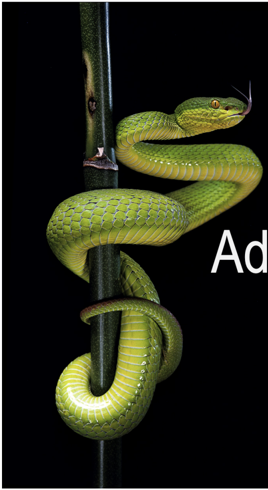



David J. Miller

Zhen Xiang

George Kesidis


# Adversarial Learning and Secure AI

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

# Appendix

## Support-Vector Machines (SVMs)

# Outline

- Background on constrained optimization and duality
- Background on distance to a hyperplane
- Linear SVMs
- Dealing with not linearly separable data
- SVMs with nonlinear kernels
- SVMs for more than two classes
- Single-class SVMs

# Background on constrained optimization and duality

- Consider a *primal* optimization problem with a set of $m$ inequality constraints: Find

$$\arg\min_{\underline{x} \in D} f_0(\underline{x}),$$

where the constrained domain of optimization is

$$D \quad \equiv \quad \{\underline{x} \in \mathbb{R}^n \mid f_i(\underline{x}) \leq 0, \forall\ i \in \{1, 2, ..., m\}\}.$$

- To study the primal problem, we define the corresponding *Lagrangian* function on $\mathbb{R}^n \times [0, \infty)^m$:

$$L(\underline{x}, \underline{v}) \quad \equiv \quad f_0(\underline{x}) + \sum_{i=1}^{m} v_i f_i(\underline{x}),$$

where, by implication, the vector of *Lagrange multipliers* (dual variables) is $\underline{v} \in [0, \infty)^m$, *i.e.*, non-negative $\underline{v} \geq \underline{0}$.

# Primal constrained optimization with Lagrange multipliers

- **Theorem:**

$$\min_{\underline{x}\in\mathbb{R}^n} \max_{\underline{v}\geq\underline{0}} L(\underline{x},\underline{v}) \quad = \quad \min_{\underline{x}\in D} f_0(\underline{x}) \;\equiv\; p^*.$$

- **Proof:** Simply,

$$\max_{\underline{v}\geq\underline{0}} L(\underline{x},\underline{v}) \quad = \quad \begin{cases} \infty & \text{if } \underline{x}\notin D, \\ f_0(\underline{x}) & \text{if } \underline{x}\in D \text{ (see complementary slackness)} \end{cases}$$

$\square$

- Note that if $\underline{x}\notin D$ then $\exists\; i>0$ s.t. $f_i(\underline{x})>0 \Rightarrow$ optimal $v_i^* = \infty$.
- So, we can maximize the Lagrangian in an *unconstrained* fashion to find the solution to the constrained primal problem.

# Complementary slackness of primal solution

- Define the maximizing values of the Lagrange multipliers,

$$\underline{v}^*(\underline{x}) \quad \equiv \quad \arg\max_{\underline{v} \geq \underline{0}} L(\underline{x}, \underline{v})$$

  and note that the *complementary slackness* conditions

$$v_i^*(\underline{x}) f_i(\underline{x}) \quad = \quad 0$$

  hold for all $\underline{x} \in D$ and $i \in \{1, 2, ..., m\}$.
- That is, if there is slackness in the $i^{\text{th}}$ constraint, *i.e.*, $f_i(\underline{x}) < 0$, then there is no slackness in the constraint of the corresponding Lagrange multiplier, *i.e.*, $v_i^*(\underline{x}) = 0$.
- Conversely, if $f_i(\underline{x}) = 0$, then the *optimal* value of the Lagrange multiplier $v_i^*(\underline{x})$ is not relevant to the Lagrangian.

# The dual problem

- Now define the *dual* function of the primal problem:

$$g(\underline{v}) \quad = \quad \min_{\underline{x} \in \mathbb{R}^n} L(\underline{x}, \underline{v}).$$

  *i.e.*, unconstrained optimization w.r.t. *primal* variables first.
- Note that $g(\underline{v})$ may be infinite for some values of $\underline{v}$ and that $g$ is always concave.
- **Theorem:** For all $\underline{x} \in D$ and $\underline{v} \geq \underline{0}$,

$$g(\underline{v}) \quad \leq \quad f_0(\underline{x}).$$

- **Proof:** For $\underline{v} \geq \underline{0}$ and $\underline{x} \in D$,

$$g(\underline{v}) \leq L(\underline{x}, \underline{v}) \leq \max_{\underline{v} \geq \underline{0}} L(\underline{x}, \underline{v}) = f_0(\underline{x}),$$

  where the last equality is the bound on $L$ assuming $\underline{x} \in D$. $\square$

# The dual problem (cont)

- By the previous theorem, if we solve the *dual problem*, *i.e.*, find

$$d^* \equiv \max_{\underline{v} \geq \underline{0}} g(\underline{v}),$$

then we will have obtained a (hopefully good) lower bound to the primal problem, *i.e.*,

$$\max_{\underline{v} \geq \underline{0}} g(\underline{v}) = \max_{\underline{v} \geq \underline{0}} \min_{\underline{x} \in \mathbb{R}^n} L(\underline{x}, \underline{v}) = d^* \leq p^* = \min_{\underline{x} \in \mathbb{R}^n} \max_{\underline{v} \geq \underline{0}} L(\underline{x}, \underline{v}) = \min_{\underline{x} \in D} f_0(\underline{x})$$

- Under certain conditions in this finite dimensional setting, in particular when the primal problem is convex and a *strictly feasible* solution exists, the *duality gap*,

$$p^* - d^* = 0.$$

# The dual problem for a linear program

- If $f_0(\underline{x}) = \sum_{j=1}^{n} \phi_j x_j$ and all $f_i(\underline{x}) = \xi_i + \sum_{j=1}^{n} \gamma_{i,j} x_j$ are linear functions, then the above primal problem, $\min_{\underline{x}} f_0(\underline{x})$ s.t. $f_i(\underline{x}) \leq 0 \; \forall i \geq 1$, is called a Linear Program (LP).
- **Exercise:** Find an *equivalent* dual LP. Hint: first show the Lagrangian of the primal problem can be written as

$$L(\underline{x}, \underline{v}) \;=\; \sum_{i=1}^{m} \xi_i v_i + \sum_{j=1}^{n} x_j \left( \phi_j + \sum_{i=1}^{m} v_i \gamma_{i,j} \right).$$

- LPs can be solved by the simplex algorithm (along feasible region boundaries) or by interior point methods.

# Iterated subgradient method

- Using duality to find $p^*$ and $\underline{x}^* = \text{argmin}_{\underline{x}\in D} f_0(\underline{x})$ in this case, consider a *slow* ascent method is used to maximize $g$,

$$\underline{v}_n = \underline{v}_{n-1} + \alpha_1 \nabla_{\underline{v}} L(\underline{x}^*(\underline{v}_{n-1}), \underline{v}_{n-1}),$$

and between steps of the ascent method, a *fast* descent method used to evaluate $g(\underline{v}_n)$ by minimizing $L(\underline{x}, \underline{v}_n)$,

$$\underline{x}_k = \underline{x}_{k-1} - \alpha_2 \nabla_{\underline{x}} L(\underline{x}_{k-1}, \underline{v}_n) \to \underline{x}^*(\underline{v}_n)$$

- The process described by such an ascent/descent method is called an iterative subgradient method.
- The step sizes $\alpha$ can be chosen dynamically, *e.g.*, steepest ascent/descent (*i.e.*, itself the result of optimization).
- Instead of slow ascent, the descent step can be projected on the feasible domain $D$.

# KKT conditions

- Consider again a *primal* optimization problem with a set of $m$ inequality constraints: Find
$$\arg\min_{\underline{x}\in D} f_0(\underline{x}),$$
where the constrained domain of optimization is
$$D \equiv \{\underline{x}\in\mathbb{R}^n \mid f_i(\underline{x}) \le 0, \forall\ i \in \{1,2,...,m\}\}.$$
- So the Lagrangian on $(\underline{x},\underline{v}) \in \mathbb{R}^n \times (\mathbb{R}^+)^m$ is
$$L(\underline{x},\underline{v}) \equiv f_0(\underline{x}) + \sum_{i=1}^{m} v_i f_i(\underline{x}).$$
and our objective is to find $\min_{\underline{x}} \max_{\underline{v}\ge\underline{0}} L$.
- If $f_0$ is convex and, $\forall i \ge 1$, $f_i$ is linear, then the following Karush-Kuhn-Tucker (KKT) conditions suffice for optimality:
$$\begin{aligned} \forall j, \quad \partial L/\partial x_j &= 0 \text{ and} \\ \forall i, \quad v_i f_i &= 0 \text{ (complementary slackness).} \end{aligned}$$

# Background: Distance to a hyperplane

- For a fixed $n$-dimensional vector $\underline{w} \in \mathbb{R}^n$, $\underline{w} \neq 0$, and scalar $b \in \mathbb{R}$, we can define the hyperplane in $\mathbb{R}^n$,

$$\mathcal{F} \quad = \quad \{\underline{x} \in \mathbb{R}^n \mid 0 = f(\underline{x}) = b + \langle \underline{x}, \underline{w} \rangle = b + \sum_{i=1}^{n} x_i w_i\}.$$

- For $n = 2$, note that the line $x_2 = b + mx_1$ can be written as $0 = b + \langle \underline{x}, \underline{w} \rangle$ where $\underline{w} = (m, -1)^{\mathrm{T}} \perp (1, m)^{\mathrm{T}}$.
- If $\underline{x}, \underline{z} \in \mathcal{F}$ and $\underline{x} \neq \underline{z}$, then the vector $\underline{x} - \underline{z}$ must be parallel to the hyperplane;
- since $0 = f(\underline{x}) - f(\underline{z}) = \langle \underline{w}, \underline{x} - \underline{z} \rangle$, $\underline{w} \perp \mathcal{F}$.
- Equivalently, we can define the hyperplane $\mathcal{F}$ given *any particular* vector $\underline{x}^*$ such that $f(\underline{x}^*) = 0$:

$$\mathcal{F} \quad = \quad \{\underline{x}^* + \underline{v} \mid \underline{v} \perp \underline{w}\}.$$

- One such $\underline{x}^* = -b\ \underline{w}/\|\underline{w}\|^2 \ \perp \mathcal{F}$.

# Background: Distance to a hyperplane

- Recall we can project $\underline{z}$ onto $\underline{w}$ to write

$$\underline{z} = \alpha\frac{\underline{w}}{\|\underline{w}\|} + (\underline{z} - \alpha\frac{\underline{w}}{\|\underline{w}\|}),$$

where the component of $\underline{z}$ in the direction $\underline{w}$ is

$$\alpha \quad = \quad \frac{\langle \underline{z}, \underline{w} \rangle}{\|\underline{w}\|}$$

and $\underline{w}/\|\underline{w}\|$ is a vector of unit (Euclidean) norm.

- The distance from $\underline{z}$ to $\mathcal{F}$ is $|d_{\underline{z}}|$ for a scalar $d_{\underline{z}} \in \mathbb{R}$ such that $\underline{z} - d_{\underline{z}}\underline{w}/\|\underline{w}\| \ \in \mathcal{F}$, *i.e.*,

$$0 = f(\underline{z} - d_{\underline{z}}\underline{w}/\|\underline{w}\|) = f(\underline{z}) - d_{\underline{z}}\|\underline{w}\|.$$

# Background: Distance to a hyperplane

- Thus, $d_{\underline{z}} = f(\underline{z})/\|\underline{w}\|$ and

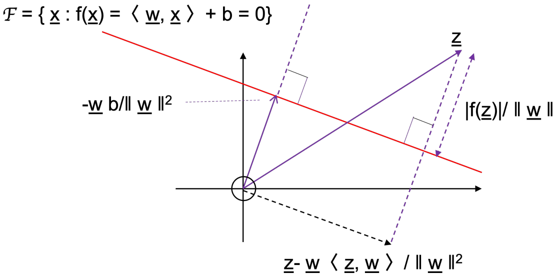
$$|d_{\underline{z}}| \quad = \quad \frac{|f(\underline{z})|}{\|\underline{w}\|}.$$

- If $d_{\underline{z}} > 0$ then $\underline{z}$ is on one side of the hyperplane $\mathcal{F}$, *i.e.*, the half of $\mathbb{R}^n$ where $f > 0$ since

$$0 = f(\underline{z} - d_{\underline{z}}\underline{w}) = f(\underline{z}) - d_{\underline{z}}\|\underline{w}\|.$$

- Else if $d_{\underline{z}} < 0$ then $\underline{z}$ is on the side of the hyperplane $\mathcal{F}$ where $f < 0$.

# Background: Distance to hyperplane (cont)



# Linear SVMs

- Support-vector machines (SVMs) are classifiers $f$ of real vector valued, two-class samples,
- *i.e.*, training set $\mathcal{X} \subset \mathbb{R}^n$ and $|\mathcal{C}| = 2$, where the classes are enumerated
$$\{-1, 1\} = \mathcal{C}.$$
- Ideally, a SVM $f : \mathbb{R}^n \to \mathbb{R}$ has the following separability property on the training set: $\forall \underline{x} \in \mathcal{X}$,
  - if $\underline{x}$'s class label $y_{\underline{x}} = 1$ then $f(\underline{x}) > 0$
  - else (when $y_{\underline{x}} = -1$) $f(\underline{x}) < 0$,
  - *i.e.*, the sign of $f(\underline{x})$ indicates the class $y_{\underline{x}}$.
- So the classifier is actually
$$\text{sgn} \circ f : \mathbb{R}^n \to \mathcal{C} = \{-1, 1\}.$$

# Linear SVMs and Separable Data

- If the data is *linearly separable*, then there is a classifier

$$f(\underline{x}) = \langle \underline{w}, \underline{x} \rangle + b$$

such that

$$y_{\underline{x}} f(\underline{x}) > 0 \text{ for all } \underline{x} \in \mathcal{X},$$

where the scalar $b \in \mathbb{R}$, the vector $\underline{w} \in \mathbb{R}^n$.

- The classification margin of a class-separating classifier $f$ is

$$\min_{\underline{x} \in \mathcal{X}} \frac{|f(\underline{x})|}{\|\underline{w}\|} = \min_{\underline{x} \in \mathcal{X}} \frac{y_{\underline{x}} f(\underline{x})}{\|\underline{w}\|}$$

# Learning Linear SVMs on Separable Data

- To maximize generalization performance, choose $\underline{w}, b$ to solve the following optimization problem:

  $$\min \|\underline{w}\|^2 \quad \text{subject to} \quad y_{\underline{x}} f(\underline{x}) = y_{\underline{x}}(\langle \underline{w}, \underline{x} \rangle + b) \geq 1 \ \ \forall \underline{x} \in \mathcal{X},$$

  where minimizing $\|\underline{w}\|^2$ (easier to differentiate than $\|\underline{w}\|$) is equivalent to maximizing classification *margin*,
- *i.e.*, maximizing the shortest distances to the decision boundary (hyperplane $\mathcal{F} = \{x : f(x) = 0\}$) among points in each class.
- Note that the linear optimization constraints are $\geq 1$ instead of $\geq 0$ so as to "normalize" the resulting the weight vector $\underline{w}$ with respect to the **closest points (support vectors)** to the hyperplane $\mathcal{F}$,

  $$\mathcal{X}^* \subset \mathcal{X}.$$

# Linear SVMs with Separable Data - finding $\underline{w}$, $b$

Define the Lagrangian with Lagrange multipliers $\lambda$:

$$L((\underline{w}, b), \underline{\lambda}) = \frac{1}{2}\|\underline{w}\|^2 - \sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}}(y_{\underline{x}}(\langle \underline{w}, \underline{x}\rangle + b) - 1)$$

Taking the dual approach:

$$\nabla_w L = 0 \Rightarrow \underline{w}^* = \sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}} y_{\underline{x}} \underline{x}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}} y_{\underline{x}} = 0$$

# Linear SVMs with Separable Data - Dual Approach

Thus by substitution we get

$$
\begin{aligned}
L((\underline{w}^*, b^*), \underline{\lambda}) &= \frac{1}{2}\|\sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}} y_{\underline{x}} \underline{x}\|^2 - \sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}} y_{\underline{x}} \langle \sum_{\underline{z}\in\mathcal{X}} \lambda_{\underline{z}} y_{\underline{z}} \underline{z}), \ \underline{x}\rangle \\
&\quad - \sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}} y_{\underline{x}} b + \sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}} \\
&= -\frac{1}{2} \sum_{\underline{x}\in\mathcal{X}} \sum_{\underline{z}\in\mathcal{X}} \lambda_{\underline{x}} \lambda_{\underline{z}} y_{\underline{x}} y_{\underline{z}} \langle \underline{x}, \underline{z} \rangle + \sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}}
\end{aligned}
$$

# Linear SVMs with Separable Data - Dual Approach (cont)

- Now need to maximize $L((\underline{w}^*, b^*), \underline{\lambda})$ over $\underline{\lambda}$, subject to

$$\sum_{\underline{x} \in \mathcal{X}} \lambda_{\underline{x}} y_{\underline{x}} = 0, \quad \lambda_{\underline{x}} \geq 0, \forall \underline{x} \in \mathcal{X}$$

  which is a quadratic program.
- This problem corresponds to solving a set of *linear* equations in the unknowns $\underline{\lambda} \geq \underline{0}$.

# Linear SVMs with Separable Data - Dual Approach (cont)

- Recall complementary slackness,

$$\forall x \in \mathcal{X},\ \lambda_{\underline{x}}(y_{\underline{x}}(\langle \underline{w}^*, \underline{x} \rangle + b^*) - 1) = 0$$

- Thus,

$$\begin{aligned} \lambda_{\underline{x}} = 0 \quad &\Leftrightarrow \quad \underline{x} \notin \mathcal{X}^* \\ \underline{w}^* \quad &= \quad \sum_{\underline{x} \in \mathcal{X}^*} \lambda_{\underline{x}} y_{\underline{x}} \underline{x} \\ b^* \quad &= \quad y_{\underline{x}} - \langle \underline{w}^*, \underline{x} \rangle = y_{\underline{x}} - \sum_{\underline{z} \in \mathcal{X}^*} \lambda_{\underline{z}} y_{\underline{z}} \langle \underline{z}, \underline{x} \rangle \end{aligned}$$

- For robustness to error, average $b^*$ over support vectors:

$$b^* \quad = \quad = \frac{1}{|\mathcal{X}^*|} \sum_{\underline{x} \in \mathcal{X}^*} \left( y_{\underline{x}} - \sum_{\underline{z} \in \mathcal{X}^*} \lambda_{\underline{z}} y_{\underline{z}} \langle \underline{z}, \underline{x} \rangle \right)$$

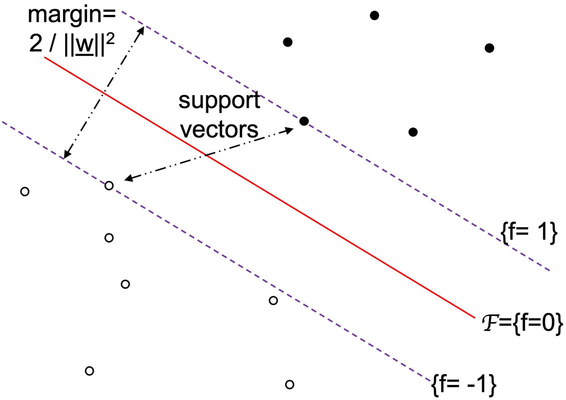
# Inference by linear SVM

- Suppose the class of test sample $\underline{\xi}$ is to be inferred.
- Evaluating the sign of the SVM at a test sample $\underline{\xi}$,

$$\operatorname{sgn}(f(\underline{\xi})) = \operatorname{sgn}\left(\langle \underline{w}^*, \underline{\xi}\rangle + b^*\right) \quad = \quad \operatorname{sgn}\left(\sum_{\underline{x}\in\mathcal{X}^*} \lambda_{\underline{x}} y_{\underline{x}} \langle \underline{x}, \underline{\xi}\rangle + b^*\right)$$

  depends only on inner products of: $\underline{\xi}$ with support vectors, $\langle \underline{\xi}, \underline{x}\rangle$ for $\underline{x} \in \mathcal{X}^*$ and support vectors with each other ($b^*$),
- see the kernel trick.

# Linear SVMs with Separable Data - Illustrative Example




The SVM is the decision boundary that maximizes classification margin for both clases for improved generalization performance.

# Linear SVM with slackness for non-separable data

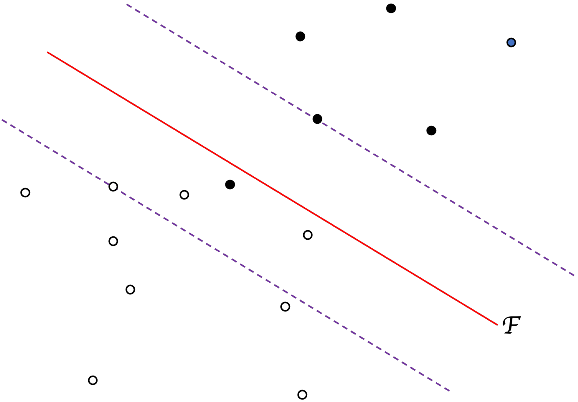
- When the labelled training data $\mathcal{X}$ is not linearly separable, instead:

$$\min \frac{1}{2}\|\underline{w}\|^2 + c\sum_{\underline{x}\in\mathcal{X}} \gamma_{\underline{x}} \text{ such that:}$$

$$y_{\underline{x}} f(\underline{x}) := y_{\underline{x}}(\langle \underline{w}, \underline{x}\rangle + b) \geq 1 - \gamma_{\underline{x}} \quad \forall \underline{x} \in \mathcal{X}.$$

  and optimize over $\underline{w}$, $b$ and $\gamma_{\underline{x}} \geq 0\ \forall \underline{x} \in \mathcal{X}$.
- The scale $c$ and slackness $\gamma$ hyperparameters determine the relative importance of the weight vector magnitude (inverse margin) in order to deal with margin violators including misclassified samples.
- $c, \gamma$ can be found by grid search using a held-out evaluation subset of the training set $\mathcal{X}$ (or training-set fold) to assess the error rate of each $(\gamma, c)$ grid point.

# Linear SVMs with non-separable data




- two white margin violators
- one black sample misclassified

# Kernel SVMs with nonlinear decision boundaries

- Again suppose that the training data $\mathcal{X}$ is not linearly separable, but suppose there is a feature mapping function $\Phi : \mathbb{R}^n \to \mathbb{R}^N$ for $N > n$ (the original number of features) such that $\{(\Phi(\underline{x}), y_{\underline{x}})\}_{\underline{x}\in\mathcal{X}}$ *is* linearly separable.
- Recall Cover's theorem.
- Repeating the above procedure gives the SVM in $\mathbb{R}^N$:

$$f^*(\underline{x}) \;=\; \langle \underline{w}^*, \Phi(\underline{x})\rangle + b^* \;=\; \sum_{\underline{\xi}\in\mathcal{X}^*} \lambda_{\underline{\xi}} y_{\underline{\xi}} K(\underline{\xi}, \underline{x}) + b^*,$$

where for any $\underline{\zeta} \in \mathcal{X}^*$

$$b^* \;=\; y_{\underline{\zeta}} - \sum_{\underline{\xi}\in\mathcal{X}^*} \lambda_{\underline{\xi}} y_{\underline{\xi}} K(\underline{\xi}, \underline{\zeta})$$

and the *kernel* $K$ is given by

$$K(\underline{\xi}, \underline{\zeta}) \;=\; \langle \Phi(\underline{\xi}), \Phi(\underline{\zeta})\rangle.$$

# Kernel SVMs (cont)

- Note that the SVM classifier $f^*$ depends on the feature map $\Phi$ only implicitly through $K$ and $\mathcal{X}^*$.
- Computation of the SVM typically *begins* by selecting the kernel $K$ (and parameter $N$), hoping that the data can be separable with the choice made, and attempting to discover the Lagrange multipliers $\lambda_{\underline{x}}$ (and hence the support vectors $\mathcal{X}^*$) without $\Phi$,
- *i.e.*, the "**kernel trick**".
- The corresponding "weight vector norm squared" (classification margin) in the nonlinear classifier case is

$$\sum_{\underline{\xi},\underline{\zeta}\in\mathcal{X}} \lambda_{\underline{\xi}}\lambda_{\underline{\zeta}}y_{\underline{\xi}}y_{\underline{\zeta}}K(\underline{\xi},\underline{\zeta})$$

clearly generalizing $\|\underline{w}^*\|^2$ for a linear SVM.

# Kernel SVMs - optimization

The optimization objective is:

$$\frac{1}{2}\sum_{\underline{\xi},\underline{\zeta}\in\mathcal{X}} \lambda_{\underline{\xi}}\lambda_{\underline{\zeta}}y_{\underline{\xi}}y_{\underline{\zeta}}K(\underline{\xi},\underline{\zeta}) + c\sum_{\underline{x}\in\mathcal{X}}\gamma_{\underline{x}} \quad \text{subject to:}$$

$$1-\lambda_{\underline{x}} \leq y_{\underline{x}}(\langle \underline{w}, \Phi(\underline{x})\rangle + b), \quad \lambda_{\underline{x}} \geq 0 \quad \forall \underline{x} \in \mathcal{X}$$

with class decision function

$$f(\underline{z}) = \sum_{\underline{x}\in\mathcal{X}} \lambda_{\underline{x}} y_{\underline{x}} K(\underline{z},\underline{x}) + b$$

where

- the **kernel** $K(\underline{z},\underline{x}) = \langle \Phi(\underline{z}), \Phi(\underline{x})\rangle$, and
- $\lambda_{\underline{x}} > 0 \Rightarrow \underline{x} \in \mathcal{X}^*$.

# Kernel SVMs - Examples

- Suppose $\Phi(x)$ contains all polynomial components that can be created with the components of $\underline{x}$ of degree $\leq 2$:
  - Specifically, if $x = (x_1, x_2, x_3)^{\mathrm{T}}$ and

    $$\Phi(\underline{x}) = (1, x_1\sqrt{2}, x_2\sqrt{2}, x_3\sqrt{2}, x_1^2, x_2^2, x_3^2, x_1x_2\sqrt{2}, x_2x_3\sqrt{2}, x_3x_1\sqrt{2})^{\mathrm{T}}$$

  - then

    $$K(\underline{x}, \underline{z}) = \langle \Phi(\underline{x}), \Phi(\underline{z}) = (1 + \langle \underline{x}, \underline{z} \rangle \rangle)^2.$$

- Gaussian radial basis functions are commonly used:

  $$K(\underline{z}, \underline{x}) = \exp\left(-\frac{\|\underline{z} - \underline{x}\|^2}{2\sigma^2}\right)$$

- Note that we don't need to know $\Phi$ to train the SVM or make inferences!

# Feature selection using SVMs

- Many features may confound or may simply not be useful for purposes of classification.
- Promote sparsity in the weights by adding a penalty term to the Lagrangian:

$$\|w\|_q^q = \sum_i |w_i|^q \quad \text{for } 0 < q \ll 1,$$

*i.e.*, penalize nonzero weights ($q \ll 1$) while preserving differentiability ($q > 0$).

- Recursive methods have been proposed to reduce features used for classification (make SVM weights sparser).
- Eliminate features associated with smallest-magnitude weight-vector components (RFE).
- Margin-based Feature Elimination (MFE) removes the feature which results in the largest margin.

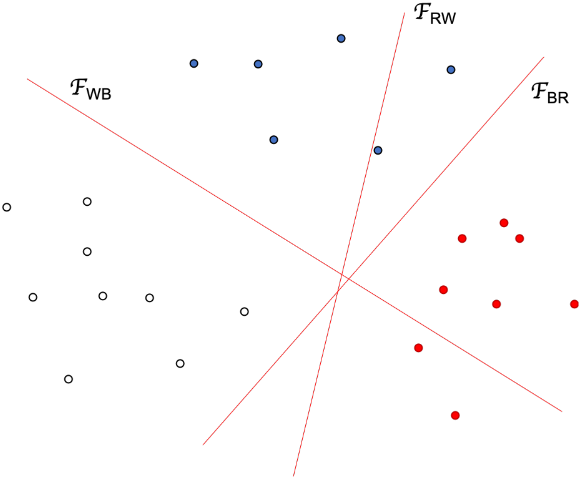
# SVMs for more than two classes

- Assume $|\mathcal{C}| \geq 2$ classes.
- Suppose we find $|\mathcal{C}|(|\mathcal{C}|-1)/2$ SVMs $f_{y,y'}$, one for each different pair of classes $y' \neq y \in \mathcal{C}$.
- For example, we can then create a multiclass SVM by labelling test samples $\underline{x}$ as follows,

$$y_{\underline{x}} \quad = \quad \arg\max_{y} \sum_{y' \in \mathcal{C},\ y' \neq y} \mathbf{1}\{f_{y,y'}(\underline{x}) \text{ decides } y\},$$

  *i.e.*, each SVM "votes" for a class label for $\underline{x}$.
- Alternatively, we could find just $|\mathcal{C}|$ SVMs each between a class $y$ and the rest of classes $\bar{y} := \mathcal{C} \backslash y = \{y' \in \mathcal{C} | y' \neq y\}$ (*i.e.*, lump together the rest of the classes giving "one versus rest").
- Ideally here, select class $y_{\underline{x}}$ for unlabelled sample $\underline{x}$ if $f_{y_{\underline{x}},\bar{y}_{\underline{x}}}$ decides $y_{\underline{x}}$ for $\underline{x}$ and $\forall y \in \mathcal{C} \backslash y_{\underline{x}}$, $f_{y,\bar{y}}$ decides $\bar{y}$ for $\underline{x}$.

# Pairwise SVMs for three linearly separable classes




For red samples, two SVMs vote "red" (R) while $\mathcal{F}_{\mathrm{WB}}$ votes either "black" (B) or "white" (W), so decide the "red" class.

# SVMs for more than two classes (cont)

- One may choose the class label for $\underline{x}$ with corresponding largest *distance* to the class-decision boundary (as $|f(\underline{x})|/\|w\|^2$ for linear SVMs).
- For the set of classes $\mathcal{C}(x)$ which have garnered votes for test sample $x$, let $y^*(x) \in \mathcal{C}(x)$ with largest distance to classification boundary, where the distance to be the class-decision boundary is $d_{y^*}(x)$.
- Can take the quantity

$$1 - \max_{y \in \mathcal{C}(x) \setminus y^*(x)} d_y(x) \Big/ d_{y^*}(x)$$

as the "confidence" in class decision-making for $x$.

# One-class SVMs

- Idea is to train the SVM to return 1 for a small region containing the training samples, otherwise return -1.
- In one approach (Scholkopf):

$$\min_{\underline{w},\gamma,\rho} \frac{\|\underline{w}\|^2}{2} + \frac{1}{\nu n}\sum_{\underline{x}\in\mathcal{X}} \gamma_{\underline{x}} - \rho \quad \text{subject to:}$$

$$\langle \underline{w}, \Phi(\underline{x})\rangle \geq \rho - \gamma_{\underline{x}},\ \ \gamma_{\underline{x}} \leq 0 \ \ \forall \underline{x} \in \mathcal{X}$$

- The parameter $\nu$ sets:
  - an upper bound on the fraction of outliers (misclassified training samples), and
  - a lower bound on the number of training samples used as support vectors.

# One-class SVMs

- In another approach (Tax and Duin):

$$\min_{R,\underline{z}} \; R^2 + c \sum_{\underline{x} \in \mathcal{X}} \gamma_{\underline{x}} \quad \text{subject to:}$$

$$\|\underline{x} - \underline{z}\|^2 \leq R^2 + \gamma_{\underline{x}}, \;\; \gamma_{\underline{x}} \leq 0 \;\; \forall \underline{x} \in \mathcal{X}$$

- Here the decision boundary is a sphere according to the norm $\| \cdot \|$.