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

user \rightarrow \text{kernel } K \rightarrow \text{algorithm} \rightarrow h

sample
Learning Kernel Framework

user \rightarrow \text{kernel family } \mathcal{K} \rightarrow \text{algorithm} \rightarrow (K, h) \rightarrow \text{sample}
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, $q \geq 1$,

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

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

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 $S = (x_1, \ldots, x_m)$, 

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

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

- **Rademacher complexity of** $H$:

  $$ \mathcal{R}_m(H) = \mathbb{E}_{S \sim D^m} [\hat{\mathcal{R}}_S(H)]. $$
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 $h \in H_1$,

$$R(h) \leq \hat{R}_\rho(h) + 2\sqrt{\frac{R^2/\rho^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$
Early Learning Kernel Bounds

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

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

$$R(h) \leq \hat{R}_\rho(h) + \frac{1}{\sqrt{m}} \left[ \sqrt{\max_{p=1}^P \text{Tr}(K_k) \max_{k=1}^P \frac{\|K_k\|}{\text{Tr}(K_k)}} + 4 + \sqrt{2 \log \frac{1}{\delta}} \right].$$

- 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 $h \in H_1$,

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

- 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 \( h \in H_1 \),

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

- 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

Theorem: for any sample \( S \) of size \( m \), and positive integer \( r \),

\[
\hat{\mathcal{R}}_S(H_1) \leq \sqrt{\frac{23}{22}} \frac{r\|u\|_r}{m},
\]

with \( u = (\text{Tr}[K_1], \ldots, \text{Tr}[K_p])^\top \).

- similarity with single kernel bound.
- can be used directly to derive an algorithm.
New Data-Dependent Bound

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

\[
\hat{\mathcal{R}}_S(H_q) = \frac{1}{m} \mathbb{E} \left[ \sup_{h \in H_q} \sum_{i=1}^{m} \sigma_i h(x_i) \right]
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{\mu \in \Delta_q, \alpha^\top K \alpha \leq 1} \sum_{i,j=1}^{m} \sigma_i \alpha_j K_{\mu}(x_i, x_j) \right]
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{\mu \in \Delta_q, \alpha^\top K \alpha \leq 1} \sigma^\top K_{\mu} \alpha \right] = \frac{1}{m} \mathbb{E} \left[ \sup_{\mu \in \Delta_q, \|\alpha\|_{K^{1/2}} \leq 1} \langle \sigma, \alpha \rangle_{K^{1/2}} \right]
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{\mu \in \Delta_q} \sqrt{\sigma^\top K_{\mu} \sigma} \right] \quad \text{(Cauchy-Schwarz)}
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{\mu \in \Delta_q} \sqrt{\mu \cdot u_\sigma} \right] \quad \text{[} u_\sigma = (\sigma^\top K_1 \sigma, \ldots, \sigma^\top K_p \sigma)^\top \text{]} \quad \text{(definition of dual norm)}
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sqrt{\|u_\sigma\|_r} \right].
\]

Corinna Cortes, Mehryar Mohri, Afshin Rostami - ICML 2011 Tutorial.
New Data-Dependent Bound

Proof: in the following, \( r \geq 1 \) is arbitrary integer.

\[
\hat{R}_S(H_1) = \frac{1}{m} E_{\sigma} \left[ \sqrt{\|u_\sigma\|_\infty} \right] \\
\leq \frac{1}{m} E_{\sigma} \left[ \sqrt{\|u_\sigma\|_r} \right] \\
= \frac{1}{m} E_{\sigma} \left[ \left( \sum_{k=1}^{p} (\sigma^\top K_k \sigma)^r \right)^{\frac{1}{2^r}} \right] \\
\leq \frac{1}{m} \left[ E_{\sigma} \left( \sum_{k=1}^{p} (\sigma^\top K_k \sigma)^r \right) \right]^{\frac{1}{2^r}} \quad \text{(Jensen’s inequality)} \\
= \frac{1}{m} \left[ \sum_{k=1}^{p} E_{\sigma} \left( (\sigma^\top K_k \sigma)^r \right) \right]^{\frac{1}{2^r}} \\
\leq \frac{1}{m} \left[ \sum_{k=1}^{p} \left( \frac{23}{22} r \text{Tr}[K_k] \right)^r \right]^{\frac{1}{2^r}} = \sqrt{\frac{23}{22} r \|u\|_r} \frac{1}{m}. 
\]
Key Lemma

Lemma: Let $K$ be a kernel matrix for a finite sample. Then, for any integer $r$,

$$E_{\sigma} \left[ (\sigma^T K \sigma)^r \right] \leq \left( \frac{23}{22} r \operatorname{Tr}[K] \right)^r.$$ 

- 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 $h \in H_1$,

$$R(h) \leq \hat{R}_\rho(h) + 2 \sqrt{\frac{23}{22} e \left[ \log p \right] R^2 / \rho^2} \sqrt{\frac{1}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

- 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

\[ \rho / R = 0.2 \]

[Srebro & Ben-David, 2006]

[Our bound, 2010]
Lower Bound

- **Tight bound:**
  - dependency $\sqrt{\log p}$ cannot be improved.
  - argument based on VC dimension or example.

- **Observations:** case $\mathcal{X}=\{-1, +1\}^p$.
  - canonical projection kernels $K_k(x, x') = x_k x'_k$.
  - $H_1$ contains $J_p = \{x \mapsto sx_k : k \in [1, p], s \in \{-1, +1\}\}$.
  - $\text{VCdim}(J_p) = \Omega(\log p)$.
  - for $\rho = 1$ and $h \in J_p$, $\hat{R}_\rho(h) = \hat{R}(h)$.
  - VC lower bound: $\Omega(\sqrt{\text{VCdim}(J_p)/m})$. 
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 → multiplicative bound.

- however: authors are preparing new version (private communication: J. Shawe-Taylor).
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 $h \in H_q$,

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

- 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.
  - $\sum_{k=1}^{p} \mu_k K_k = (\sum_{k=1}^{p} \mu_k) K_1$.
  - thus, $\|h\|_{\mathcal{H}_{K_1}}^2 = (\sum_{k=1}^{p} \mu_k)\|h\|_{\mathcal{H}_{K}}^2$ for $\sum_{k=1}^{p} \mu_k \neq 0$.
  - $\sum_{k=1}^{p} \mu_k \leq p^{\frac{1}{r}} \|\mu\|_q = p^{\frac{1}{r}}$ (Hölder’s inequality).
  - $H_q$ coincides with $\{h \in \mathcal{H}_{K_1}: \|h\|_{\mathcal{H}_{K_1}} \leq p^{\frac{1}{2r}}\}$. 
Comparison L1 vs L2

![Graph comparing L1 and L2 norms, showing bounds for different values of p: p=m, p=m^{1/3}, and p=20. The x-axis represents m in millions, and the y-axis represents the bound on the norm.]
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: \( \text{Pdim}(\mathcal{K}) < \infty \).
  - general hypothesis set:

\[
H_{\mathcal{K}} = \left\{ h \in \mathbb{H}_K : K \in \mathcal{K}, \|h\|_{\mathbb{H}_K} \leq 1 \right\}.
\]
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

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

(Pollard, 1984)

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

**Definition** (equivalent, see also (Vapnik, 1995)):

$$\text{Pdim}(H) = \text{VCdim}\left(\{(x, t) \mapsto 1_{(h(x)-t)>0} : h \in H\}\right).$$
Pseudo-Dimension - Properties

- **Theorem**: Pseudo-dimension of hyperplanes.

\[
P_{\text{dim}}(x \mapsto w \cdot x + b: w \in \mathbb{R}^N, b \in \mathbb{R}) = N + 1.
\]

- **Theorem**: Pseudo-dimension of a vector space of real-valued functions \(H\).

\[
P_{\text{dim}}(H) = \dim(H).
\]

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

\[
P_{\text{dim}}(\phi(H)) \leq \dim(H).
\]
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 $h \in H_\mathcal{K}$,

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

- 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):

\[
\mathcal{K}_{\text{lin}} = \left\{ K_\mu = \sum_{k=1}^{p} \mu_k K_k : \left( \sum_{k=1}^{p} \mu_k = 1 \right) \land \left( K_\mu \succeq 0 \right) \right\}
\]

\[
\mathcal{K}_1 = \left\{ K_\mu = \sum_{k=1}^{p} \mu_k K_k : \left( \sum_{k=1}^{p} \mu_k = 1 \right) \land \left( \mu \succeq 0 \right) \right\}
\]

- Since \( \mathcal{K}_{\text{lin}} \subseteq \mathcal{K}_1 \subseteq \left\{ \sum_{k=1}^{p} \mu_k K_k \right\} \),

\[
P\dim(\mathcal{K}_1) \leq P\dim(\mathcal{K}_{\text{lin}}) \leq \dim \left( \left\{ \sum_{k=1}^{p} \mu_k K_k \right\} \right) = p.
\]
Application: Gaussian Kernels

- Gaussian kernels with a fixed covariance matrix:

\[ \mathcal{K}_{\text{Gaussian}} = \left\{ (x_1, x_2) \mapsto \exp(-(x_2 - x_1)^\top A (x_2 - x_1)) : A \in \mathbb{S}_+^N \right\}. \]

- since \( \exp \) is monotone and since

\[ \left\{ (x_1, x_2) \mapsto (x_2 - x_1)^\top A (x_2 - x_1) : A \in \mathbb{S}_+^N \right\} = \left\{ (x_1, x_2) \mapsto \sum_{i,j=1}^{n} A_{ij} (x_2 - x_1)_i (x_2 - x_1)_j : A \in \mathbb{S}_+^N \right\} \subseteq \text{span}\left\{ (x_1, x_2) \mapsto (x_2 - x_1)_i (x_2 - x_1)_j : 1 \leq i \leq j \leq N \right\}, \]

- \( \text{Pdim}(\mathcal{K}_{\text{Gaussian}}) \leq \frac{N(N-1)}{2}. \)

- Similar for \( A \) diagonal, \( \text{Pdim}(\mathcal{K}_{\text{Gaussian}}) \leq N. \)
References


References

