Advanced Machine Learning
Learning Kernels

MEHRYAR MOHRI
MOHRI@
COURANT INSTITUTE & GOOGLE RESEARCH
Outline

- Kernel methods.

- Learning kernels
  - scenario.
  - learning bounds.
  - algorithms.
Machine Learning Components

user → features → algorithm → sample → $h$
Machine Learning Components

- **User** → **Features**
- **Sample** → **Algorithm**
- **Algorithm** → **$h$**

Critical task: main focus of ML literature
Kernel Methods

- Features $\Phi : X \rightarrow \mathbb{H}$ implicitly defined via the choice of a PDS kernel $K$
  \[ \forall x, y \in X, \quad \Phi(x) \cdot \Phi(y) = K(x, y). \]

- $K$ interpreted as a similarity measure.

- Flexibility: PDS kernel can be chosen arbitrarily.

- Help extend a variety of algorithms to non-linear predictors, e.g., SVMs, KRR, SVR, KPCA.

- PDS condition directly related to convexity of optimization problem.
Example - Polynomial Kernels

- **Definition:**

\[
\forall x, y \in \mathbb{R}^N, \ K(x, y) = (x \cdot y + c)^d, \quad c > 0.
\]

- **Example:** for \( N = 2 \) and \( d = 2 \),

\[
K(x, y) = (x_1 y_1 + x_2 y_2 + c)^2
\]

\[
= \begin{bmatrix}
x_1^2 & \sqrt{2} x_1 x_2 \\
x_2^2 & \sqrt{2} c x_1 \\
\sqrt{2} x_1 x_2 & \sqrt{2} c x_2 \\
\sqrt{2} c x_1 & \sqrt{2} c x_2 \\
c & c
\end{bmatrix}
\begin{bmatrix}
y_1^2 \\
y_2^2 \\
\sqrt{2} y_1 y_2 \\
\sqrt{2} c y_1 \\
\sqrt{2} c y_2 \\
c
\end{bmatrix}.
\]
XOR Problem

Use second-degree polynomial kernel with $c = 1$:

$\sqrt{2} x_1 x_2$

$\sqrt{2} x_1$

Linearly non-separable   Linearly separable by $x_1 x_2 = 0$. 
Other Standard PDS Kernels

- **Gaussian kernels:**

  \[ K(x, y) = \exp \left( -\frac{||x - y||^2}{2\sigma^2} \right), \quad \sigma \neq 0. \]

  - Normalized kernel of \((x, x') \mapsto \exp \left( \frac{x \cdot x'}{\sigma^2} \right)\).

- **Sigmoid Kernels:**

  \[ K(x, y) = \tanh(a(x \cdot y) + b), \quad a, b \geq 0. \]
SVM

(Cortes and Vapnik, 1995; Boser, Guyon, and Vapnik, 1992)

Primal:

$$
\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \left( 1 - y_i (\mathbf{w} \cdot \Phi_K(x_i) + b) \right)_+.
$$

Dual:

$$
\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)
$$

subject to: $0 \leq \alpha_i \leq C \land \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]$. 
Kernel Ridge Regression

(Hoerl and Kennard, 1970; Sanders et al., 1998)

- **Primal:**

\[
\min_w \lambda \|w\|^2 + \sum_{i=1}^{m} \left( w \cdot \Phi_K(x_i) + b - y_i \right)^2.
\]

- **Dual:**

\[
\max_{\alpha \in \mathbb{R}^m} -\alpha^T (K + \lambda I) \alpha + 2\alpha^T y.
\]
Questions

How should the user choose the kernel?

- problem similar to that of selecting features for other learning algorithms.
- poor choice → learning made very difficult.
- good choice → even poor learners could succeed.

The requirement from the user is thus critical.

- can this requirement be lessened?
- is a more automatic selection of features possible?
Outline

- Kernel methods.
- Learning kernels
  - scenario.
  - learning bounds.
  - algorithms.
Standard Learning with Kernels

user \rightarrow \text{kernel } K \rightarrow \text{algorithm } h \rightarrow \text{sample}
Learning Kernel Framework

user \rightarrow \text{kernel family } \mathcal{K} \rightarrow \text{algorithm} \rightarrow (K, h) \rightarrow \text{sample}
Kernel Families

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

$$
\mathcal{K}_q = \left\{ K_\mu : K_\mu = \sum_{k=1}^{p} \mu_k K_k, \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \in \Delta_q \right\}
$$

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

Hypothesis sets:

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

Lemma: for $p, q \in (0, +\infty]$, the following holds:

$$\forall \mathbf{x} \in \mathbb{R}^N, p \leq q \Rightarrow \|\mathbf{x}\|_q \leq \|\mathbf{x}\|_p \leq N^{\frac{1}{p} - \frac{1}{q}} \|\mathbf{x}\|_q.$$ 

Proof: for the left inequalities, observe that for $\mathbf{x} \neq 0$,

$$\left(\frac{\|\mathbf{x}\|_p}{\|\mathbf{x}\|_q}\right)^p \leq \sum_{i=1}^N \left(\frac{|x_i|}{\|\mathbf{x}\|_q}\right)^p \geq \sum_{i=1}^N \left(\frac{|x_i|}{\|\mathbf{x}\|_q}\right)^q = 1.$$ 

- Right inequalities follow immediately Hölder’s inequality:

$$\|\mathbf{x}\|_p = \left[\sum_{i=1}^N |x_i|^p\right]^{\frac{1}{p}} \leq \left[\left(\sum_{i=1}^N (|x_i|^p)^{\frac{p}{q}}\right)^{\frac{p}{q}} \left(\sum_{i=1}^N (1)^{\frac{q}{q-p}}\right)^{1-\frac{p}{q}}\right]^{\frac{1}{p}} = \|\mathbf{x}\|_q N^{\frac{1}{p} - \frac{1}{q}}.$$
Single Kernel Guarantee

(Theorem) fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H_1$,

$$R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \sqrt{\text{Tr}[K]} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$
Multiple Kernel Guarantee

(Cortes, MM, and Rostamizadeh, 2010)

Theorem: fix $\rho > 0$. Let $q, r \geq 1$ with $\frac{1}{q} + \frac{1}{r} = 1$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H_q$ and any integer $1 \leq s \leq r$:

$$
R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \frac{\sqrt{s \| u \|_s}}{m} + \sqrt{\log \frac{1}{2\delta}},
$$

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

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

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

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

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

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{\mu \in \Delta_q} \sqrt{\sigma^\top \mathbf{K}_\mu \sigma} \right]
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{\mu \in \Delta_q} \sqrt{\mu \cdot u_\sigma} \right] = \frac{1}{m} \mathbb{E} \left[ \sqrt{\|u_\sigma\|_r} \right].
\]

(Cauchy-Schwarz)

\[
[u_\sigma = (\sigma^\top \mathbf{K}_1 \sigma, \ldots, \sigma^\top \mathbf{K}_p \sigma)^\top]
\]

(definition of dual norm)
Lemma

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

$$E_{\sigma} \left[ (\sigma^T K \sigma)^r \right] \leq \left( r \text{ Tr}[K] \right)^r .$$

Proof: combinatorial argument.
Proof

For any $1 \leq s \leq r$,

$$
\hat{R}_S(H_q) = \frac{1}{m} \mathbb{E}_{\sigma} \left[ \sqrt{\| u_\sigma \|_r} \right]
\leq \frac{1}{m} \mathbb{E}_{\sigma} \left[ \sqrt{\| u_\sigma \|_S} \right]
= \frac{1}{m} \mathbb{E}_{\sigma} \left[ \left( \sum_{k=1}^{p} (\sigma^\top K_k \sigma)^s \right)^{1/2s} \right]
\leq \frac{1}{m} \left[ \mathbb{E}_{\sigma} \left( \sum_{k=1}^{p} (\sigma^\top K_k \sigma)^s \right)^{1/2s} \right] (\text{Jensen's inequality})
= \frac{1}{m} \left[ \sum_{k=1}^{p} \mathbb{E}_{\sigma} \left( (\sigma^\top K_k \sigma)^s \right)^{1/2s} \right]
\leq \frac{1}{m} \left[ \sum_{k=1}^{p} \left( s \text{Tr}[K_k]^s \right)^{1/2s} \right]^{1/2s} = \frac{\sqrt{s \| u \|_S}}{m}. \quad \text{(lemma)}
$$
Corollary: fix $\rho > 0$. For any $\delta > 0$, with probability $1 - \delta$, the following holds for all $h \in H_1$:

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

- weak dependency on $p$.
- bound valid for $p \gg m$.
- $\text{Tr}[K_k] \leq m \max_x K_k(x, x)$.
Proof

For \( q = 1 \), the bound holds for any integer \( s \geq 1 \)

\[
R(h) \leq \hat{R}_\rho(h) + \frac{2\sqrt{s \| u \|_s}}{\rho m} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},
\]

with \( s \| u \|_s = s \left[ \sum_{k=1}^{p} \text{Tr}[K_k]^s \right]^{1/s} \leq s p^{1/s} \max_{k=1}^{p} \text{Tr}[K_k] \).

The function \( s \mapsto sp^{1/s} \) reaches its minimum at \( \log p \).
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 $J_p = \{x \mapsto s x_k : k \in [1, p], s \in \{-1, +1\}\}$.
- $\text{VCdim}(J_p) = \Omega(\log p)$.
- for $\rho = 1$ and $h \in J_p$, $\hat{R}_\rho(h) = \hat{R}(h)$.
- VC lower bound: $\Omega\left(\sqrt{\text{VCdim}(J^p)/m}\right)$. 
Pseudo-Dimension 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 \( h \in H_1 \),

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

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

\[ \frac{\rho}{R} = .2 \]
Corollary: fix $\rho > 0$. Let $q, r \geq 1$ with $\frac{1}{q} + \frac{1}{r} = 1$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H_q$:

$$
R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \sqrt{r p^{\frac{1}{r}} \max_{k=1}^p \text{Tr}[K_k]} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},
$$

- mild dependency on $p$.
- $\text{Tr}[K_k] \leq m \max_x K_k(x, x)$. 

(Cortes, MM, and Rostamizadeh, 2010)
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.
  - $\sum_{k=1}^{p} \mu_k K_k = \left( \sum_{k=1}^{p} \mu_k \right) K_1$.
  - thus, $\|h\|_{\mathbb{H}K_1}^2 = \left( \sum_{k=1}^{p} \mu_k \right) \|h\|_{\mathbb{H}K}^2$ 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 $\{ h \in \mathbb{H}K_1 : \|h\|_{\mathbb{H}K_1} \leq p^{\frac{1}{2r}} \}$. 
Outline

- Kernel methods.
- Learning kernels
  - scenario.
  - learning bounds.
  - algorithms.
General LK Formulation - SVMs

Notation:

- $\mathcal{K}$ set of PDS kernel functions.
- $\overline{\mathcal{K}}$ kernel matrices associated to $\mathcal{K}$, assumed convex.
- $\mathbf{Y} \in \mathbb{R}^{m \times m}$ diagonal matrix with $Y_{ii} = y_i$.

Optimization problem:

$$\min_{\mathcal{K} \in \overline{\mathcal{K}}} \max_{\alpha} \ 2 \alpha^\top 1 - \alpha^\top Y^\top K Y \alpha$$

subject to: $0 \leq \alpha \leq C \land \alpha^\top y = 0$.

- convex problem: function linear in $K$, convexity of pointwise maximum.
Parameterized LK Formulation

Notation:

• \((K_\mu)_{\mu \in \Delta}\) parameterized set of PDS kernel functions.
• \(\Delta\) convex set, \(\mu \mapsto K_\mu\) concave function.
• \(Y \in \mathbb{R}^{m \times m}\) diagonal matrix with \(Y_{ii} = y_i\).

Optimization problem:

\[
\min_{\mu \in \Delta} \max_{\alpha} \ 2 \alpha^\top 1 - \alpha^\top Y^\top K_\mu Y \alpha
\]

subject to: \(0 \leq \alpha \leq C \land \alpha^\top y = 0\).

• convex problem: function convex in \(\mu\), convexity of pointwise maximum.
Non-Negative Combinations

1. \( K_\mu = \sum_{k=1}^{p} \mu_k K_k, \mu \in \Delta_1. \)

2. By von Neumann’s generalized minimax theorem (convexity wrt \( \mu \), concavity wrt \( \alpha \), \( \Delta_1 \) convex and compact, \( \mathcal{A} \) convex and compact):

\[
\begin{align*}
\min_{\mu \in \Delta_1} \max_{\alpha \in \mathcal{A}} & \quad 2 \alpha^\top 1 - \alpha^\top Y^\top K_\mu Y \alpha \\
= \max_{\alpha \in \mathcal{A}} \min_{\mu \in \Delta_1} & \quad 2 \alpha^\top 1 - \alpha^\top Y^\top K_\mu Y \alpha \\
= \max_{\alpha \in \mathcal{A}} & \quad 2 \alpha^\top 1 - \max_{\mu \in \Delta_1} \alpha^\top Y^\top K_\mu Y \alpha \\
= \max_{\alpha \in \mathcal{A}} & \quad 2 \alpha^\top 1 - \max_{k \in [1,p]} \alpha^\top Y^\top K_k Y \alpha.
\end{align*}
\]
Non-Negative Combinations

(Lanckriet et al., 2004)

- **Optimization problem**: in view of the previous analysis, the problem can be rewritten as the following QCQP.

\[
\begin{align*}
\max_{\alpha, t} & \quad 2\alpha^\top 1 - t \\
\text{subject to: } & \forall k \in [1, p], t \geq \alpha^\top Y^\top K_k Y \alpha; \\
& 0 \leq \alpha \leq C \land \alpha^\top y = 0.
\end{align*}
\]

- complexity (interior-point methods): \(O(pm^3)\).
Equivalent Primal Formulation

- Optimization problem:

\[
\min_{w, \mu \in \Delta_q} \frac{1}{2} \sum_{k=1}^{p} \frac{\|w_k\|_2^2}{\mu_k} + C \sum_{i=1}^{m} \max \left\{ 0, 1 - y_i \left( \sum_{k=1}^{p} w_k \cdot \Phi_k(x_i) \right) \right\}.
\]
Lots of Optimization Solutions

- QCQP (Lanckriet et al., 2004).
- Wrapper methods — interleaving call to SVM solver and update of $\mu$:
  - SILP (Sonnenburg et al., 2006).
  - Reduced gradient (SimpleML) (Rakotomamonjy et al., 2008).
  - Newton’s method (Kloft et al., 2009).
  - Mirror descent (Nath et al., 2009).
- Many other methods proposed.
Does It Work?

Experiments:

- this algorithm and its different optimization solutions often do not significantly outperform the simple uniform combination kernel in practice!
- observations corroborated by NIPS workshops.

Alternative algorithms: significant improvement (see empirical results of (Gönen and Alpaydin, 2011)).

- centered alignment-based LK algorithms (Cortes, MM, and Rostamizadeh, 2010 and 2012).
LK Formulation - KRR

(Cortes, MM, and Rostamizadeh, 2009)

Kernel family:
- non-negative combinations.
- \( L_q \) regularization.

Optimization problem:

\[
\min_{\mu} \max_{\alpha} \quad -\lambda \alpha^\top \alpha - \sum_{k=1}^{p} \mu_k \alpha^\top K_k \alpha + 2\alpha^\top y
\]

subject to: \( \mu \geq 0 \land \|\mu - \mu_0\|_q \leq \Lambda \).

- convex optimization: linearity in \( \mu \) and convexity of pointwise maximum.
Projected Gradient

- Solving maximization problem in $\alpha$, closed-form solution $\alpha = (K_\mu + \lambda I)^{-1}y$, reduces problem to

$$\min_\mu y^\top (K_\mu + \lambda I)^{-1}y$$

subject to: $\mu \geq 0 \land \|\mu - \mu_0\|_2 \leq \Lambda$.

- Convex optimization problem, one solution using projection-based gradient descent:

$$\frac{\partial F}{\partial \mu_k} = \text{Tr} \left[ \frac{\partial y^\top (K_\mu + \lambda I)^{-1}y}{\partial (K_\mu + \lambda I)} \frac{\partial (K_\mu + \lambda I)}{\partial \mu_k} \right] = -\text{Tr} \left[ (K_\mu + \lambda I)^{-1}yy^\top (K_\mu + \lambda I)^{-1} \frac{\partial (K_\mu + \lambda I)}{\partial \mu_k} \right]$$

$$= -\text{Tr} \left[ (K_\mu + \lambda I)^{-1}yy^\top (K_\mu + \lambda I)^{-1}K_k \right]$$

$$= -y^\top (K_\mu + \lambda I)^{-1}K_k(K_\mu + \lambda I)^{-1}y = -\alpha^\top K_k \alpha. \quad \Box$$
Proj. Grad. KRR - L$_2$ Reg.

**Projec”tionBasedGradientDescent**((K$_k$)$_{k\in[1,p]}$, µ$_0$)

1. µ ← µ$_0$
2. µ′ ← ∞
3. while ||µ′ − µ|| > ϵ do
   4. µ ← µ′
   5. α ← (K$_μ$ + λI)$^{-1}$y
   6. µ′ ← µ + η (α$^TK_1α$, …, α$^TK_pα$)$^T$
   7. for k ← 1 to p do
      8. µ′$_k$ ← max(0, µ′$_k$)
   9. µ′ ← µ$_0$ + Λ $\frac{µ′−µ_0}{||µ′−µ_0||}$
10. return µ′
Interpolated Step KRR - $L_2$ Reg.

**INTERPOLATED ITERATIVE ALGORITHM**\((\{K_k\}_{k \in [1,p]}, \mu_0)\)

1. \(\alpha \leftarrow \infty\)
2. \(\alpha' \leftarrow (K_{\mu_0} + \lambda I)^{-1} y\)
3. while \(\|\alpha' - \alpha\| > \epsilon\) do
4. \(\alpha \leftarrow \alpha'\)
5. \(v \leftarrow (\alpha^\top K_1 \alpha, \ldots, \alpha^\top K_p \alpha)^\top\)
6. \(\mu \leftarrow \mu_0 + \Lambda \frac{v}{\|v\|}\)
7. \(\alpha' \leftarrow \eta \alpha + (1 - \eta)(K_\mu + \lambda I)^{-1} y\)
8. return \(\alpha'\)

Simple and very efficient: few iterations (less than 15).
Dense combinations are beneficial when using many kernels.

Combining kernels based on single features, can be viewed as principled feature weighting.

(Cortes, MM, and Rostamizadeh, 2009)
Conclusion

- Solid theoretical guarantees suggesting the use of a large number of base kernels.
- Broad literature on optimization techniques but often no significant improvement over uniform combination.
- Recent algorithms with significant improvements, in particular non-linear combinations.
- Still many theoretical and algorithmic questions left to explore.
References


References


- Sham M. Kakade, Shai Shalev-Shwartz, Ambuj Tewari: Regularization Techniques for Learning with Matrices. JMLR 13: 1865-1890, 2012.


References

