Optimality Implies Kernel Sum Classifiers are Statistically Efficient

Raphael A. Meyer[1] Jean Honorio[1]

[1]Dept. of Computer Science, Purdue University

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

- Define a set of possible estimators to map inputs to labels

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

- Define a set of possible estimators to map inputs to labels

$$\mathcal{F} \subseteq \{f \colon \mathcal{X} \mapsto \{-1, 1\}\}$$

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

- Define a set of possible estimators to map inputs to labels

$$\mathcal{F} \subseteq \{f : \mathcal{X} \mapsto \{-1, 1\}\}$$

- Prove the empirical error is close to the expected error

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

- Define a set of possible estimators to map inputs to labels

$$\mathcal{F} \subseteq \{f \colon \mathcal{X} \mapsto \{-1, 1\}\}$$

- Prove the empirical error is close to the expected error
    - Rademacher Complexity, PAC Bayes, etc.

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

- Define a set of possible estimators to map inputs to labels

$$\mathcal{F} \subseteq \{f : \mathcal{X} \mapsto \{-1, 1\}\}$$

- Prove the empirical error is close to the expected error
    - Rademacher Complexity, PAC Bayes, etc.
    - Prove this for all $f \in \mathcal{F}$

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

- Define a set of possible estimators to map inputs to labels

$$\mathcal{F} \subseteq \{f : \mathcal{X} \mapsto \{-1, 1\}\}$$

- Prove the empirical error is close to the expected error
  - Rademacher Complexity, PAC Bayes, etc.
  - Prove this for all $f \in \mathcal{F}$

$$\mathop{\mathbb{E}}_{\mathbf{x},y}[\ell(f(\mathbf{x}), y)] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) + \varepsilon$$

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

- Define a set of possible estimators to map inputs to labels

$$\mathcal{F} \subseteq \{f \colon \mathcal{X} \mapsto \{-1, 1\}\}$$

- Prove the empirical error is close to the expected error
  - Rademacher Complexity, PAC Bayes, etc.
  - Prove this for all $f \in \mathcal{F}$

$$\mathop{\mathbb{E}}_{\mathbf{x},y}[\ell(f(\mathbf{x}), y)] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) + \varepsilon$$

Generalization Error Proofs

- Assume we have n i.i.d. labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\mathbf{x} \in \mathcal{X}, \qquad y \in \{-1, 1\}$$

- Define a set of possible estimators to map inputs to labels

$$\mathcal{F} \subseteq \{f : \mathcal{X} \mapsto \{-1, 1\}\}$$

- Prove the empirical error is close to the expected error
  - Rademacher Complexity, PAC Bayes, etc.
  - Prove this for all $f \in \mathcal{F}$
  - This includes low-accuracy estimators f

Optimization Theory

- Optimization is often used to find estimators in ML

Optimization Theory

- Optimization is often used to find estimators in ML
  - Regression, Kernel SVM, etc.

Optimization Theory

- Optimization is often used to find estimators in ML
  - Regression, Kernel SVM, etc.
- These tools are used in practice

Optimization Theory

- Optimization is often used to find estimators in ML
  - Regression, Kernel SVM, etc.
- These tools are used in practice
- Strong theoretical guarantees in polynomial time

Optimization Theory

- Optimization is often used to find estimators in ML
  - Regression, Kernel SVM, etc.
- These tools are used in practice
- Strong theoretical guarantees in polynomial time
  - Karush-Kuhn-Tucker (KKT) Conditions

Optimization Theory

- Optimization is often used to find estimators in ML
  - Regression, Kernel SVM, etc.
- These tools are used in practice
- Strong theoretical guarantees in polynomial time
  - Karush-Kuhn-Tucker (KKT) Conditions
- How can considering optimal estimators help us understand statistical efficiency?

Given:

Multiple Kernel Learning & Classification

Given:

- Dataset with n samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

Multiple Kernel Learning & Classification

Given:

- Dataset with n samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- m kernels $k_1, \ldots, k_m$

Multiple Kernel Learning & Classification

Given:

- Dataset with n samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- m kernels $k_1, \ldots, k_m$

Learn:

Multiple Kernel Learning & Classification

Given:

- Dataset with n samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- m kernels $k_1, \ldots, k_m$

Learn:

- Linear combination $\boldsymbol{\theta}$ of kernels

Multiple Kernel Learning & Classification

Given:

- Dataset with n samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- m kernels $k_1, \ldots, k_m$

Learn:

- Linear combination $\boldsymbol{\theta}$ of kernels s.t. the resulting kernel

$$k_{\Sigma}(\cdot, \cdot) := \sum_{t=1}^{m} \theta_t k_t(\cdot, \cdot)$$

Classifies the dataset well

Multiple Kernel Learning & Classification

Given:

- Dataset with n samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- m kernels $k_1, \ldots, k_m$

Learn:

- Linear combination $\boldsymbol{\theta}$ of kernels s.t. the resulting kernel

$$k_\Sigma(\cdot, \cdot) := \sum_{t=1}^{m} \theta_t k_t(\cdot, \cdot)$$

  Classifies the dataset well

- Constraints on $\boldsymbol{\theta}$ vary among papers

Multiple Kernel Learning & Classification

Given:

- Dataset with n samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- m kernels $k_1, \ldots, k_m$

Learn:

- Linear combination $\boldsymbol{\theta}$ of kernels s.t. the resulting kernel

$$k_\Sigma(\cdot,\cdot) := \sum_{t=1}^{m} \theta_t k_t(\cdot,\cdot)$$

  Classifies the dataset well

- Constraints on $\boldsymbol{\theta}$ vary among papers
  - $\boldsymbol{\theta}$ may be convex combination, 0/1 vector, etc.

Prior Work: Generalization Error

- Estimators are uniquely identified by $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{\theta}$

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i k_{\Sigma}(\mathbf{x}_i, \mathbf{x})$$

Prior Work: Generalization Error

- Estimators are uniquely identified by $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{\theta}$

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i k_\Sigma(\mathbf{x}_i, \mathbf{x})$$

Theorem [CMR10]

For all kernels, vectors $\boldsymbol{\alpha}$, and convex combinations $\boldsymbol{\theta}$ where

Prior Work: Generalization Error

- Estimators are uniquely identified by $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{\theta}$

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i k_{\Sigma}(\mathbf{x}_i, \mathbf{x})$$

Theorem [CMR10]

For all kernels, vectors $\boldsymbol{\alpha}$, and convex combinations $\boldsymbol{\theta}$ where

- $k_t(\mathbf{x}_i, \mathbf{x}_i) \leq R^2$ for all $\mathbf{x}_i$ and k_t

Prior Work: Generalization Error

- Estimators are uniquely identified by $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{\theta}$

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i k_\Sigma(\mathbf{x}_i, \mathbf{x})$$

Theorem [CMR10]

For all kernels, vectors $\boldsymbol{\alpha}$, and convex combinations $\boldsymbol{\theta}$ where

- $k_t(\mathbf{x}_i, \mathbf{x}_i) \leq R^2$ for all $\mathbf{x}_i$ and k_t
- $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha} \leq C^2$ for kernel matrix $\tilde{\boldsymbol{K}}_\Sigma = \sum_{t=1}^{m} \theta_t \tilde{\boldsymbol{K}}_t$

Prior Work: Generalization Error

- Estimators are uniquely identified by $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{\theta}$

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i k_{\Sigma}(\mathbf{x}_i, \mathbf{x})$$

Theorem [CMR10]

For all kernels, vectors $\boldsymbol{\alpha}$, and convex combinations $\boldsymbol{\theta}$ where

- $k_t(\mathbf{x}_i, \mathbf{x}_i) \leq R^2$ for all $\mathbf{x}_i$ and k_t
- $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_{\Sigma} \boldsymbol{\alpha} \leq C^2$ for kernel matrix $\tilde{\boldsymbol{K}}_{\Sigma} = \sum_{t=1}^{m} \theta_t \tilde{\boldsymbol{K}}_t$

We have

$$\mathop{\mathbb{E}}_{\mathbf{x},y}[\ell(f(\mathbf{x}; \boldsymbol{\alpha}), y)] \leq \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i; \boldsymbol{\alpha}), y_i) + O\left(\frac{CR}{\sqrt{n}} \sqrt{\ln m}\right)$$

Prior Work: Generalization Error

- Estimators are uniquely identified by $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{\theta}$

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i k_\Sigma(\mathbf{x}_i, \mathbf{x})$$

Theorem [CMR10]

For all kernels, vectors $\boldsymbol{\alpha}$, and convex combinations $\boldsymbol{\theta}$ where

- $k_t(\mathbf{x}_i, \mathbf{x}_i) \leq R^2$ for all $\mathbf{x}_i$ and k_t
- $\boldsymbol{\alpha}^\mathsf{T} \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha} \leq C^2$ for kernel matrix $\tilde{\boldsymbol{K}}_\Sigma = \sum_{t=1}^{m} \theta_t \tilde{\boldsymbol{K}}_t$

We have

$$\mathbb{E}_{\mathbf{x},y}[\ell(f(\mathbf{x}; \boldsymbol{\alpha}), y)] \leq \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i; \boldsymbol{\alpha}), y_i) + O\left(\frac{CR}{\sqrt{n}}\sqrt{\ln m}\right)$$

How can we understand $\alpha^\intercal \tilde{\boldsymbol{K}}_\Sigma \alpha$?

How does $\alpha^\intercal \tilde{\boldsymbol{K}}_\Sigma \alpha$ depend on

- The number of kernels?

How can we understand $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}$?

How does $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}$ depend on

- The number of kernels?
- The conditioning of the kernels?

How can we understand $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}$?

How does $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}$ depend on

- The number of kernels?
- The conditioning of the kernels?

These questions are ill-posed:

How can we understand $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}$?

How does $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}$ depend on

- The number of kernels?
- The conditioning of the kernels?

These questions are ill-posed:

- All $\boldsymbol{\alpha} \in \mathbb{R}^n$ define a feasible estimator

How can we understand $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}$?

How does $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}$ depend on

- The number of kernels?
- The conditioning of the kernels?

These questions are ill-posed:

- All $\boldsymbol{\alpha} \in \mathbb{R}^n$ define a feasible estimator
- So there always exists a feasible estimator with large $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}$

How can we understand $\alpha^\top \tilde{K}_\Sigma \alpha$?

How does $\alpha^\top \tilde{K}_\Sigma \alpha$ depend on

- The number of kernels?
- The conditioning of the kernels?

These questions are ill-posed:

- All $\alpha \in \mathbb{R}^n$ define a feasible estimator
- So there always exists a feasible estimator with large $\alpha^\top \tilde{K} \alpha$
- But Support Vector Machines pick α in practice

How can we understand $\alpha^\intercal \tilde{\boldsymbol{K}}_\Sigma \alpha$?

If $\boldsymbol{\alpha}_\Sigma$ solves the SVM problem with $\tilde{\boldsymbol{K}}_\Sigma$,
How does $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma$ depend on

- The number of kernels?
- The conditioning of the kernels?

These questions are ill-posed:

- All $\boldsymbol{\alpha} \in \mathbb{R}^n$ define a feasible estimator
- So there always exists a feasible estimator with large $\alpha^\intercal \tilde{\boldsymbol{K}} \alpha$
- But Support Vector Machines pick α in practice

Our Approach

Given: $\tilde{\boldsymbol{K}}_1 \quad \tilde{\boldsymbol{K}}_2 \quad \cdots \quad \tilde{\boldsymbol{K}}_m \quad \tilde{\boldsymbol{K}}_\Sigma$

Our Approach

Given: $\tilde{\boldsymbol{K}}_1 \quad \tilde{\boldsymbol{K}}_2 \quad \cdots \quad \tilde{\boldsymbol{K}}_m \quad \tilde{\boldsymbol{K}}_\Sigma$

Dual SVM

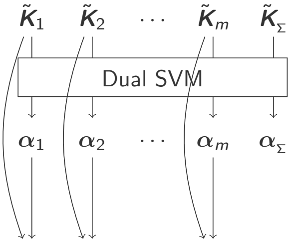
Optimize: $\boldsymbol{\alpha}_1 \quad \boldsymbol{\alpha}_2 \quad \cdots \quad \boldsymbol{\alpha}_m \quad \boldsymbol{\alpha}_\Sigma$

Our Approach

Given:

Optimize:

Assume:



For all $t = 1, 2, \ldots, m$,
Assume $\boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \leq B^2$

Our Approach

Given: $\tilde{\boldsymbol{K}}_1 \quad \tilde{\boldsymbol{K}}_2 \quad \cdots \quad \tilde{\boldsymbol{K}}_m \quad \tilde{\boldsymbol{K}}_\Sigma$

Dual SVM

Optimize: $\boldsymbol{\alpha}_1 \quad \boldsymbol{\alpha}_2 \quad \cdots \quad \boldsymbol{\alpha}_m \quad \boldsymbol{\alpha}_\Sigma$

Assume: For all $t = 1, 2, \ldots, m$,
Assume $\boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \leq B^2$

KKT Conditions

Then $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq 3m^{-0.58} B^2$

Our Approach

Given: $\tilde{K}_1 \quad \tilde{K}_2 \quad \cdots \quad \tilde{K}_m \quad \tilde{K}_\Sigma$

Dual SVM

Optimize: $\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_m \quad \alpha_\Sigma$

Assume: For all $t = 1, 2, \ldots, m$,

Assume $\alpha_t^\intercal \tilde{K}_t \alpha_t \leq B^2$

KKT Conditions

Then $\alpha_\Sigma^\intercal \tilde{K}_\Sigma \alpha_\Sigma \leq 3m^{-0.58} B^2$

Rademacher Complexity

Then: Estimator $y(\mathbf{x}; \alpha_\Sigma)$ generalizes well

$O(\frac{BRm^{0.208}\sqrt{\ln m}}{\sqrt{n}})$

Conclusion

- Leverage Optimization Theory to ask and answer well-posed questions about the statistics of practical estimators.

Conclusion

- Leverage Optimization Theory to ask and answer well-posed questions about the statistics of practical estimators.
- Consider $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}$ in Multiple Kernel Learning as a specific case

Conclusion

- Leverage Optimization Theory to ask and answer well-posed questions about the statistics of practical estimators.
- Consider $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}$ in Multiple Kernel Learning as a specific case
- Several possible applications of this idea beyond kernels

THANK
YOU

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh.
Generalization Bounds for Learning Kernels.
In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 247–254. Omnipress, 2010.

Raphael Meyer and Jean Honorio.
Optimality Implies Kernel Sum Classifiers are Statistically Efficient.
In *International Conference on Machine Learning*, pages 4566–4574, 2019.