Advanced Machine Learning

Learning Kernels

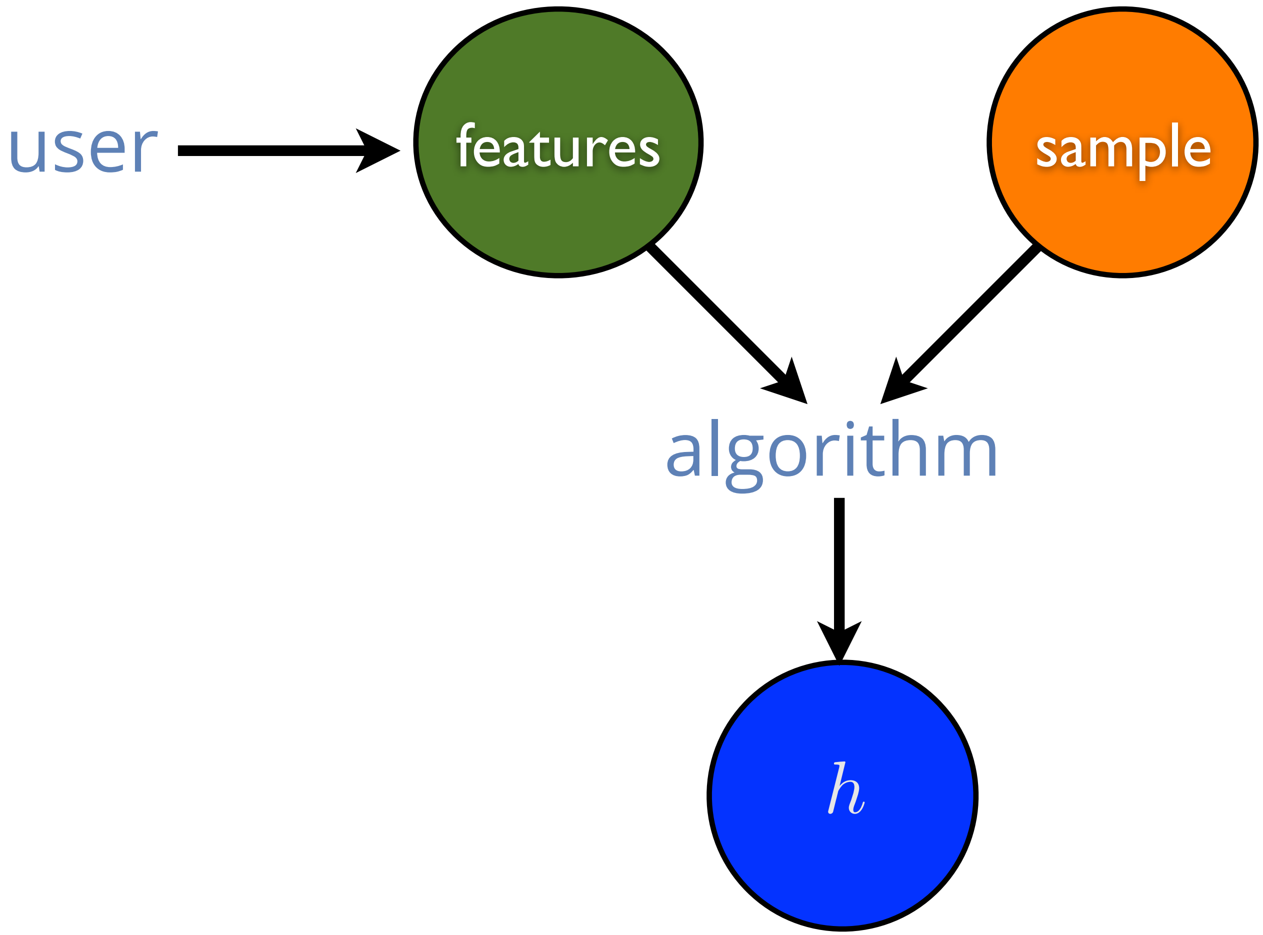
MEHRYAR MOHRI MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

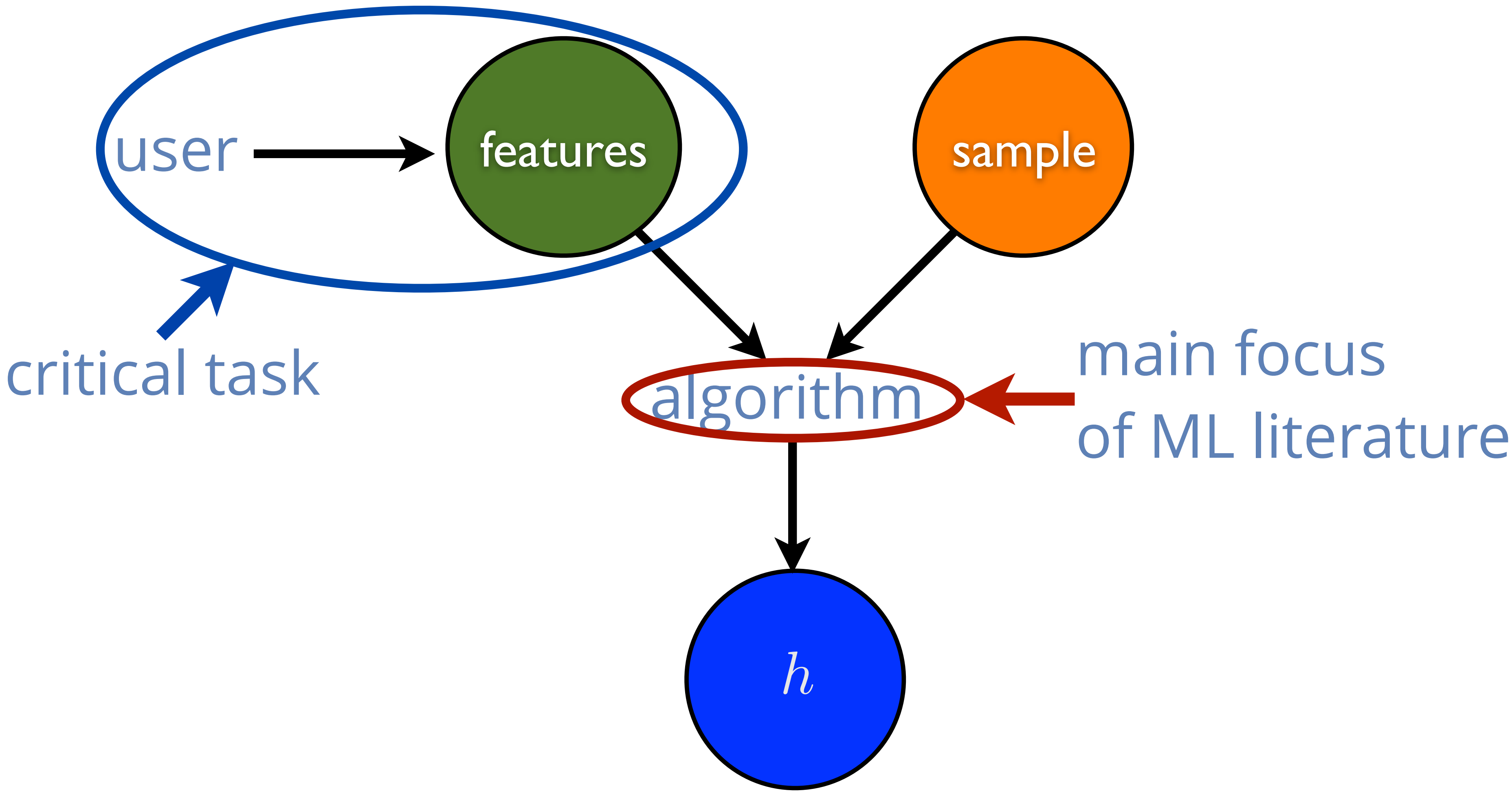
Outline

- Kernel methods.
- Learning kernels
  - scenario.
  - learning bounds.
  - algorithms.

Machine Learning Components



Machine Learning Components



Kernel Methods

- Features $\mathbf{\Phi}\colon X \to \mathbb{H}$ 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:

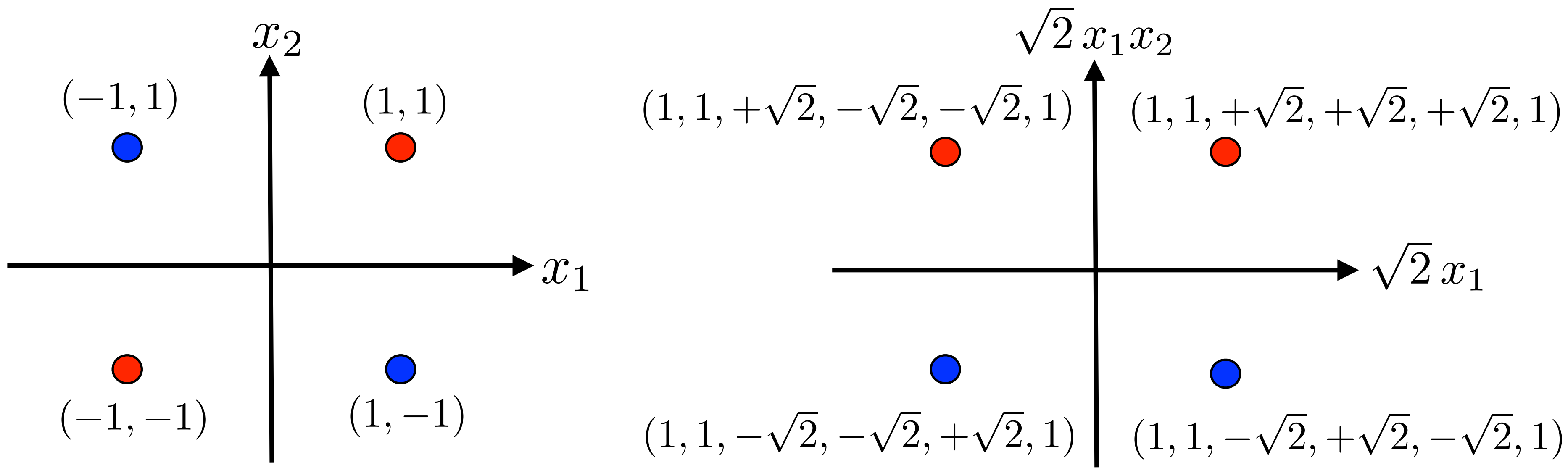
$$\forall x, y \in \mathbb{R}^N, \; K(x, y) = (x \cdot y + c)^d, \quad c > 0.$$

- Example: for $N = 2$ and $d = 2$,

$$K(x, y) = (x_1 y_1 + x_2 y_2 + c)^2$$
$$= \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \\ \sqrt{2c}\, x_1 \\ \sqrt{2c}\, x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2}\, y_1 y_2 \\ \sqrt{2c}\, y_1 \\ \sqrt{2c}\, y_2 \\ c \end{bmatrix}.$$

XOR Problem

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



Linearly non-separable

Linearly separable by $x_1 x_2 = 0$.

Other Standard PDS Kernels

- **Gaussian kernels:**

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

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

- **Sigmoid Kernels:**

$$K(x, y) = \tanh(a(x \cdot y) + b), \ \ a, b \geq 0.$$

SVM

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

- Primal:

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

- Dual:

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

$$\text{subject to: } 0 \leq \alpha_i \leq C \wedge \sum_{i=1}^{m}\alpha_i y_i = 0, i \in [1, m].$$

Kernel Ridge Regression

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

- Primal:

$$\min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \left(\mathbf{w} \cdot \mathbf{\Phi}_K(x_i) + b - y_i\right)^2 .$$

- Dual:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} -\boldsymbol{\alpha}^\top (\mathbf{K} + \lambda \mathbf{I}) \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{y}.$$

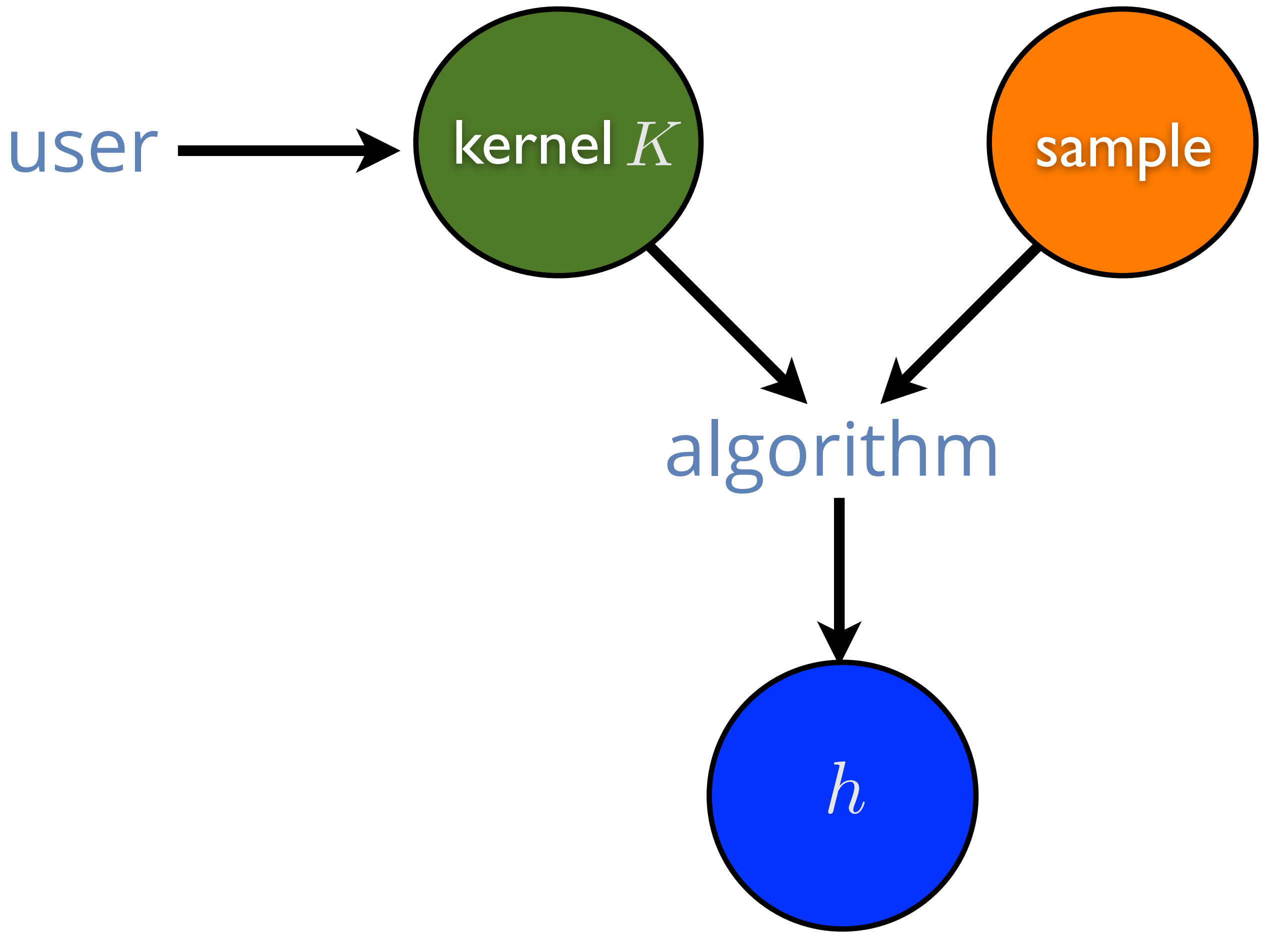
Questions

- 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?

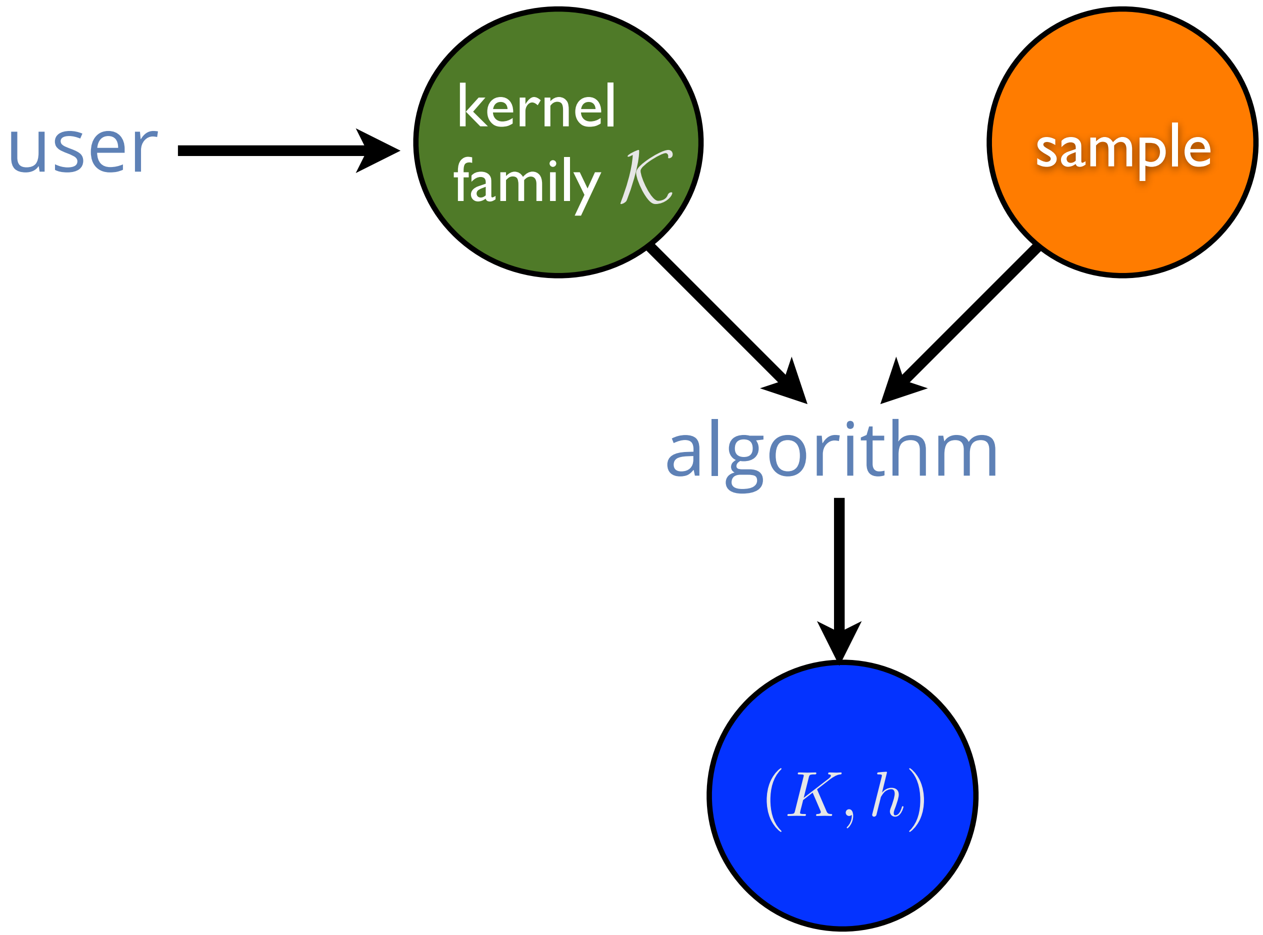
Outline

- Kernel methods.
- Learning kernels
  - scenario.
  - learning bounds.
  - algorithms.

Standard Learning with Kernels



Learning Kernel Framework



Kernel Families

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

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

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

- Hypothesis sets:

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

Relation between Norms

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

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

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

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

  - Right inequalities follow immediately Hölder's inequality:

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

Single Kernel Guarantee

(Koltchinskii and Panchenko, 2002)

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

$$R(h) \leq \widehat{R}_\rho(h) + \frac{2}{\rho} \frac{\sqrt{\mathrm{Tr}[\mathbf{K}]}}{m} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

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 $1 \leq s \leq r$:

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

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

Proof

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

$$\widehat{\mathfrak{R}}_S(H_q) = \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in H_q} \sum_{i=1}^{m} \sigma_i h(x_i) \right]$$

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

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

$$= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\mu} \in \Delta_q} \sqrt{\boldsymbol{\sigma}^\top \mathbf{K}_{\boldsymbol{\mu}} \boldsymbol{\sigma}} \right] \qquad \text{(Cauchy-Schwarz)}$$

$$= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\mu} \in \Delta_q} \sqrt{\boldsymbol{\mu} \cdot \mathbf{u}_{\boldsymbol{\sigma}}} \right] \qquad \left[\mathbf{u}_{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}^\top \mathbf{K}_1 \boldsymbol{\sigma}, \ldots, \boldsymbol{\sigma}^\top \mathbf{K}_p \boldsymbol{\sigma})^\top) \right]$$

$$= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \left[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r} \right]. \qquad \text{(definition of dual norm)}$$

Lemma

(Cortes, MM, and Rostamizadeh, 2010)

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

$$\operatorname*{E}_{\boldsymbol{\sigma}}\Big[(\boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\sigma})^r\Big] \leq \Big(r \operatorname{Tr}[\mathbf{K}]\Big)^r.$$

- Proof: combinatorial argument.

Proof

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

$$
\begin{aligned}
\widehat{\mathfrak{R}}_S(H_q) &= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r} \Big] \\
&\leq \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_s} \Big] \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \bigg[\Big[\sum_{k=1}^{p} (\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^s \Big]^{\frac{1}{2s}} \bigg] \\
&\leq \frac{1}{m} \bigg[\operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sum_{k=1}^{p} (\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^s \Big] \bigg]^{\frac{1}{2s}} \quad \text{(Jensen's inequality)} \\
&= \frac{1}{m} \bigg[\sum_{k=1}^{p} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[(\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^s \Big] \bigg]^{\frac{1}{2s}} \\
&\leq \frac{1}{m} \bigg[\sum_{k=1}^{p} \Big(s \operatorname{Tr}[\mathbf{K}_k] \Big)^s \bigg]^{\frac{1}{2s}} = \frac{\sqrt{s \|\mathbf{u}\|_s}}{m}. \qquad \text{(lemma)}
\end{aligned}
$$

L_1 Learning Bound

(Cortes, MM, and Rostamizadeh, 2010)

- Corollary: fix $\rho > 0$. For any $\delta > 0$, with probability $1 - \delta$, the following holds for all $h \in H_1$:

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

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

Proof

- For $q = 1$, the bound holds for any integer $s \geq 1$

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

with $$s\|\mathbf{u}\|_s = s \left[\sum_{k=1}^p \mathrm{Tr}[\mathbf{K}_k]^s \right]^{\frac{1}{s}} \leq s p^{\frac{1}{s}} \max_{k=1}^p \mathrm{Tr}[\mathbf{K}_k].$$

- The function $s \mapsto s p^{\frac{1}{s}}$ reaches it minimum at $\log p$.

Lower Bound

- Tight bound:
  - dependency $\sqrt{\log p}$ cannot be improved.
  - argument based on VC dimension or example.
- Observations: case $\mathcal{X} = \{\text{-}1, +1\}^p$.
  - canonical projection kernels $K_k(\mathbf{x}, \mathbf{x}') = x_k x'_k$.
  - H_1 contains $J_p = \{\mathbf{x} \mapsto s x_k : k \in [1, p], s \in \{\text{-}1, +1\}\}$.
  - $\text{VCdim}(J_p) = \Omega(\log p)$.
  - for $\rho = 1$ and $h \in J_p$, $\widehat{R}_\rho(h) = \widehat{R}(h)$.
  - VC lower bound: $\Omega\big(\sqrt{\text{VCdim}(J^p)/m}\big)$.

Pseudo-Dimension Bound

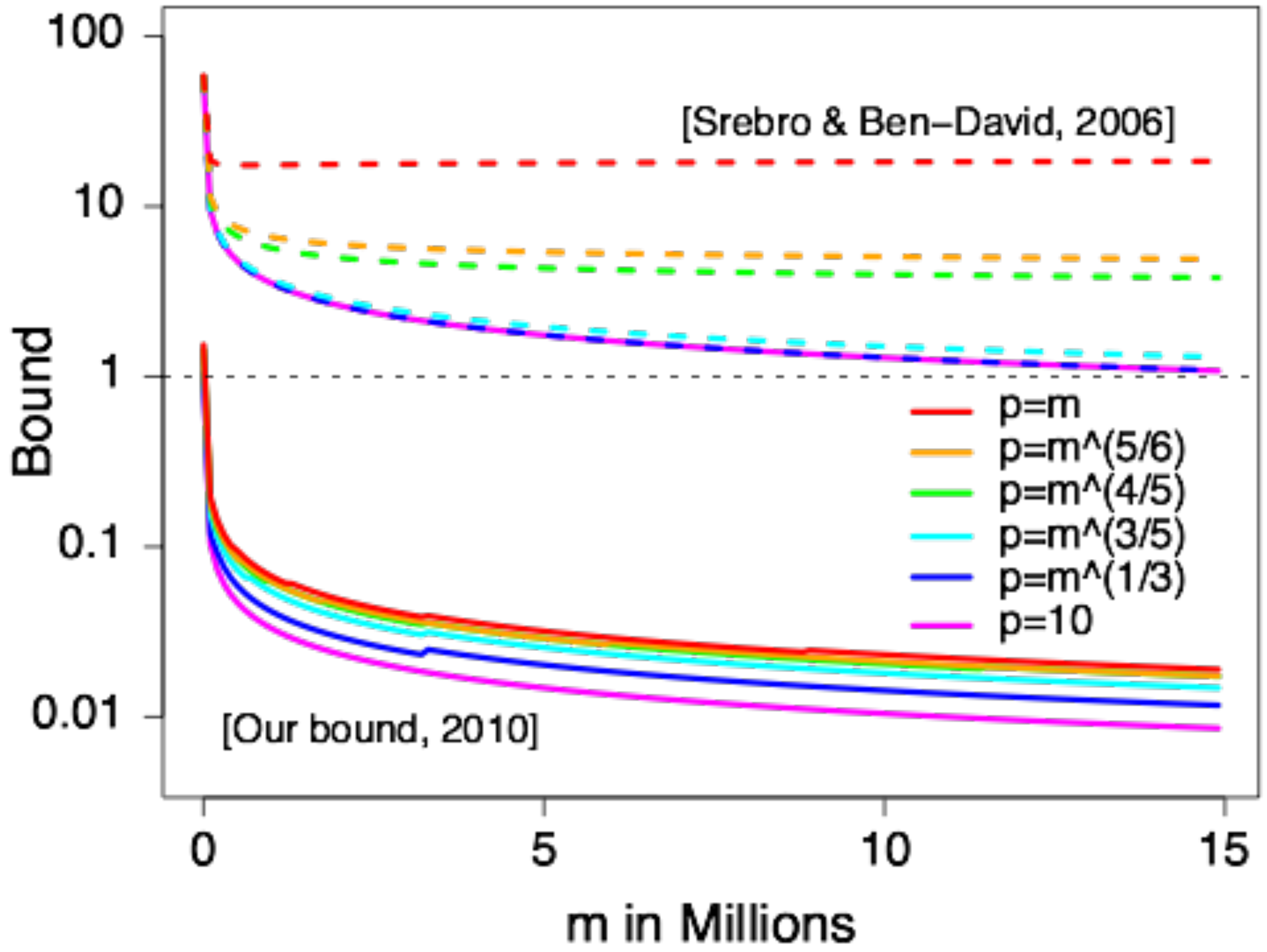
(Srebro and Ben-David, 2006)

- 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 $h \in H_1$,

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

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

Comparison



$$\rho/R = .2$$

L_q Learning Bound

(Cortes, MM, and Rostamizadeh, 2010)

- 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 $h \in H_q$:

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

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

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 $\{h \in \mathbb{H}_{K_1} : \|h\|_{\mathbb{H}_{K_1}} \leq p^{\frac{1}{2r}}\}$.

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 $\mathbf{Y}_{ii} = \mathbf{y}_i$.
- Optimization problem:

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

$$\text{subject to: } \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0.$$

  - convex problem: function linear in $\mathbf{K}$, convexity of pointwise maximum.

Parameterized LK Formulation

- Notation:
  - $(K_{\boldsymbol{\mu}})_{\boldsymbol{\mu} \in \Delta}$ parameterized set of PDS kernel functions.
  - Δ convex set, $\mu \mapsto \mathbf{K}_\mu$ concave function.
  - $\mathbf{Y} \in \mathbb{R}^{m \times m}$ diagonal matrix with $\mathbf{Y}_{ii} = \mathbf{y}_i$.
- Optimization problem:

$$\min_{\mu \in \Delta} \max_{\boldsymbol{\alpha}} 2\, \boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_{\boldsymbol{\mu}} \mathbf{Y} \boldsymbol{\alpha}$$

$$\text{subject to: } \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0.$$

  - convex problem: function convex in $\boldsymbol{\mu}$, convexity of pointwise maximum.

Non-Negative Combinations

- $K_\mu = \sum_{k=1}^{p} \mu_k K_k$, $\mu \in \Delta_1$.
- By von Neumann's generalized minimax theorem (convexity wrt $\boldsymbol{\mu}$, concavity wrt $\boldsymbol{\alpha}$, Δ_1 convex and compact, $\mathcal{A}$ convex and compact):

$$\begin{aligned}
&\min_{\mu \in \Delta_1} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \; 2\,\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_{\boldsymbol{\mu}} \mathbf{Y} \boldsymbol{\alpha} \\
&= \max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\mu \in \Delta_1} \; 2\,\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_{\boldsymbol{\mu}} \mathbf{Y} \boldsymbol{\alpha} \\
&= \max_{\boldsymbol{\alpha} \in \mathcal{A}} \; 2\,\boldsymbol{\alpha}^\top \mathbf{1} - \max_{\mu \in \Delta_1} \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_{\boldsymbol{\mu}} \mathbf{Y} \boldsymbol{\alpha} \\
&= \max_{\boldsymbol{\alpha} \in \mathcal{A}} \; 2\,\boldsymbol{\alpha}^\top \mathbf{1} - \max_{k \in [1,p]} \; \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha}.
\end{aligned}$$

Non-Negative Combinations

(Lanckriet et al., 2004)

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

$$\max_{\boldsymbol{\alpha}, t} \; 2\boldsymbol{\alpha}^\top \mathbf{1} - t$$

$$\text{subject to: } \forall k \in [1, p], t \geq \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha};$$

$$\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0.$$

  - complexity (interior-point methods): $O(pm^3)$.

Equivalent Primal Formulation

- Optimization problem:

$$\min_{w,\mu \in \Delta_q} \frac{1}{2} \sum_{k=1}^{p} \frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C \sum_{i=1}^{m} \max \left\{ 0, 1 - y_i \left(\sum_{k=1}^{p} \mathbf{w}_k \cdot \mathbf{\Phi}_k(x_i) \right) \right\}.$$

Lots of Optimization Solutions

- QCQP (Lanckriet et al., 2004).
- Wrapper methods — interleaving call to SVM solver and update of $\boldsymbol{\mu}$:
  - 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).
- On-line method (Orabona & Jie, 2011).
- 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).
  - **non-linear combination of kernels** (Cortes, MM, and Rostamizadeh, 2009).

LK Formulation - KRR

(Cortes, MM, and Rostamizadeh, 2009)

- Kernel family:
  - non-negative combinations.
  - L_q regularization.
- Optimization problem:

$$\min_{\boldsymbol{\mu}} \max_{\boldsymbol{\alpha}} \; -\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \sum_{k=1}^{p} \mu_k \boldsymbol{\alpha}^\top \mathbf{K}_k \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{y}$$

$$\text{subject to: } \boldsymbol{\mu} \geq 0 \wedge \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|_q \leq \Lambda.$$

  - convex optimization: linearity in $\boldsymbol{\mu}$ and convexity of pointwise maximum.

Projected Gradient

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

$$\min_{\boldsymbol{\mu}} \ \mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1}\mathbf{y}$$

$$\text{subject to: } \boldsymbol{\mu} \geq 0 \wedge \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|_2 \leq \Lambda.$$

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

$$\begin{aligned}
\frac{\partial F}{\partial \mu_k} &= \mathrm{Tr}\left[\frac{\partial \mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1}\mathbf{y}}{\partial(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})} \frac{\partial(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})}{\partial \mu_k}\right] \\
&= -\,\mathrm{Tr}\left[(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1}\mathbf{y}\mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1} \frac{\partial(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})}{\partial \mu_k}\right] \\
&= -\,\mathrm{Tr}\left[(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1}\mathbf{y}\mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1}\mathbf{K}_k\right] \\
&= -\,\mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1}\mathbf{K}_k(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1}\mathbf{y} = -\boldsymbol{\alpha}^\top\mathbf{K}_k\boldsymbol{\alpha}. \qquad \square
\end{aligned}$$

Proj. Grad. KRR - L_2 Reg.

$\textsc{ProjectionBasedGradientDescent}((\mathbf{K}_k)_{k\in[1,p]}, \boldsymbol{\mu}_0)$

$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_0$

$\boldsymbol{\mu}' \leftarrow \infty$

while $\|\boldsymbol{\mu}' - \boldsymbol{\mu}\| > \epsilon$ **do**

    $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}'$

    $\boldsymbol{\alpha} \leftarrow (\mathbf{K}_{\boldsymbol{\mu}} + \lambda \mathbf{I})^{-1}\mathbf{y}$

    $\boldsymbol{\mu}' \leftarrow \boldsymbol{\mu} + \eta\,(\boldsymbol{\alpha}^\top \mathbf{K}_1 \boldsymbol{\alpha}, \ldots, \boldsymbol{\alpha}^\top \mathbf{K}_p \boldsymbol{\alpha})^\top$

    for $k \leftarrow 1$ **to** p **do**

        $\boldsymbol{\mu}'_k \leftarrow \max(0, \boldsymbol{\mu}'_k)$

    $\boldsymbol{\mu}' \leftarrow \boldsymbol{\mu}_0 + \Lambda \frac{\boldsymbol{\mu}' - \boldsymbol{\mu}_0}{\|\boldsymbol{\mu}' - \boldsymbol{\mu}_0\|}$

return $\boldsymbol{\mu}'$

Interpolated Step KRR - L_2 Reg.

$\textsc{InterpolatedIterativeAlgorithm}((\mathbf{K}_k)_{k\in[1,p]}, \boldsymbol{\mu}_0)$

$\boldsymbol{\alpha} \leftarrow \infty$

$\boldsymbol{\alpha}' \leftarrow (\mathbf{K}_{\boldsymbol{\mu}_0} + \lambda\mathbf{I})^{-1}\mathbf{y}$

while $\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\| > \epsilon$ **do**

$\quad \boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}'$

$\quad \mathbf{v} \leftarrow (\boldsymbol{\alpha}^\top \mathbf{K}_1 \boldsymbol{\alpha}, \ldots, \boldsymbol{\alpha}^\top \mathbf{K}_p \boldsymbol{\alpha})^\top$

$\quad \boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_0 + \Lambda \frac{\mathbf{v}}{\|\mathbf{v}\|}$

$\quad \boldsymbol{\alpha}' \leftarrow \eta\boldsymbol{\alpha} + (1-\eta)(\mathbf{K}_{\boldsymbol{\mu}} + \lambda\mathbf{I})^{-1}\mathbf{y}$

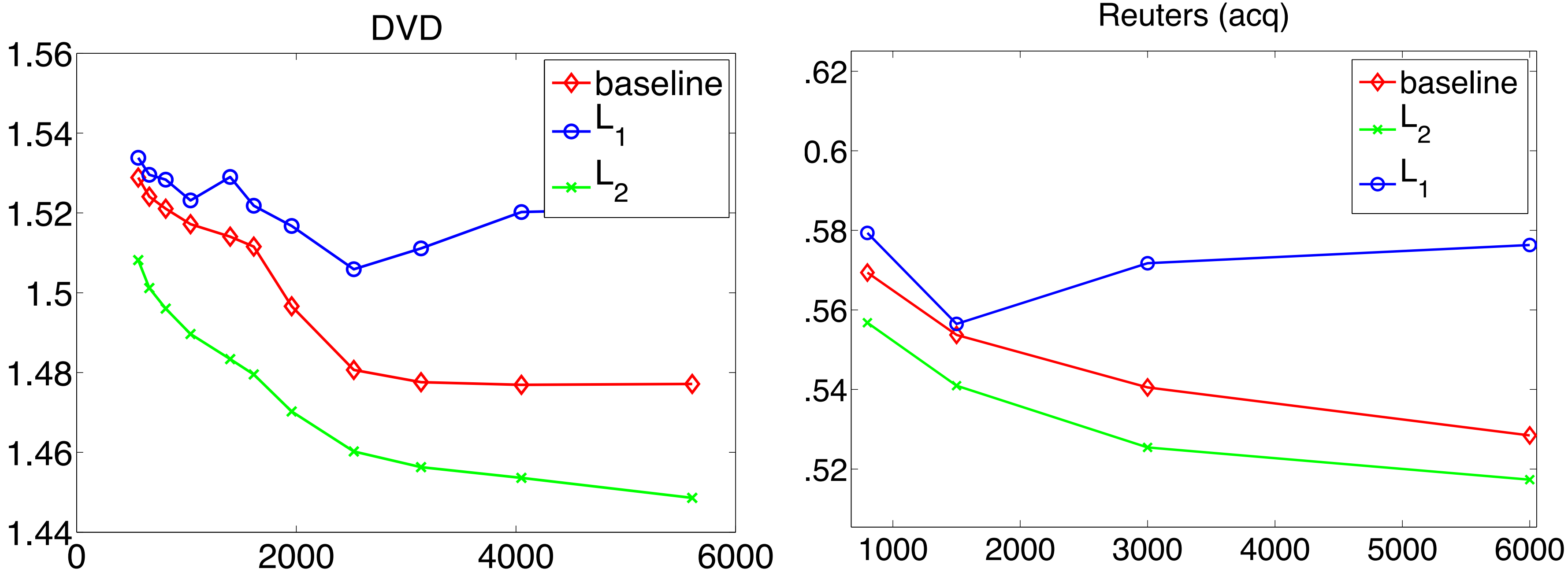
return $\boldsymbol{\alpha}'$

Simple and very efficient: few iterations (less than15).

L_2-Regularized Combinations

(Cortes, MM, and Rostamizadeh, 2009)

- Dense combinations are beneficial when using many kernels.
- Combining kernels based on single features, can be viewed as principled feature weighting.



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

- Bousquet, Olivier and Herrmann, Daniel J. L. On the complexity of learning the kernel matrix. In NIPS, 2002.
- Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local Rademacher complexity. In Proceedings of NIPS, 2013.
- Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Learning non-linear combinations of kernels. In NIPS, 2009.
- 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.
- Corinna Cortes, Mehryar Mohri, Afshin Rostamizadeh. Algorithms for Learning Kernels Based on Centered Alignment. JMLR 13: 795-828, 2012.

References

- Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Tutorial: Learning Kernels. ICML 2011, Bellevue, Washington, July 2011.
- Zakria Hussain, John Shawe-Taylor. Improved Loss Bounds For Multiple Kernel Learning. In AISTATS, 2011 [see arxiv for corrected version].
- Sham M. Kakade, Shai Shalev-Shwartz, Ambuj Tewari: Regularization Techniques for Learning with Matrices. JMLR 13: 1865-1890, 2012.
- 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.
- Lanckriet, Gert, Cristianini, Nello, Bartlett, Peter, Ghaoui, Laurent El, and Jordan, Michael. Learning the kernel matrix with semidefinite programming. JMLR, 5, 2004.

References

- Mehmet Gönen, Ethem Alpaydin: Multiple Kernel Learning Algorithms. JMLR 12: 2211-2268 (2011).
- 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.