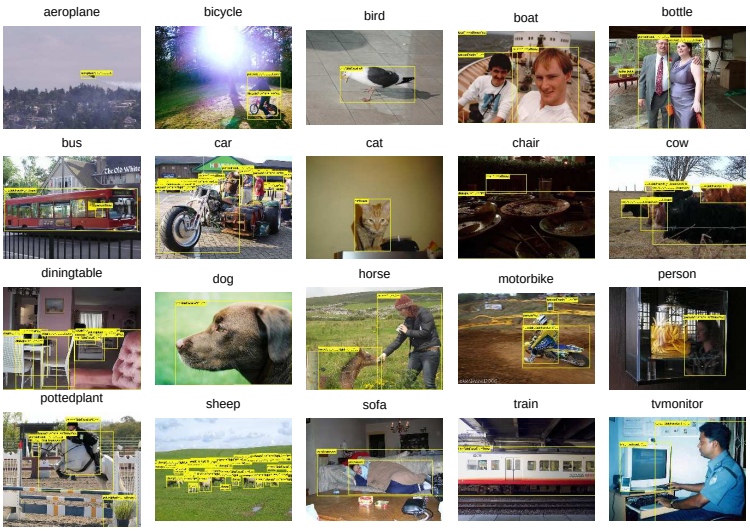
- Can we get bounds with **mild** dependence on #classes?
- What would we learn from such bounds?
  $\Rightarrow$ Novel algorithms?

Theory

Multi-class Classification

Given:

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



Formal Problem Setting

Aim:

- Define a hypothesis class H of functions $h = (h_1, \ldots, h_c)$
- Find an $h \in H$ that "predicts well" via

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

Multi-class SVMs:

- $h_y(x) = \langle \mathbf{w}_y, \phi(x) \rangle$
- Introduce notion of the **(multi-class) margin**

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

  - the larger the margin, the better

Want: large expected margin $\mathbb{E}\rho_h(X, Y)$.

Types of Generalization bounds for Multi-class Classification

Data-independent bounds

- based on covering numbers
  (Guermeur, 2002; Zhang, 2004a,b; Hill and Doucet, 2007)
- \- conservative
  - 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)
- \+ tighter
  - able to capture the real data
  - computable from the data

Rademacher & Gaussian Complexity

Definition

- Let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher variables (taking only values ± 1, with equal probability).
- The **Rademacher complexity** (RC) is defined as

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

Definition

- Let $g_1, \ldots, g_n \sim N(0,1)$.
- The **Gaussian complexity** (GC) is defined as

$$\mathfrak{G}(H) = \mathbb{E}_{\boldsymbol{g}}\Big[\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \boxed{g_i} h(z_i)\Big]$$

Interpretation: RC and GC reflect the **ability of the hypothesis class to correlate with random noise**.

Existing Data-Dependent Analysis

The key step is estimating $\mathfrak{R}(\{\rho_h : h \in H\})$ induced from the **margin operator** ρ_h and class H.

Existing bounds build on the structural result:

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

The correlation among class-wise components is ignored.

Best known dependence on the number of classes:

- **quadratic** dependence Koltchinskii and Panchenko (2002); Mohri et al. (2012); Cortes et al. (2013)
- **linear** dependence Kuznetsov et al. (2014)

Can we do better?

A New Structural Lemma on Gaussian Complexities

We consider Gaussian complexity.

- H is a vector-valued function class, $g_{11}, \ldots, g_{nc} \sim N(0, 1)$
- We show:

$$\mathfrak{G}\big(\{\max\{h_1, \ldots, h_c\} : h = (h_1, \ldots, h_c) \in H\}\big) \leq$$

$$\boxed{\frac{1}{n}\mathbb{E}_{\boldsymbol{g}} \sup_{h=(h_1,\ldots,h_c)\in H} \sum_{i=1}^{n}\sum_{j=1}^{c} g_{ij} h_j(x_i)}. \qquad (2)$$

Core idea: **Comparison inequality** on GPs: (Slepian, 1962)

$$\mathfrak{X}_h := \sum_{i=1}^{n} g_i \max\{h_1(x_i), \ldots, h_c(x_i)\}, \mathfrak{Y}_h := \sum_{i=1}^{n}\sum_{j=1}^{c} g_{ij} h_j(x_i), \forall h \in H.$$

$$\mathbb{E}[(\mathfrak{X}_\theta - \mathfrak{X}_{\bar{\theta}})^2] \leq \mathbb{E}[(\mathfrak{Y}_\theta - \mathfrak{Y}_{\bar{\theta}})^2] \Longrightarrow \mathbb{E}[\sup_{\theta\in\Theta} \mathfrak{X}_\theta] \leq \mathbb{E}[\sup_{\theta\in\Theta} \mathfrak{Y}_\theta].$$

Eq. (2) preserves the coupling among class-wise components!

Example on Comparison of the Structural Lemma

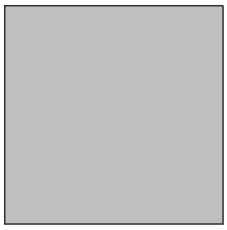
- Consider

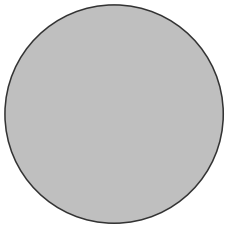
$$H := \{(x_1, x_2) \to (h_1, h_2)(x_1, x_2) = (w_1 x_1, w_2 x_2) : \|(w_1, w_2)\|_2 \leq 1\}$$

- For the function class $\{\max\{h_1, h_2\} : h = (h_1, h_2) \in H\}$,

$$\sup_{(h_1,h_2)\in H} \sum_{i=1}^n \sigma_i h_1(x_i) + \sup_{(h_1,h_2)\in H} \sum_{i=1}^n \sigma_i h_2(x_i)$$

$$\sup_{(h_1,h_2)\in H} \sum_{i=1}^n [g_{i1} h_1(x_i) + g_{i2} h_2(x_i)]$$





Preserving the coupling means supremum in a smaller space!

Estimating Multi-class Gaussian Complexity

- Consider a **vector-valued** function class defined by

$$H := \{h^{\mathbf{w}} = (\langle \mathbf{w}_1, \phi(x)\rangle, \ldots, \langle \mathbf{w}_c, \phi(x)\rangle) : f(\mathbf{w}) \leq \Lambda\},$$

where f is β-**strongly convex** w.r.t. $\|\cdot\|$

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

Theorem

$$\frac{1}{n}\mathbb{E}_g \sup_{h^{\mathbf{w}}\in H} \sum_{i=1}^n \sum_{j=1}^c g_{ij} h_j^{\mathbf{w}}(x_i) \leq \frac{1}{n}\sqrt{\frac{2\pi\Lambda}{\beta}\mathbb{E}_g \sum_{i=1}^n \left\| \left(g_{ij}\phi(x_i)\right)_{j=1}^c \right\|_*^2}, \quad (3)$$

where $\|\cdot\|_*$ is the **dual norm** of $\|\cdot\|$.

Features of the complexity bound

- Applies to a **general** function class defined through a strongly-convex regularizer f
- Class-wise components $h_1, \ldots, h_c$ are correlated through the term $\left\| \left(g_{ij} \phi(x_i) \right)_{j=1}^{c} \right\|_*^2$
- Consider class $H_{p,\Lambda} := \{ h^{\mathbf{w}} : \|\mathbf{w}\|_{2,p} \leq \Lambda \}$, $(\frac{1}{p} + \frac{1}{p^*} = 1)$; then:

$$\frac{1}{n} \mathbb{E}_{\boldsymbol{g}} \sup_{h^{\mathbf{w}} \in H_{p,\Lambda}} \sum_{i=1}^{n} \sum_{j=1}^{c} g_{ij} h_j^{\mathbf{w}}(x_i) \leq \frac{\Lambda}{n} \sqrt{\sum_{i=1}^{n} k(x_i, x_i)} \times \begin{cases} \sqrt{e} (4 \log c)^{1 + \frac{1}{2 \log c}}, & \text{if } p^* \geq 2 \log c, \\ (2p^*)^{1 + \frac{1}{p^*}} \boxed{c^{\frac{1}{p^*}}}, & \text{otherwise.} \end{cases}$$

The dependence is **sublinear** for $1 \leq p \leq 2$, and even **logarithmic** when p approaches to 1!

Algorithms

ℓ_p-norm Multi-class SVM

Motivated by the **mild dependence** on c as $p \to 1$, we consider

(ℓ_p-norm) Multi-class SVM, $1 \leq p \leq 2$

$$\begin{aligned} \min_{\mathbf{w}} \ & \frac{1}{2}\Big[\sum_{j=1}^{c} \|\mathbf{w}_j\|_2^p\Big]^{\frac{2}{p}} + C\sum_{i=1}^{n}(1-t_i)_+, \\ \text{s.t. } & t_i = \langle \mathbf{w}_{y_i}, \phi(x_i)\rangle - \max_{y:y\neq y_i} \langle \mathbf{w}_y, \phi(x_i)\rangle, \end{aligned} \tag{P}$$

Dual Problem

$$\begin{aligned} \sup_{\boldsymbol{\alpha}\in\mathbb{R}^{n\times c}} \ & -\frac{1}{2}\Big[\sum_{j=1}^{c}\Big\|\sum_{i=1}^{n}\alpha_{ij}\phi(x_i)\Big\|_2^{\frac{p}{p-1}}\Big]^{\frac{2(p-1)}{p}} + \sum_{i=1}^{n}\alpha_{iy_i} \\ \text{s.t. } & \boldsymbol{\alpha}_i \leq \mathbf{e}_{y_i}\cdot C \ \wedge \ \boldsymbol{\alpha}_i \cdot \mathbf{1} = 0, \quad \forall i = 1,\ldots,n. \end{aligned} \tag{D}$$

(D) is not quadratic if $p \neq 2$; how to optimize?

Equivalent Formulation

We introduce class weights $\beta_1, \ldots, \beta_c$ to get quadratic dual

$\min_{\boldsymbol{\beta}} \ \frac{1}{2}\sum_{j=1}^{c}\frac{\|\mathbf{w}_j\|^2}{\beta_j} + \lambda\|\boldsymbol{\beta}\|_p^p$ has optimum for $\beta_j \propto \sqrt[p+1]{\|\mathbf{w}_j\|^2}$.

Equivalent Problem

$$\begin{aligned} \min_{\mathbf{w},\,\boldsymbol{\beta}} \ & \sum_{j=1}^{c}\frac{\|\mathbf{w}_j\|_2^2}{2\beta_j} + C\sum_{i=1}^{n}(1-t_i)_+ \\ \text{s.t. } & t_i \leq \langle \mathbf{w}_{y_i}, \phi(x_i)\rangle - \langle \mathbf{w}_y, \phi(x_i)\rangle, \quad y \neq y_i, i = 1,\ldots,n, \\ & \|\boldsymbol{\beta}\|_{\bar{p}} \leq 1, \bar{p} = p(2-p)^{-1}, \beta_j \geq 0. \end{aligned} \tag{E}$$

Alternating optimization w.r.t. $\boldsymbol{\beta}$ and to $\mathbf{w}$

Empirical Results

Description of datasets used in the experiments:

Dataset	# Classes	# Training Examples	# Test Examples	# Attributes
Sector	105	6,412	3,207	55,197
News 20	20	15,935	3,993	62,060
Rcv1	53	15,564	518,571	47,236
Birds 50	200	9,958	1,830	4,096
Caltech 256	256	12,800	16,980	4,096

Empirical Results:

Method / Dataset	Sector	News 20	Rcv1	Birds 50	Caltech 256
ℓ_p-norm MC-SVM	94.2 ± 0.3	86.2 ± 0.1	85.7 ± 0.7	27.9 ± 0.2	56.0 ± 1.2
Crammer & Singer	93.9 ± 0.3	85.1 ± 0.3	85.2 ± 0.3	26.3 ± 0.3	55.0 ± 1.1

Proposed ℓ_p-norm MC-SVM consistently better on benchmark datasets.

Future Directions

Theory: A data-dependent bound **independent** of the class size?

$\Rightarrow$ Need more powerful structural result on Gaussian complexity for functions induced by **maximum operator**.

- Might be worth to look into **ℓ_∞-norm covering numbers**.

Algorithms: New models & efficient solvers

- **Novel models** motivated by theory
  - top-k MC-SVM (Lapin et al., 2015), nuclear norm regularization, ...
- **Scalable** algorithms
- Analyze $p > 2$ regime
- Extensions to **multi-label** learning

References I

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

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In **Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on**, pages 248–255. IEEE, 2009.

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.

S. S. Keerthi, S. Sundararajan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. A sequential dual method for large scale multi-class linear svms. In **14th ACM SIGKDD**, pages 408–416. ACM, 2008.

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.

M. Lapin, M. Hein, and B. Schiele. Top-k multiclass SVM. **CoRR**, abs/1511.06683, 2015. URL `http://arxiv.org/abs/1511.06683`.

M. Ledoux and M. Talagrand. **Probability in Banach Spaces: isoperimetry and processes**, volume 23. Springer, Berlin, 1991.

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

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. **Mathematical Programming SERIES A and B**, 5, (to appear).

D. Slepian. The one-sided barrier problem for gaussian noise. **Bell System Technical Journal**, 41(2):463–501, 1962.

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.