Generalization Bounds for Learning Kernels

Corinna Cortes, Mehryar Mohri, and Afshin Rostami
Google Research and Courant Institute
mohri@cs.nyu.edu

Learning Kernels

- Idea:
  - standard: user commits to a single specific kernel.
  - learning kernels: user specifies a family of kernels.
  - let learning algorithm use sample to select both an appropriate kernel and a hypothesis.
- Questions:
  - what is the price to pay for relaxing the requirement from the user?
  - how does the choice of the family affect generalization?

Hypothesis Sets

- Based on non-negative combinations of p base kernels:

$$H_p^q = \left\{ h \in \mathbb{H}_K : K = \sum_{k=1}^{p} \mu_k K_k, \|h\|_{\mathbb{H}_K} \leq 1, \boldsymbol{\mu} \in \Delta_q \right\}$$

$$\Delta_q = \{ \boldsymbol{\mu} : \boldsymbol{\mu} \geq 0, \sum_{k=1}^{p} \mu_k^q = 1 \}.$$

  - Most previous learning kernel studies, e.g., (Bousquet and Herrmann, 2003; Crammer et al., 2003; Lanckriet et al., 2004; Sonnenburg et al, 2005; Argyriou at al., 2005; Cortes et al., 2008-2009).

Single Kernel Margin Bound

- **Theorem** (Koltchinskii and Panchenko, 2002): fix $\rho > 0$. Assume that $K(x,x) \leq R^2$ for all x, then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H_1^1$,

$$R(h) \leq \widehat{R}_\rho(h) + 2\sqrt{\frac{R^2/\rho^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Rademacher Complexity

- Empirical Rademacher complexity of H: for a sample $S = (x_1, \ldots, x_m)$,

$$\widehat{\mathfrak{R}}_S(H) = \mathop{\mathrm{E}}_{\sigma}\left[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i)\right],$$

  where σ_is are independent uniform random variables taking values in $\{-1, +1\}$.

- Rademacher complexity of H:

$$\mathfrak{R}_m(H) = \mathop{\mathrm{E}}_{S \sim D^m}[\widehat{\mathfrak{R}}_S(H)].$$

Single Kernel Margin Bound

- Lemma: for any sample S of size m,

$$\widehat{\mathfrak{R}}_S(H_1^1) \leq \frac{\sqrt{\mathrm{Tr}[\mathbf{K}]}}{m}.$$

- Theorem (Koltchinskii and Panchenko, 2002): fix $\rho > 0$. Assume that $K(x,x) \leq R^2$ for all x, then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H_1^1$,

$$R(h) \leq \widehat{R}_\rho(h) + 2\sqrt{\frac{R^2/\rho^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

This Talk

- Previous work
- Novel guarantees

Early Learning Kernel Bounds

(Bousquet and Herrmann 2003; Lanckriet et al., 2004)

- For any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H_p^1$,

$$R(h) \leq \widehat{R}_\rho(h) + \frac{1}{\sqrt{m}} \left[\sqrt{\frac{\max_{k=1}^p \operatorname{Tr}(\mathbf{K}_k) \max_{k=1}^p \frac{\|\mathbf{K}_k\|}{\operatorname{Tr}(\mathbf{K}_k)}}{\rho^2} + 4 + \sqrt{2 \log \frac{1}{\delta}} \right].$$

  - but, bound always greater than one (Srebro and Ben-David, 2006)!
  - other bound of (Lanckriet et al., 2004) for linear combination case also always greater than one!

Multiplicative Learning Bound

(Lanckriet et al., 2004)

- Assume that for all $k \in [1, p]$, $K_k(x, x) \leq R^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H_p^1$,

$$R(h) \leq \widehat{R}_\rho(h) + O\left(\sqrt{\frac{p\, R^2/\rho^2}{m}}\right).$$

  - bound multiplicative in p (number of kernels).

Additive Learning Bound

(Srebro and Ben-David, 2006)

- Assume that for all $k \in [1, p]$, $K_k(x, x) \leq R^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H_p^1$,

$$R(h) \leq \widehat{R}_\rho(h) + \sqrt{8 \frac{2 + p \log \frac{128 e m^3 R^2}{\rho^2 p} + 256 \frac{R^2}{\rho^2} \log \frac{\rho e m}{8R} \log \frac{128 m R^2}{\rho^2} + \log(1/\delta)}{m}}.$$

  - bound additive in p (modulo log terms).
  - not informative for $p > m$.
  - based on pseudo-dimension of kernel family.
  - similar guarantees for other families.

This Talk

- Previous work
- Novel guarantees

New Data-Dependent Bound

- Theorem: for any sample S of size m, and positive integer r,

$$\widehat{\mathfrak{R}}_S(H_p^1) \leq \frac{\sqrt{\frac{23}{22} r \|\mathbf{u}\|_r}}{m} ,$$

with $\mathbf{u} = (\mathrm{Tr}[\mathbf{K}_1], \ldots, \mathrm{Tr}[\mathbf{K}_p])^\top$.

  - similarity with single kernel bound.

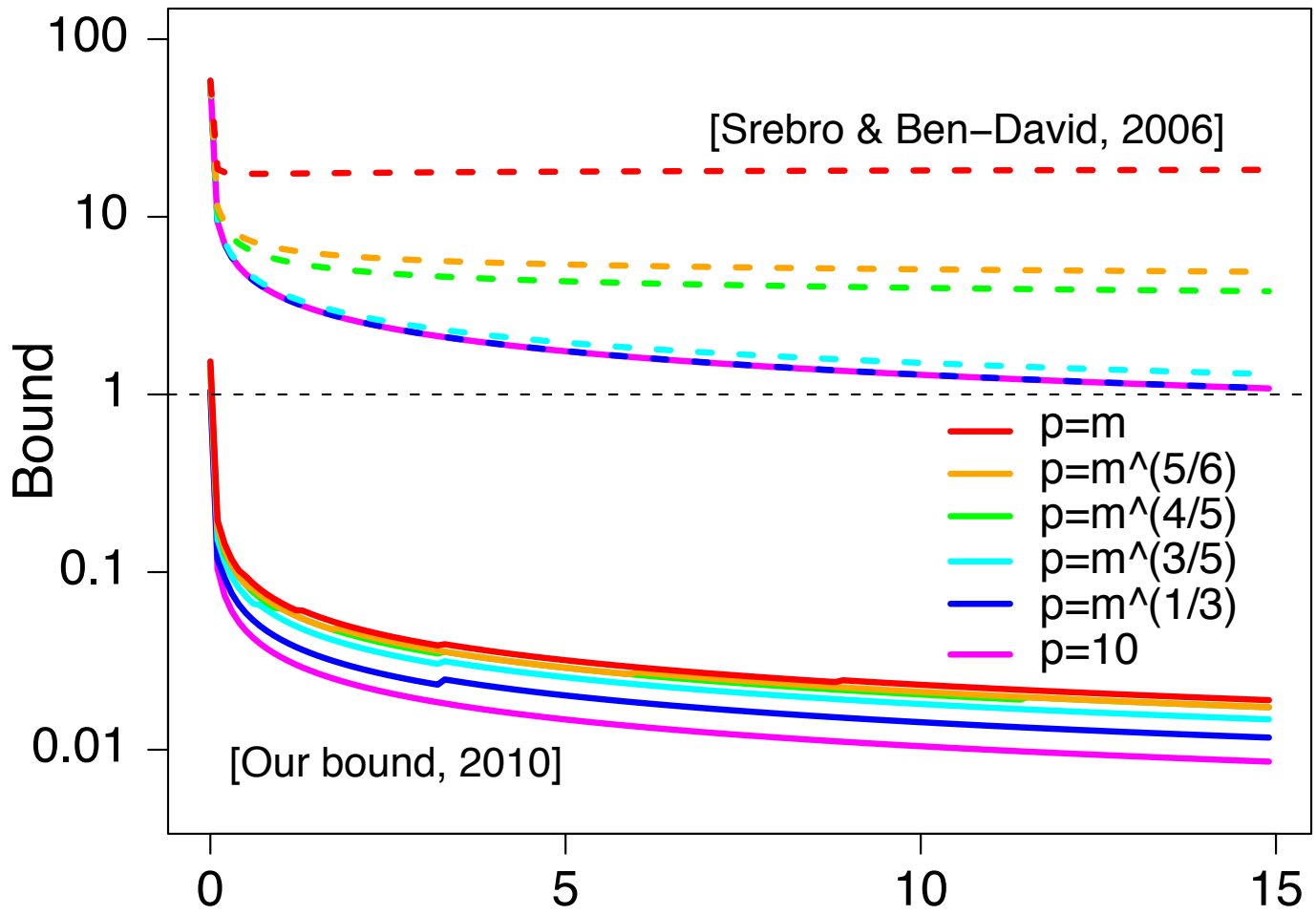
New Learning Bound - L1

- **Theorem**: assume that for all $k \in [1, p]$, $K_k(x, x) \leq R^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H_p^1$,

$$R(h) \leq \widehat{R}_\rho(h) + 2\sqrt{\frac{\frac{23}{22} e \lceil \log p \rceil R^2/\rho^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

  - very weak dependency on p, no extra log terms.
  - analysis based on Rademacher complexity.
  - bound valid for $p \gg m$.
  - similar bound, less favorable const. (Kakade et al., 2010).

Comparison



$\rho/R = .2$

Lower Bound

- Tight bound:
  - dependency $\sqrt{\log p}$ cannot be improved.
  - argument based on VC dimension or example.
- Observations: case $\mathcal{X}=\{\text{-}1, +1\}^p$.
  - canonical projection kernels $K_k(\mathbf{x}, \mathbf{x}') = x_k x'_k$.
  - H_p^1 contains $J_p = \{\mathbf{x} \mapsto s x_k : k \in [1, p], s \in \{\text{-}1, +1\}\}$.
  - $\mathrm{VCdim}(J_p) = \Omega(\log p)$.
  - for $\rho = 1$ and $h \in J_p$, $\widehat{R}_\rho(h) = \widehat{R}(h)$.
  - VC lower bound: $\Omega\left(\sqrt{\mathrm{VCdim}(J^p)/m}\right)$.

Key Lemma

- Lemma: Let $\mathbf{K}$ be a kernel matrix for a finite sample. Then, for any integer r ,

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[(\boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\sigma})^r\right] \leq \left(\frac{23}{22}\, r\, \mathrm{Tr}[\mathbf{K}]\right)^r.$$

  - proof based on combinatorial argument.

New Learning Bound - Lq

- **Theorem**: let $q, r \geq 1$ with $\frac{1}{q} + \frac{1}{r} = 1$ and r integer. Assume that for all $k \in [1, p]$, $K_k(x, x) \leq R^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H_p^q$,

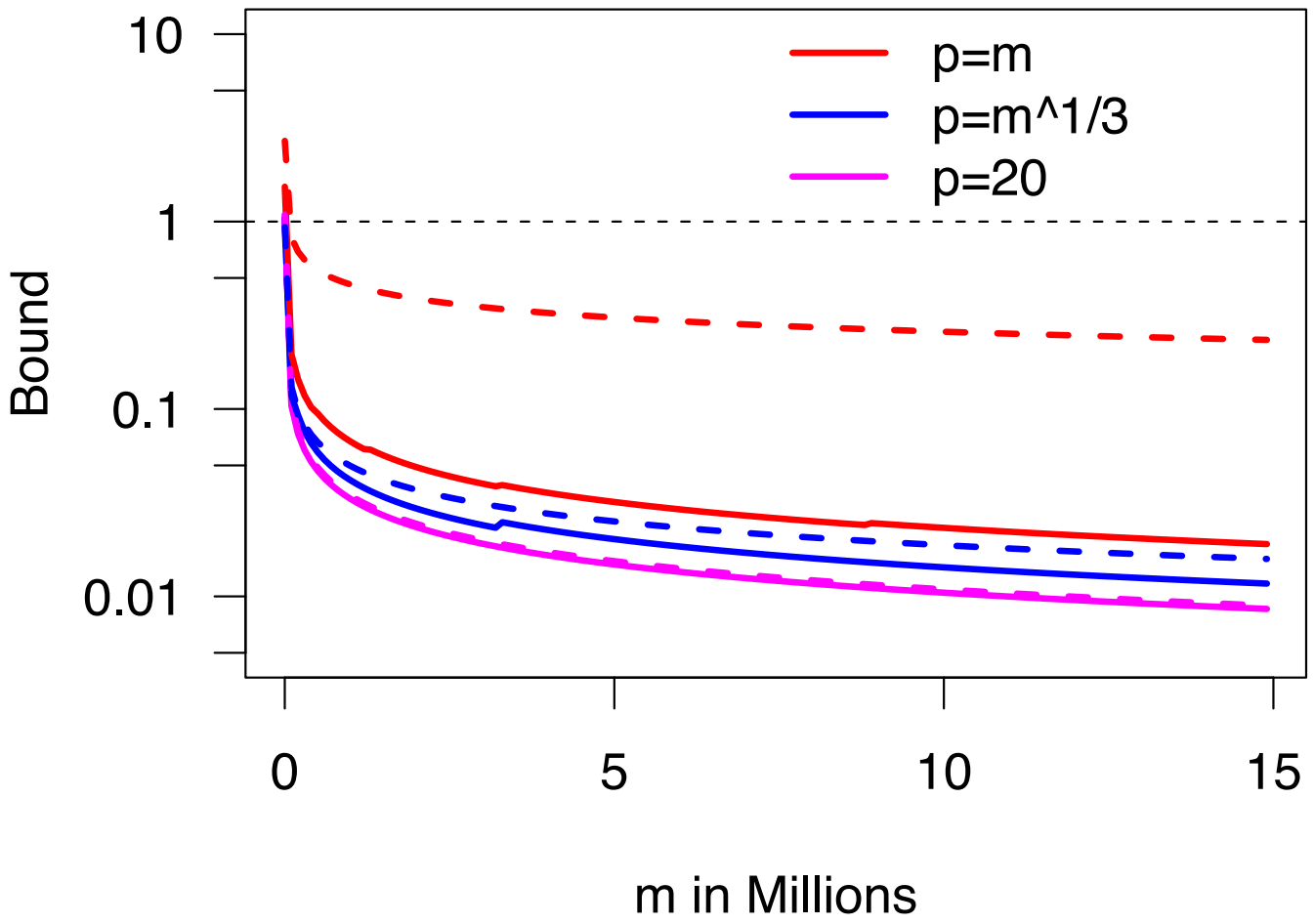
$$R(h) \leq \widehat{R}_\rho(h) + 2p^{\frac{1}{2r}} \sqrt{\frac{\frac{23}{22} r R^2 / \rho^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

  - mild dependency on p.
  - analysis based on Rademacher complexity.

Lower Bound

- Tight bound:
  - dependency $p^{\frac{1}{2r}}$ cannot be improved.
  - in particular $p^{\frac{1}{4}}$ tight for L_2 regularization.
- Observations: equal kernels.
  - $\sum_{k=1}^{p} \mu_k K_k = \big(\sum_{k=1}^{p} \mu_k\big) K_1$.
  - thus, $\|h\|^2_{\mathbb{H}_{K_1}} = (\sum_{k=1}^{p} \mu_k) \|h\|^2_{\mathbb{H}_K}$ for $\sum_{k=1}^{p} \mu_k \neq 0$.
  - $\sum_{k=1}^{p} \mu_k \leq p^{\frac{1}{r}} \|\boldsymbol{\mu}\|_q = p^{\frac{1}{r}}$ (Hölder's inequality).
  - H_p^q coincides with $\{h \in \mathbb{H}_{K_1} \colon \|h\|_{\mathbb{H}_{K_1}} \leq p^{\frac{1}{2r}}\}$.

Comparison L1 vs L2



Conclusion

- Theory: tight generalization bounds for learning kernels with L_1 or L_q regularization.
  - mild dependency on p.
  - similar proof and analysis for other regularizations.
- Applications: can learning kernels improve performance? (Cortes, ICML 2009).
  - results suggest using large number of kernels.
  - recent results show significant improvements (Cortes, MM, Rostamizadeh, ICML 2010).