Learning Kernels -Tutorial

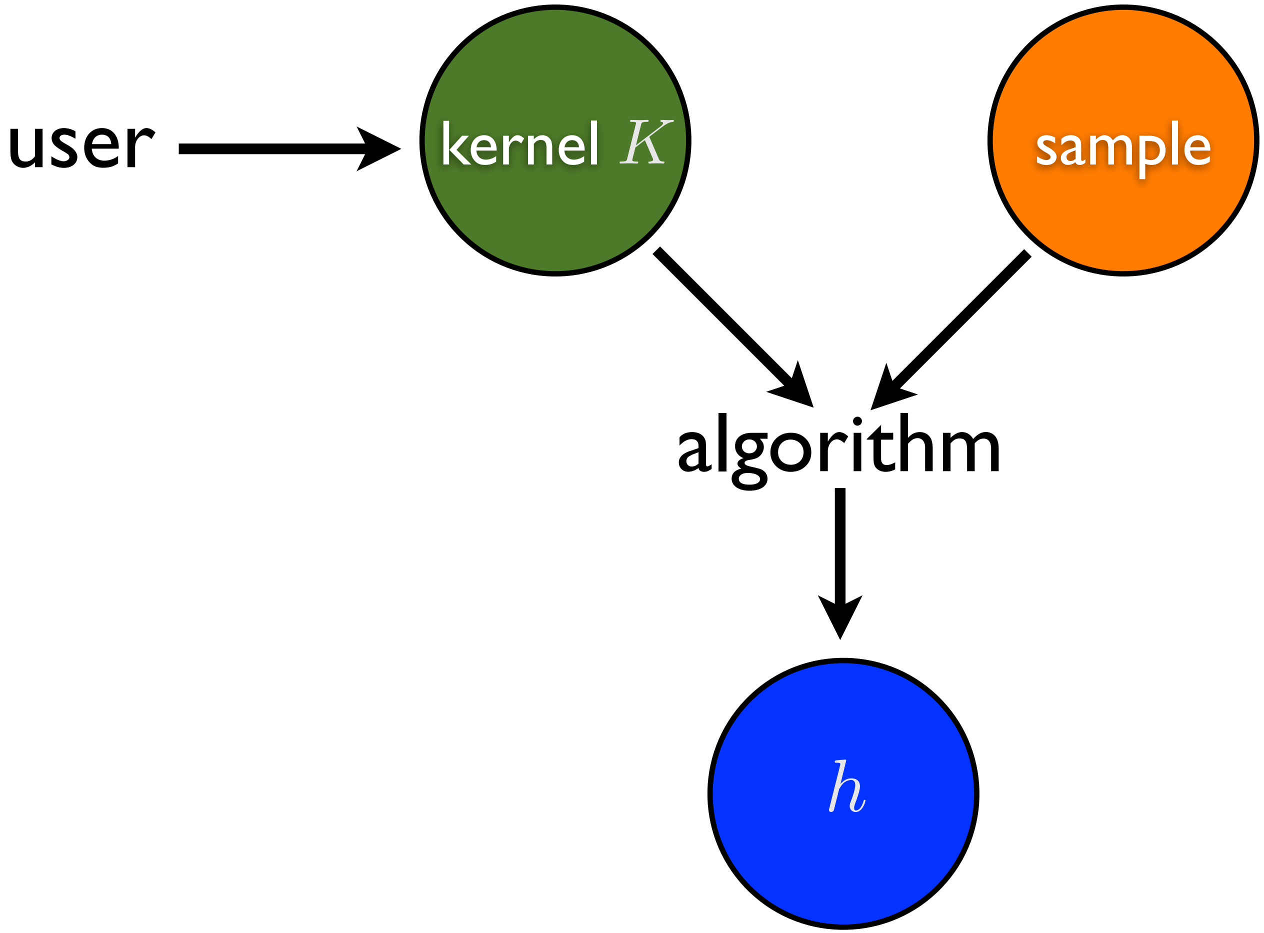
Part III: Theoretical Guarantees.

Corinna Cortes
Google Research
corinna@google.com

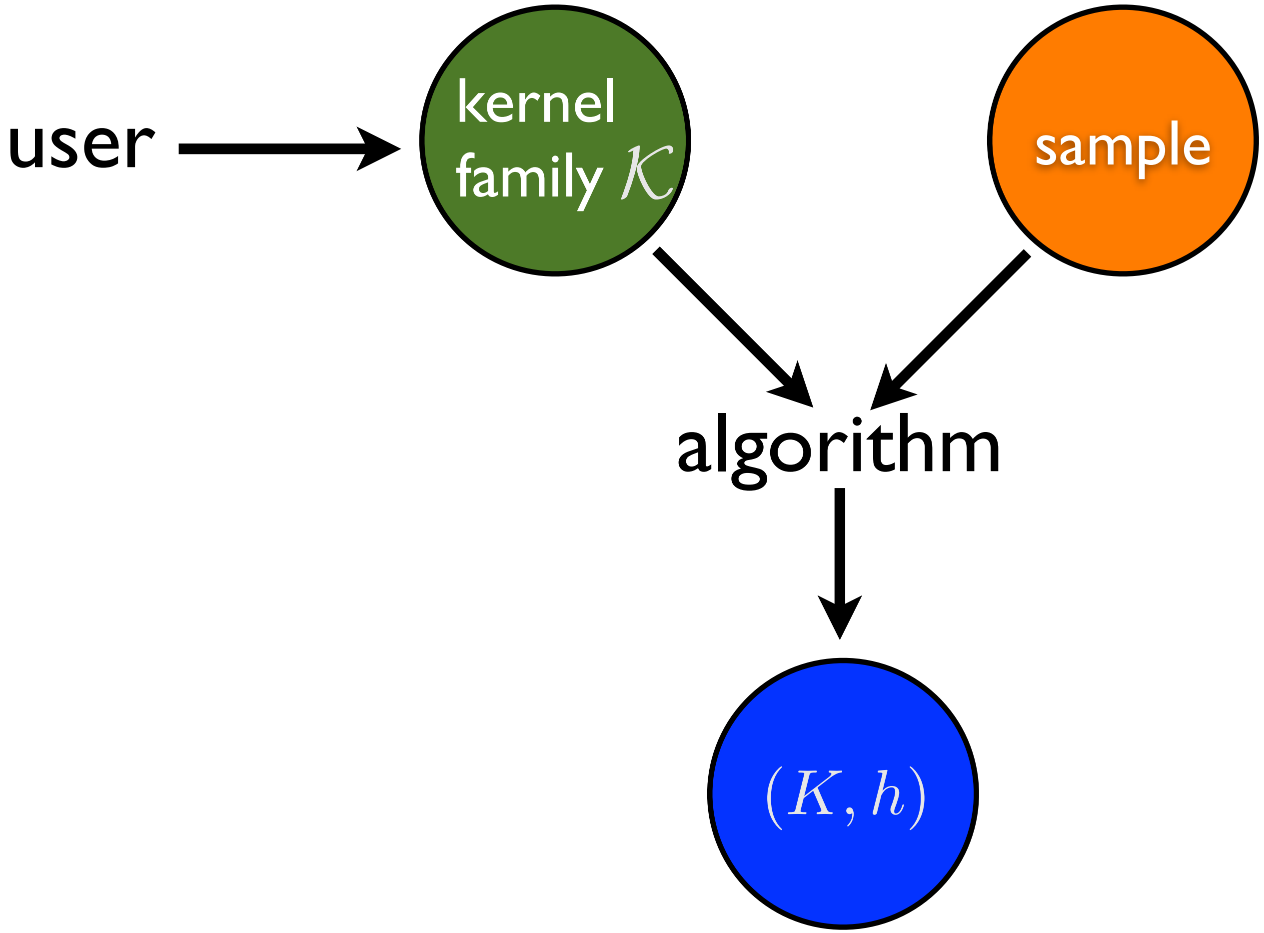
Mehryar Mohri
Courant Institute &
Google Research
mohri@cims.nyu.edu

Afshin Rostami
UC Berkeley
arostami@eecs.
berkeley.edu

Standard Learning with Kernels



Learning Kernel Framework



Learning Kernels

- Theoretical questions:
  - what is the price to pay for relaxing the requirement from the user to specify a kernel?
  - how does the choice of the kernel family affect generalization?

Part III

- Non-negative combinations.
- General case.

Kernel Families

- Most frequently used kernel families, $q \geq 1$,

$$\mathcal{K}_q = \left\{ \sum_{k=1}^{p} \mu_k K_k : \boldsymbol{\mu} \in \Delta_q \right\}$$

with $\Delta_q = \left\{ \boldsymbol{\mu} : \boldsymbol{\mu} \geq 0, \|\boldsymbol{\mu}\|_q = 1 \right\}$.

- Hypothesis sets:

$$H_q = \left\{ h \in \mathbb{H}_K : K \in \mathcal{K}_q, \|h\|_{\mathbb{H}_K} \leq 1 \right\}.$$

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

- **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 $h \in H_1$,

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

Early Learning Kernel Bounds

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

- For any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H_1$,

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

  - 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 $h \in H_1$,

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

  - 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 $h \in H_1$,

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

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

New Data-Dependent Bound

(CC, MM, and AR, 2010)

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

$$\widehat{\mathfrak{R}}_S(H_1) \leq \frac{\sqrt{\frac{23}{22} r \|\mathbf{u}\|_r}}{m},$$

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

  - similarity with single kernel bound.
  - can be used directly to derive an algorithm.

New Data-Dependent Bound

- **Proof**: Let $q, r \geq 1$ with $\frac{1}{q} + \frac{1}{r} = 1$.

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H_q) &= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sup_{h \in H_q} \sum_{i=1}^m \sigma_i h(x_i) \Big] \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \Delta_q, \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} \sum_{i,j=1}^m \sigma_i \alpha_j K_{\boldsymbol{\mu}}(x_i, x_j) \Big] \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \Delta_q, \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} \boldsymbol{\sigma}^\top \mathbf{K}_{\boldsymbol{\mu}} \boldsymbol{\alpha} \Big] = \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \Delta_q, \|\boldsymbol{\alpha}\|_{\mathbf{K}^{1/2}} \leq 1} \langle \boldsymbol{\sigma}, \boldsymbol{\alpha} \rangle_{\mathbf{K}^{1/2}} \Big] \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \Delta_q} \sqrt{\boldsymbol{\sigma}^\top \mathbf{K}_{\boldsymbol{\mu}} \boldsymbol{\sigma}} \Big] && \text{(Cauchy-Schwarz)} \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \Delta_q} \sqrt{\boldsymbol{\mu} \cdot \mathbf{u}_{\boldsymbol{\sigma}}} \Big] && [\mathbf{u}_{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}^\top \mathbf{K}_1 \boldsymbol{\sigma}, \ldots, \boldsymbol{\sigma}^\top \mathbf{K}_p \boldsymbol{\sigma})^\top)] \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r} \Big]. && \text{(definition of dual norm)}
\end{aligned}$$

New Data-Dependent Bound

- Proof: in the following, $r \geq 1$ is arbitrary integer.

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H_1) &= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_\infty}\Big] \\
&\leq \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r}\Big] && (\forall r \geq 1, \|\mathbf{u}_{\boldsymbol{\sigma}}\|_\infty \leq \|\mathbf{u}_{\boldsymbol{\sigma}}\|_r) \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Bigg[\Big[\sum_{k=1}^p (\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^r\Big]^{\frac{1}{2r}}\Bigg] \\
&\leq \frac{1}{m} \Bigg[\operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sum_{k=1}^p (\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^r\Big]\Bigg]^{\frac{1}{2r}} \text{ (Jensen's inequality)} \\
&= \frac{1}{m} \Bigg[\sum_{k=1}^p \operatorname*{E}_{\boldsymbol{\sigma}} \Big[(\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^r\Big]\Bigg]^{\frac{1}{2r}} \\
&\leq \frac{1}{m} \Bigg[\sum_{k=1}^p \Big(\tfrac{23}{22} r \operatorname{Tr}[\mathbf{K}_k]\Big)^r\Bigg]^{\frac{1}{2r}} = \frac{\sqrt{\tfrac{23}{22} r \|\mathbf{u}\|_r}}{m}. && \text{(lemma)}
\end{aligned}$$

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 - L1

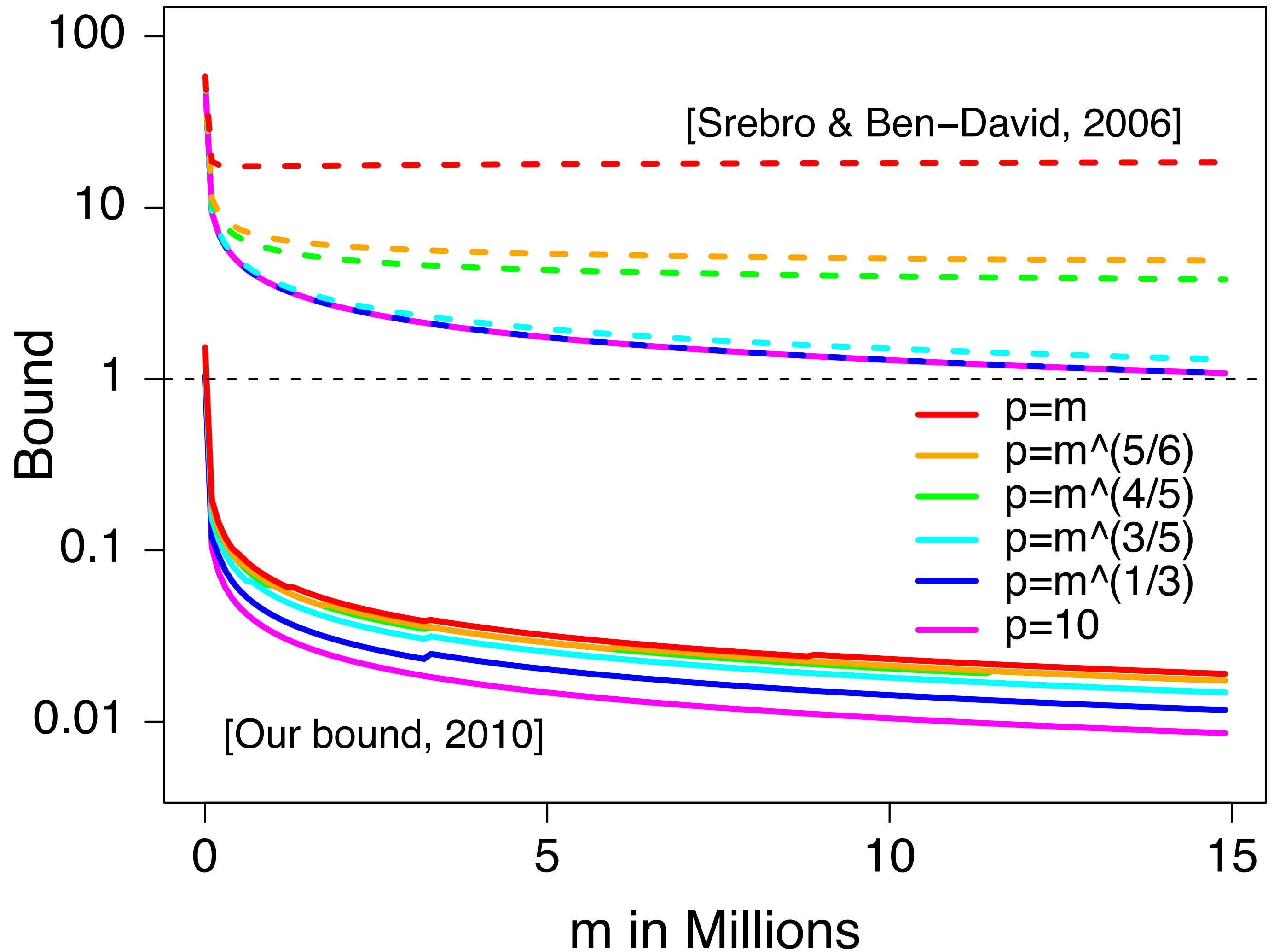
(CC, MM, and AR, 2010)

- 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 $h \in H_1$,

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

  - very weak dependency on p, no extra log terms.
  - analysis based on Rademacher complexity.
  - bound valid for $p \gg m$.
  - see also (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 $\mathcal{X} = \{\text{-}1, \text{+}1\}^p$.
  - canonical projection kernels $K_k(\mathbf{x}, \mathbf{x}') = x_k x'_k$.
  - H_1 contains $J_p = \{\mathbf{x} \mapsto s x_k \colon k \in [1, p], s \in \{\text{-}1, \text{+}1\}\}$.
  - $\text{VCdim}(J_p) = \Omega(\log p)$.
  - for $\rho = 1$ and $h \in J_p$, $\widehat{R}_\rho(h) = \widehat{R}(h)$.
  - VC lower bound: $\Omega\big(\sqrt{\text{VCdim}(J^p)/m}\big)$.

New Paper

- Recent claim (Hussain and Shawe-Taylor, AISTATS 2011): additive bound in terms of $\log p$, instead of multiplicative.
  - main proof incorrect: probabilistic bound on Rademacher complexity, but slack term left out of proof of theorem 8. Adding it $\longrightarrow$ multiplicative bound.
  - however: authors are preparing new version (private communication: J. Shawe-Taylor).

New Learning Bound - Lq

(CC, MM, and AR, 2010)

- 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 $h \in H_q$,

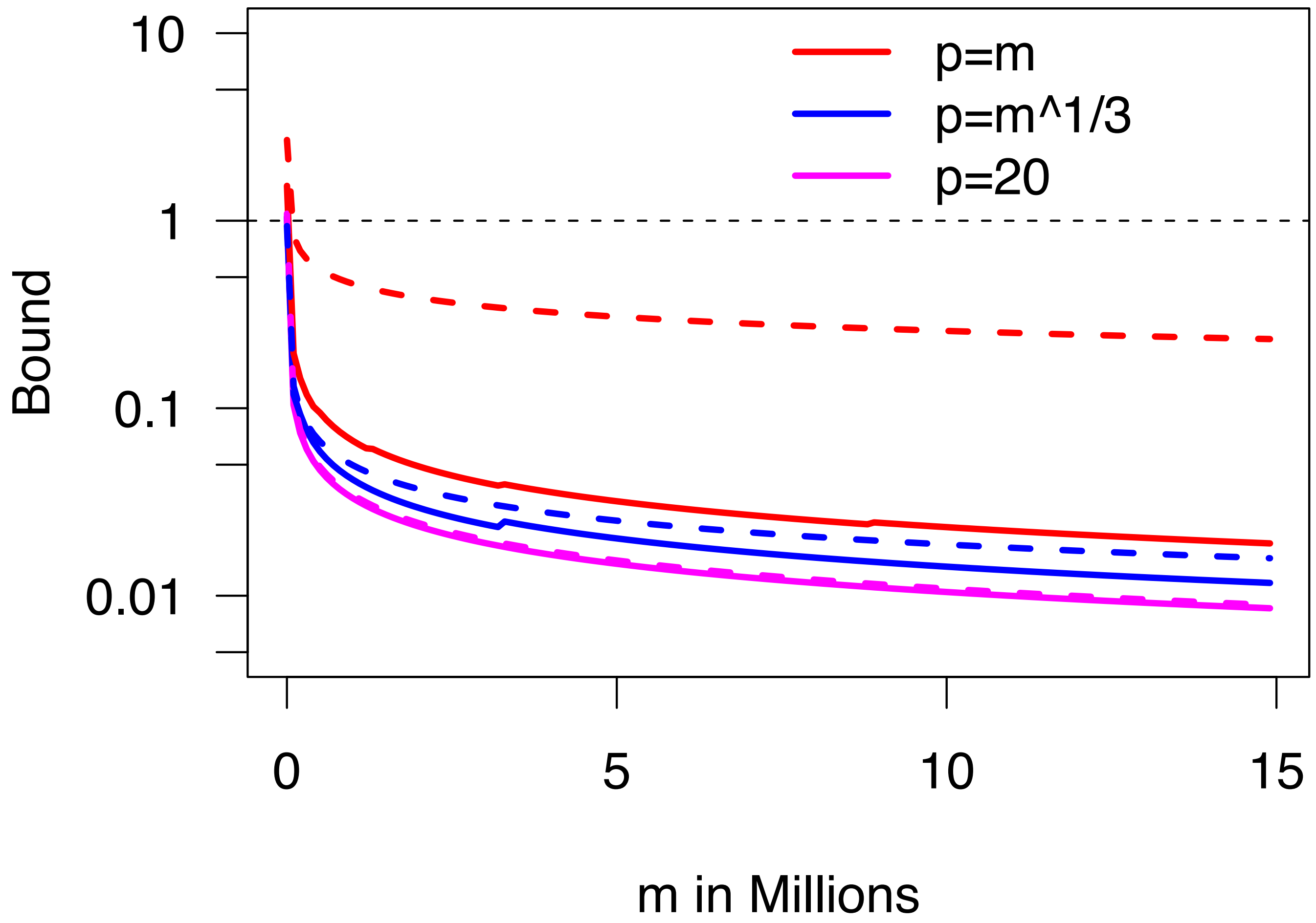
$$R(h) \leq \widehat{R}_\rho(h) + 2p^{\frac{1}{2r}} \sqrt{\frac{\frac{23}{22} r R^2 / \rho^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

  - 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_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 (p dependency).
  - mild dependency on p.
  - similar proof and analysis for other regularizations.
- Applications:
  - results suggest using large number of kernels.
  - recent results show significant improvements (CC, MM, AR, ICML 2010).

Part III

- Non-negative combinations.
- General case.

Kernel Family

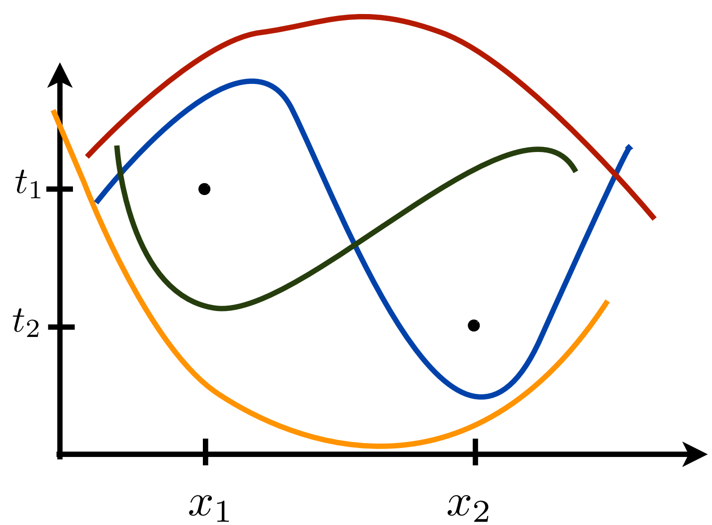
- General case:
  - $\mathcal{K}$ a family of kernels bounded by R.
  - finite pseudo-dimension: $\mathrm{Pdim}(\mathcal{K}) < \infty$.
  - general hypothesis set:

$$H_{\mathcal{K}} = \Big\{ h \in \mathbb{H}_K : K \in \mathcal{K}, \|h\|_{\mathbb{H}_K} \leq 1 \Big\}.$$

Shattering

- Definition: Let H be a hypothesis set of functions from X to $\mathbb{R}$. $A = \{x_1, \ldots, x_m\}$ is shattered by H if there exist $t_1, \ldots, t_m \in \mathbb{R}$ such that

$$\left| \left\{ \begin{bmatrix} \operatorname{sgn}\big(L(h(x_1), f(x_1)) - t_1\big) \\ \vdots \\ \operatorname{sgn}\big(L(h(x_m), f(x_m)) - t_m\big) \end{bmatrix} : h \in H \right\} \right| = 2^m.$$

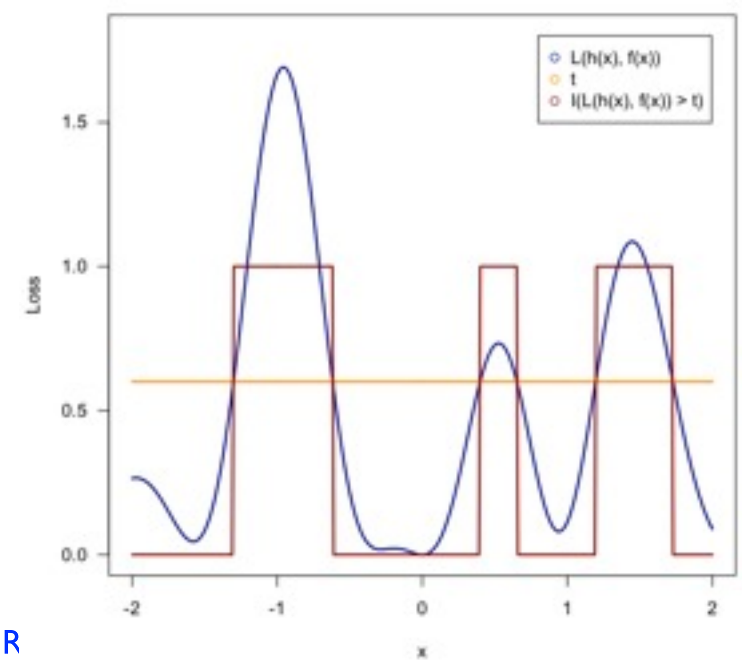


Pseudo-Dimension

(Pollard, 1984)

- Definition: Let H be a hypothesis set of functions from X to $\mathbb{R}$. The pseudo-dimension of H, $\mathrm{Pdim}(H)$, is the size of the largest set shattered by H.
- Definition (equivalent, see also (Vapnik, 1995)):

$$\mathrm{Pdim}(H) = \mathrm{VCdim}\Big(\big\{(x, t) \mapsto 1_{(h(x)-t)>0} \colon h \in H\big\}\Big).$$



Pseudo-Dimension - Properties

- Theorem: Pseudo-dimension of hyperplanes.

$$\mathrm{Pdim}(\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} + b \colon \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}) = N + 1.$$

- Theorem: Pseudo-dimension of a vector space of real-valued functions H.

$$\mathrm{Pdim}(H) = \dim(H).$$

- Theorem: Pseudo-dimension of $\phi(H) = \{\phi \circ h \colon h \in H\}$ where ϕ is a monotone function:

$$\mathrm{Pdim}(\phi(H)) \leq \dim(H).$$

General Pdim Learning Bound

(Srebro and Ben-David, 2006)

- Let $\mathcal{K}$ a family of kernel functions bounded by R. Let $d = \text{Pdim}(\mathcal{K})$, then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H_{\mathcal{K}}$,

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

  - bound additive in d (modulo log terms).
  - not informative for $d > m$.

Application: Linear Combinations

- Linear and non-negative combination of base kernels (previous section):

$$\mathcal{K}_{\text{lin}} = \Big\{ K_{\boldsymbol{\mu}} = \sum_{k=1}^{p} \mu_k K_k \colon \big(\sum_{k=1}^{p} \mu_k = 1\big) \wedge \big(\mathbf{K}_{\boldsymbol{\mu}} \succeq \mathbf{0}\big) \Big\}$$

$$\mathcal{K}_1 = \Big\{ K_{\boldsymbol{\mu}} = \sum_{k=1}^{p} \mu_k K_k \colon \big(\sum_{k=1}^{p} \mu_k = 1\big) \wedge \big(\boldsymbol{\mu} \geq \mathbf{0}\big) \Big\}$$

- Since $\mathcal{K}_{\text{lin}} \subseteq \mathcal{K}_1 \subseteq \Big\{ \sum_{k=1}^{p} \mu_k K_k \Big\}$,

$$\text{Pdim}(\mathcal{K}_1) \leq \text{Pdim}(\mathcal{K}_{\text{lin}}) \leq \dim \Big(\Big\{ \sum_{k=1}^{p} \mu_k K_k \Big\} \Big) = p.$$

Application: Gaussian Kernels

- Gaussian kernels with a fixed covariance matrix:

$$\mathcal{K}_{\text{Gaussian}} = \left\{(\mathbf{x}_1, \mathbf{x}_2) \mapsto \exp(-(\mathbf{x}_2 - \mathbf{x}_1)^\top \mathbf{A}(\mathbf{x}_2 - \mathbf{x}_1)) \colon \mathbf{A} \in \mathbb{S}_+^N\right\}.$$

- since $\exp$ is monotone and since

$$\left\{(\mathbf{x}_1, \mathbf{x}_2) \mapsto (\mathbf{x}_2 - \mathbf{x}_1)^\top \mathbf{A}(\mathbf{x}_2 - \mathbf{x}_1) \colon \mathbf{A} \in \mathbb{S}_+^N\right\}$$
$$= \left\{(\mathbf{x}_1, \mathbf{x}_2) \mapsto \sum_{i,j=1}^{n} \mathbf{A}_{ij}(\mathbf{x}_2 - \mathbf{x}_1)_i(\mathbf{x}_2 - \mathbf{x}_1)_j \colon \mathbf{A} \in \mathbb{S}_+^N\right\}$$
$$\subseteq \text{span}\left\{(\mathbf{x}_1, \mathbf{x}_2) \mapsto (\mathbf{x}_2 - \mathbf{x}_1)_i(\mathbf{x}_2 - \mathbf{x}_1)_j \colon 1 \leq i \leq j \leq N\right\},$$

- $\text{Pdim}(\mathcal{K}_{\text{Gaussian}}) \leq \frac{N(N-1)}{2}$.
- Similar for **A** diagonal, $\text{Pdim}(\mathcal{K}_{\text{Gaussian}}) \leq N$.

References

- Bousquet, Olivier and Herrmann, Daniel J. L. On the complexity of learning the kernel matrix. In NIPS, 2002.
- Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. Generalization Bounds for Learning Kernels. In ICML, 2010.
- Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. Two-stage learning kernel methods. In ICML, 2010.
- Zakria Hussain, John Shawe-Taylor. Improved Loss Bounds For Multiple Kernel Learning. In AISTATS, 2011.
- Kakade, Sham M., Shalev-Shwartz, Shai, and Tewari, Ambuj. Applications of strong convexity–strong smoothness duality to learning with matrices, 2010. arXiv:0910.0610v1.
- Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics, 30, 2002.
- Koltchinskii, Vladimir and Yuan, Ming. Sparse recovery in large ensembles of kernel machines on-line learning and bandits. In COLT, 2008.

References

- Lanckriet, Gert, Cristianini, Nello, Bartlett, Peter, Ghaoui, Laurent El, and Jordan, Michael. Learning the kernel matrix with semidefinite programming. JMLR, 5, 2004.
- Srebro, Nathan and Ben-David, Shai. Learning bounds for support vector machines with learned kernels. In COLT, 2006.
- Ying, Yiming and Campbell, Colin. Generalization bounds for learning the kernel problem. In COLT, 2009.