Support vector machines

CS 446 / ECE 449

2022-01-27 15:49:20 -0600 (906ddb6)
Another algorithm for linear prediction. Why?!
Support vector machines (SVMs) have three purposes for us.

1. Demonstrate maximum margin predictors, an example of “low complexity models”, which appear throughout machine learning (not just linear predictors).
2. Demonstrate nonlinear kernels, also pervasive.
3. Exercise convex optimization and duality.
Plan for SVM

- Hard-margin SVM.
- Soft-margin SVM.
- SVM duality.
- Nonlinear SVM: kernels
Maximum margin linear separators

Which linear separator is best?
Maximum margin linear separators

Which linear separator is best?

Which linear separator is best?
The max margin separator is one choice for a good predictor. It is not always the best idea. Recall from lecture 3:

Even so, the maximum margin concept is pervasive in machine learning.
Input $x \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $w \in \mathbb{R}^d$ with $H := \{z \in \mathbb{R}^d : w^T z = 0\}$. 

Formulation #1. 

- Single margin $y x^T w / \|w\|_2$.
- Overall margin $\min_i y_i x_i^T w / \|w\|_2$.
- Max margin $\max \|u\|_2 \min_i y_i x_i^T u$.

Formulation #2. Consider any $v$ with $\min_i y_i x_i^T v \geq 1$. Since margin scales with $1 / \|v\|_2$, choose $\min 1 / 2 \|v\|_2^2$ subject to $v \in \mathbb{R}^d$, $y_i x_i^T v \geq 1 \forall i$.

These two are equivalent (up to scaling).
How to write “maximum margin classifier”?

Input $x \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $w \in \mathbb{R}^d$ with $H := \{z \in \mathbb{R}^d : w^Tz = 0\}$.

**Formulation #1.**

$y^x T w \|w\|_2$
How to write “maximum margin classifier”?

Input $x \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $w \in \mathbb{R}^d$ with $H := \{z \in \mathbb{R}^d : w^T z = 0\}$.

Formulation #1.
Single margin $y x^T w / \|w\|_2$.
Overall margin $\min_i y_i x_i^T w / \|w\|_2$.
Max margin $\max \|u\|_2 = 1 \min_i y_i x_i^T u$. 

Formulation #2.
Consider any $v$ with $\min_i y_i x_i^T v \geq 1$.
Since margin scales with $1 / \|v\|_2$, choose $\min 1 / 2 \|v\|_2$ subject to $v \in \mathbb{R}^d$, $y_i x_i^T v \geq 1 \forall i$.
These two are equivalent (up to scaling).
How to write “maximum margin classifier”?

Input $x \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $w \in \mathbb{R}^d$ with $H := \{z \in \mathbb{R}^d : w^T z = 0\}$.

**Formulation #1.**
- Single margin $y x^T w / \|w\|_2$.
- Overall margin $\min_i y_i x_i^T w / \|w\|$.
- Max margin $\max_{\|u\| = 1} \min_i y_i x_i^T u$.

**Formulation #2.**

Consider any $v$ with $\min_i y_i x_i^T v \geq 1$.
Since margin scales with $\frac{1}{\|v\|}$, choose $\min \frac{1}{2} \|v\|_2$ subject to $v \in \mathbb{R}^d$, $y_i x_i^T v \geq 1 \forall i$.
These two are equivalent (up to scaling).
How to write “maximum margin classifier”?

Input \( x \in \mathbb{R}^d \), label \( y \in \{\pm 1\} \), predictor \( w \in \mathbb{R}^d \) with \( H := \{ z \in \mathbb{R}^d : w^T z = 0 \} \).

**Formulation #1.**
Single margin \( \frac{y x^T w}{\|w\|} \).
Overall margin \( \min_i y_i x_i^T w \div \|w\| \).
Max margin \( \max_{\|u\|=1} \min_i y_i x_i^T u \).

**Formulation #2.**
Consider any \( v \) with \( \min_i y_i x_i^T v \geq 1 \).
Since margin scales with \( \frac{1}{\|v\|} \), choose

\[
\min \frac{1}{2} \|v\|^2 \\
\text{subject to} \quad v \in \mathbb{R}^d, \quad y_i x_i^T v \geq 1 \quad \forall i.
\]
How to write “maximum margin classifier”? 

Input \( x \in \mathbb{R}^d \), label \( y \in \{\pm 1\} \), predictor \( w \in \mathbb{R}^d \) with \( H := \{ z \in \mathbb{R}^d : w^T z = 0 \} \).

**Formulation #1.**

Single margin \( \frac{y x^T w}{\|w\|} \).

Overall margin \( \min_i \frac{y_i x_i^T w}{\|w\|} \).

Max margin \( \max_{\|u\| = 1} \min_i y_i x_i^T u \).

**Formulation #2.**

Consider any \( v \) with \( \min_i y_i x_i^T v \geq 1 \).

Since margin scales with \( \frac{1}{\|v\|} \), choose \( \min \frac{1}{2} \|v\|^2 \)

subject to \( v \in \mathbb{R}^d \), \( y_i x_i^T v \geq 1 \ \forall i \).

These two are equivalent (up to scaling).
Take the solution to either optimization problem:

$$\begin{align*}
\max & \quad \min_i y_i x_i^T u, \\
\text{subject to} & \quad u \in \mathbb{R}^d, \quad \|u\| = 1;
\end{align*}$$

$$\begin{align*}
\min & \quad \frac{1}{2} \|v\|^2 \\
\text{subject to} & \quad v \in \mathbb{R}^d, \quad y_i x_i^T v \geq 1 \quad \forall i.
\end{align*}$$

Remarks.

▶ Since the second is a convex program, many approaches exist.

Homework will investigate a simple SGD-based strategy.

▶ What happens if the second formulation is infeasible? (That is, what if no vector $v$ satisfies $y_i x_i^T v \geq 1$?)
Hard-margin SVM.

Take the solution to either optimization problem:

\[
\begin{align*}
\max & \quad \min_i y_i x_i^T u, \\
\text{subject to} & \quad u \in \mathbb{R}^d, \\
& \quad \|u\| = 1;
\end{align*}
\]

\[
\begin{align*}
\min & \quad \frac{1}{2} \|v\|^2 \\
\text{subject to} & \quad v \in \mathbb{R}^d, \\
& \quad y_i x_i^T v \geq 1 \quad \forall i.
\end{align*}
\]

Remarks.

▶ Since the second is a convex program, many approaches exist. Homework will investigate a simple SGD-based strategy.

▶ What happens if the second formulation is infeasible? (That is, what if no vector \(v\) satisfies \(\min_i y_i x_i^T v \geq 1\)?)
What is the max margin predictor for the following data?
What is the max margin predictor for the following data?
What is the max margin predictor for the following data?

Idea. pay a price for each $y_i x_i^T v < 1$ with slack variables $(\xi_i)_{i=1}^n$: 
Soft-margin SVM

What is the max margin predictor for the following data?

**Idea.** pay a price for each $y_i x_i^T v < 1$ with slack variables $(\xi_i)_{i=1}^n$:

$$
\begin{align*}
\min_{w \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} & \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i \\
\text{s.t.} & \quad y_i x_i^T w \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n, \\
& \quad \xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, n.
\end{align*}
$$
Re-formulation as regularized ERM.

Formulation with slack variables \((\xi_i)_{i=1}^n\).

\[
\begin{align*}
\min_{w \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} & \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i \\
\text{s.t.} & \quad y_i x_i^T w \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n, \\
& \quad \xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, n.
\end{align*}
\]
Re-formulation as regularized ERM.

Formulation with slack variables \((\xi_i)_{i=1}^n\).

\[
\begin{align*}
\min_{w \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} & \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i \\
\text{s.t.} & \quad y_i x_i^T w \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n, \\
& \quad \xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, n.
\end{align*}
\]

Regularized ERM formulation.
Given any \(w\), choose \(x_i := \max\{0, 1 - y_i x_i^T w\}\), whereby

\[
\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \max\left\{0, 1 - y_i x_i^T w\right\},
\]

where \(\ell_{\text{hinge}}(y_i x_i^T w) := \max\{0, 1 - y_i x_i^T w\}\) is the hinge loss.
Re-formulation as regularized ERM.

Formulation with slack variables \((\xi_i)_{i=1}^n\).

\[
\min_{w \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i \\
\text{s.t.} \quad y_i x_i^T w \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n, \\
\xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, n.
\]

Regularized ERM formulation.
Given any \(w\), choose \(x_i := \max\{0, 1 - y_i x_i^T w\}\), whereby

\[
\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \max\{0, 1 - y_i x_i^T w\},
\]

where \(\ell_{\text{hinge}}(y_i x_i^T w) := \max\{0, 1 - y_i x_i^T w\}\) is the hinge loss.

Remarks.

- Normally we’d write \(\frac{1}{n} \sum_{i=1}^n \ell_{\text{hinge}}(y_i x_i^T w) + \frac{\lambda}{2} \|w\|^2\).
- \(C\) (or \(\lambda\)) is a hyper-parameter; it has no good search procedure.
A convex program is an optimization problem (minimization or maximization) where a convex objective is minimized over a convex constraint (feasible) set. Every convex program has a corresponding dual program.

- Clarifies the role of support vectors.
- Leads to a nice nonlinear approach via kernels.
- Gives another choice for optimization algorithms.
A convex program is an optimization problem (minimization or maximization) where a convex objective is minimized over a convex constraint (feasible) set.
A **convex program** is an optimization problem (minimization or maximization) where a convex objective is minimized over a convex constraint (feasible) set.

Every convex program has a corresponding **dual program**. For the SVM, the dual has many nice properties:

- Clarifies the role of **support vectors**.
- Leads to a nice nonlinear approach via **kernels**.
- Gives another choice for optimization algorithms.
**SVM hard-margin duality.**
Define the two optimization problems

\[
\min \left\{ \frac{1}{2} \|w\|^2 : w \in \mathbb{R}^d, \forall i \cdot 1 - y_i x_i w \leq 0 \right\} \quad \text{(primal)},
\]

\[
\max \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j : \alpha \in \mathbb{R}^n, \alpha \geq 0 \right\} \quad \text{(dual)}.
\]

If the primal is feasible, then primal optimal value = dual optimal value.
Given a primal optimum \(\tilde{w}\) and a dual optimum \(\tilde{\alpha}\), then

\[
\tilde{w} = \sum_{i=1}^{n} \tilde{\alpha}_i y_i x_i.
\]
**SVM hard-margin duality.**

Define the two optimization problems

\[
\begin{align*}
\min & \left\{ \frac{1}{2}\|w\|^2 : w \in \mathbb{R}^d, \forall i \cdot 1 - y_i x_i w \leq 0 \right\} \quad \text{(primal)}, \\
\max & \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j : \alpha \in \mathbb{R}^n, \alpha \geq 0 \right\} \quad \text{(dual)}.
\end{align*}
\]

If the primal is feasible, then primal optimal value = dual optimal value.

Given a primal optimum \( \bar{w} \) and a dual optimum \( \bar{\alpha} \), then

\[
\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i.
\]

- The dual variables \( \alpha \) have dimension \( n \), same as examples.
- We can write the primal optimum as a linear combination of examples.
- The dual objective is a **concave quadratic**.
- We will derive this duality using Lagrange multipliers.
Lagrange multipliers

Move constraints to objective using Lagrange multipliers.

Original problem:

\[
\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|_2^2 \\
\text{s.t.} \quad 1 - y_i x_i^T w \leq 0 \quad \text{for all } i = 1, \ldots, n.
\]

- For each constraint \(1 - y_i x_i^T w \leq 0\), associate a **dual variable** (Lagrange multiplier) \(\alpha_i \geq 0\).

- Move constraints to objective by adding \(\sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w)\) and maximizing over \(\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n\) s.t. \(\alpha \geq 0\) (i.e., \(\alpha_i \geq 0\) for all \(i\)).
Lagrange multipliers

Move constraints to objective using Lagrange multipliers.

Original problem: \[
\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|_2^2 \\
\text{s.t.} \quad 1 - y_i x_i^T w \leq 0 \quad \text{for all } i = 1, \ldots, n.
\]

▶ For each constraint \(1 - y_i x_i^T w \leq 0\), associate a dual variable (Lagrange multiplier) \(\alpha_i \geq 0\).

▶ Move constraints to objective by adding \(\sum_{i=1}^n \alpha_i (1 - y_i x_i^T w)\) and maximizing over \(\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \text{ s.t. } \alpha \geq 0\) (i.e., \(\alpha_i \geq 0\) for all \(i\)).

Lagrangian \(L(w, \alpha)\):

\[
L(w, \alpha) := \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^n \alpha_i (1 - y_i x_i^T w).
\]

Maximizing over \(\alpha \geq 0\) recovers primal problem: for any \(w \in \mathbb{R}^d\),

\[
P(w) := \sup_{\alpha \geq 0} L(w, \alpha) = \begin{cases} 
\frac{1}{2} \|w\|_2^2 & \text{if } \min_i y_i x_i^T w \geq 1, \\
\infty & \text{otherwise}.
\end{cases}
\]
Lagrange multipliers

Move constraints to objective using Lagrange multipliers.

Original problem: \[
\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|_2^2 \\
\text{s.t.} \quad 1 - y_i x_i^T w \leq 0 \quad \text{for all } i = 1, \ldots, n.
\]

- For each constraint \(1 - y_i x_i^T w \leq 0\), associate a dual variable (Lagrange multiplier) \(\alpha_i \geq 0\).
- Move constraints to objective by adding \(\sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w)\) and maximizing over \(\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \text{ s.t. } \alpha \geq 0\) (i.e., \(\alpha_i \geq 0\) for all \(i\)).

Lagrangian \(L(w, \alpha)\):

\[
L(w, \alpha) := \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w).
\]

Maximizing over \(\alpha \geq 0\) recovers primal problem: for any \(w \in \mathbb{R}^d\),

\[
P(w) := \sup_{\alpha \geq 0} L(w, \alpha) = \begin{cases} 
\frac{1}{2} \|w\|_2^2 & \text{if } \min_i y_i x_i^T w \geq 1, \\
\infty & \text{otherwise}.
\end{cases}
\]

What if we leave \(\alpha\) fixed, and minimize \(w\)?
Dual problem

Lagrangian

\[ L(w, \alpha) := \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w). \]

Primal hard-margin SVM

\[ P(w) = \sup_{\alpha \geq 0} L(w, \alpha) = \sup_{\alpha \geq 0} \left[ \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \right]. \]
Dual problem

Lagrangian

\[ L(w, \alpha) := \frac{1}{2} \| w \|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w). \]

Primal hard-margin SVM

\[ P(w) = \sup_{\alpha \geq 0} L(w, \alpha) = \sup_{\alpha \geq 0} \left[ \frac{1}{2} \| w \|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \right]. \]

Dual problem \( D(\alpha) = \min_w L(w, \alpha) \): given \( \alpha \geq 0 \), then \( w \mapsto L(w, \alpha) \) is a convex quadratic with minimum \( w = \sum_{i=1}^{n} \alpha_i y_i x_i \), giving

\[ D(\alpha) = \min_{w \in \mathbb{R}^d} L(w, \alpha) = L \left( \sum_{i=1}^{n} \alpha_i y_i x_i, \alpha \right) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|_2^2 = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j. \]
Summarizing,

\[ L(w, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \]  

\[ P(w) = \max_{\alpha \geq 0} L(w, \alpha) \]  

\[ D(\alpha) = \min_w L(w, \alpha) \]

Lagrangian, primal problem, dual problem.
Summarizing,

\[ L(w, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \]  

Lagrangian,

\[ P(w) = \max_{\alpha \geq 0} L(w, \alpha) \]  

primal problem,

\[ D(\alpha) = \min_w L(w, \alpha) \]  

dual problem.

> For general Lagrangians, have weak duality

\[ P(w) \geq D(\alpha), \]

since \( P(w) = \max_{\alpha' \geq 0} L(w, \alpha') \geq L(w, \alpha) \geq \min_{w'} L(w', \alpha) = D(\alpha). \]
Summarizing,

\[ L(w, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \]  

Lagrangian,

\[ P(w) = \max_{\alpha \geq 0} L(w, \alpha) \]  

primal problem,

\[ D(\alpha) = \min_w L(w, \alpha) \]  

dual problem.

- For general Lagrangians, have weak duality

\[ P(w) \geq D(\alpha), \]

since \( P(w) = \max_{\alpha' \geq 0} L(w, \alpha') \geq L(w, \alpha) \geq \min_{w'} L(w', \alpha) = D(\alpha). \)

- By convexity, have strong duality \( \min_w P(w) = \max_{\alpha \geq 0} D(\alpha), \) and an optimum \( \bar{\alpha} \) for \( D \) gives an optimum \( \bar{w} \) for \( P \) via

\[ \bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i = \arg \min_w L(w, \bar{\alpha}). \]
Support vectors

Optimal solutions $\tilde{w}$ and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)$ satisfy

$\begin{align*}
\tilde{w} &= \sum_{i=1}^{n} \tilde{\alpha}_i y_i x_i = \sum_{i: \tilde{\alpha}_i > 0} \tilde{\alpha}_i y_i x_i, \\
\tilde{\alpha}_i > 0 \Rightarrow y_i x_i^T \tilde{w} &= 1 \text{ for all } i = 1, \ldots, n \text{ (complementary slackness).}
\end{align*}$
Optimal solutions $\tilde{w}$ and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)$ satisfy

\[ \tilde{w} = \sum_{i=1}^{n} \tilde{\alpha}_i y_i x_i = \sum_{i: \tilde{\alpha}_i > 0} \tilde{\alpha}_i y_i x_i, \]

\[ \tilde{\alpha}_i > 0 \Rightarrow y_i x_i^T \tilde{w} = 1 \text{ for all } i = 1, \ldots, n \] (complementary slackness).

The $y_i x_i$ where $\tilde{\alpha}_i > 0$ are called support vectors.
Support vectors

Optimal solutions $\bar{w}$ and $\bar{\alpha} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_n)$ satisfy

$\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i = \sum_{i: \bar{\alpha}_i > 0} \bar{\alpha}_i y_i x_i,$

$\bar{\alpha}_i > 0 \Rightarrow y_i x_i^T \bar{w} = 1$ for all $i = 1, \ldots, n$ (complementary slackness).

The $y_i x_i$ where $\bar{\alpha}_i > 0$ are called support vectors.

Support vector examples satisfy “margin” constraints with equality.
Optimal solutions $\tilde{w}$ and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)$ satisfy

$\tilde{w} = \sum_{i=1}^{n} \tilde{\alpha}_i y_i x_i = \sum_{i: \tilde{\alpha}_i > 0} \tilde{\alpha}_i y_i x_i,$

$\tilde{\alpha}_i > 0 \implies y_i x_i^T \tilde{w} = 1$ for all $i = 1, \ldots, n$ (complementary slackness).

The $y_i x_i$ where $\tilde{\alpha}_i > 0$ are called support vectors.

- Support vector examples satisfy “margin” constraints with equality.
- Get same solution if non-support vectors deleted.
Support vectors

Optimal solutions \( \bar{w} \) and \( \bar{\alpha} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_n) \) satisfy

\[
\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i = \sum_{i: \bar{\alpha}_i > 0} \bar{\alpha}_i y_i x_i,
\]

\[
\bar{\alpha}_i > 0 \implies y_i x_i^T \bar{w} = 1 \text{ for all } i = 1, \ldots, n \text{ (complementary slackness)}.
\]

The \( y_i x_i \) where \( \bar{\alpha}_i > 0 \) are called support vectors.

- Support vector examples satisfy “margin” constraints with equality.
- Get same solution if non-support vectors deleted.
- Primal optimum is a linear combination of support vectors.
Proof of complementary slackness

For the optimal (feasible) solutions $\bar{w}$ and $\bar{\alpha}$, we have

\[ P(\bar{w}) = D(\bar{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \bar{\alpha}) \quad \text{(by strong duality)} \]
Proof of complementary slackness

For the optimal (feasible) solutions \( \tilde{w} \) and \( \tilde{\alpha} \), we have

\[
P(\tilde{w}) = D(\tilde{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \tilde{\alpha}) \quad \text{(by strong duality)}
\]

\[
\leq L(\tilde{w}, \tilde{\alpha})
\]
Proof of complementary slackness

For the optimal (feasible) solutions \( \bar{w} \) and \( \bar{\alpha} \), we have

\[
P(\bar{w}) = D(\bar{\alpha}) = \min_{\w \in \mathbb{R}^d} L(\w, \bar{\alpha}) \quad \text{(by strong duality)}
\]
\[
\leq L(\bar{w}, \bar{\alpha})
\]
\[
= \frac{1}{2} \| \bar{w} \|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i \x_i^T \bar{w})
\]
Proof of complementary slackness

For the optimal (feasible) solutions \( \bar{w} \) and \( \bar{\alpha} \), we have

\[
P(\bar{w}) = D(\bar{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \bar{\alpha}) \quad \text{(by strong duality)}
\]

\[
\leq L(\bar{w}, \bar{\alpha})
\]

\[
= \frac{1}{2} \|\bar{w}\|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i x_i^T \bar{w})
\]

\[
\leq \frac{1}{2} \|\bar{w}\|_2^2 \quad \text{(constraints are satisfied)}
\]

\[
= P(\bar{w}).
\]
Proof of complementary slackness

For the optimal (feasible) solutions \( \bar{w} \) and \( \bar{\alpha} \), we have

\[
P(\bar{w}) = D(\bar{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \bar{\alpha}) \quad \text{(by strong duality)}
\]

\[
\leq L(\bar{w}, \bar{\alpha})
\]

\[
= \frac{1}{2} \|\bar{w}\|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i x_i^T \bar{w})
\]

\[
\leq \frac{1}{2} \|\bar{w}\|_2^2 \quad \text{(constraints are satisfied)}
\]

\[
= P(\bar{w}).
\]

Therefore, every term in sum \( \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i x_i^T \bar{w}) \) must be zero:

\[
\bar{\alpha}_i (1 - y_i x_i^T \bar{w}) = 0 \quad \text{for all } i = 1, \ldots, n.
\]
Proof of complementary slackness

For the optimal (feasible) solutions $\bar{w}$ and $\bar{\alpha}$, we have

$$P(\bar{w}) = D(\bar{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \bar{\alpha}) \quad \text{(by strong duality)}$$

$$\leq L(\bar{w}, \bar{\alpha})$$

$$= \frac{1}{2} \|\bar{w}\|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i x_i^T \bar{w})$$

$$\leq \frac{1}{2} \|\bar{w}\|_2^2 \quad \text{(constraints are satisfied)}$$

$$= P(\bar{w}).$$

Therefore, every term in sum $\sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i x_i^T \bar{w})$ must be zero:

$$\bar{\alpha}_i (1 - y_i x_i^T \bar{w}) = 0 \quad \text{for all } i = 1, \ldots, n.$$

If $\bar{\alpha}_i > 0$, then must have $1 - y_i x_i^T \bar{w} = 0$. (Not iff!)
Lagrangian

\[ L(w, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w). \]

Primal maximum margin problem was

\[ P(w) = \sup_{\alpha \geq 0} L(w, \alpha) = \sup_{\alpha \geq 0} \left[ \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \right]. \]

Dual problem

\[ D(\alpha) = \min_{w \in \mathbb{R}^d} L(w, \alpha) = \sum_{i=1}^{n} \alpha_i y_i x_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|_2^2. \]

Given dual optimum \( \bar{\alpha} \),

- Corresponding primal optimum \( \bar{w} = \sum_{i=1}^{n} \alpha_i y_i x_i \);
- Strong duality \( P(\bar{w}) = D(\bar{\alpha}) \);
- \( \bar{\alpha}_i > 0 \) implies \( y_i x_i^T \bar{w} = 1 \),
and these \( y_i x_i \) are support vectors.
SVM soft-margin dual

Similarly, $L(w, \xi, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i x_i^T w)$ (Lagrangian), $P(w, \xi) = \sup_{\alpha \geq 0} L(w, \xi, \alpha)$ (Primal), $D(\alpha) = \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n \geq 0} L(w, \xi, \alpha)$ (Dual).

Remarks. ▶ Dual solution $\bar{\alpha}$ still gives primal solution $\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i$.
▶ Can take $C \to \infty$ to recover hard-margin case.
▶ Dual is still a constrained concave quadratic (used in many solvers).
▶ Some presentations include bias in primal ($x_i^T w + b$); this introduces a constraint $\sum_{i=1}^{n} \alpha_i y_i = 0$ in dual.
Similarly,

\[
L(w, \xi, \alpha) = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i x_i^T w) \quad \text{(Lagrangian)},
\]

\[
P(w, \xi) = \sup_{\alpha \geq 0} L(w, \xi, \alpha) \quad \text{(Primal)},
\]

\[
P(w, \xi, \alpha) = \begin{cases} 
\frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i & \forall i, 1 - \xi_i - y_i x_i^T w \leq 0, \\
\infty & \text{otherwise},
\end{cases}
\]

\[
D(\alpha) = \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n_{\geq 0}} L(w, \xi, \alpha) \quad \text{(Dual)},
\]

\[
D(\alpha) = \max_{\alpha \in \mathbb{R}^n_{0 \leq \alpha_i \leq C}} \left[ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i \right]^2.
\]
SVM soft-margin dual

Similarly,

\[ L(w, \xi, \alpha) = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i x_i^T w) \quad \text{(Lagrangian)}, \]

\[ P(w, \xi) = \sup_{\alpha \geq 0} L(w, \xi, \alpha) \quad \text{(Primal)}, \]

\[ = \begin{cases} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i & \forall i, 1 - \xi_i - y_i x_i^T w \leq 0, \\ \infty & \text{otherwise}, \end{cases} \]

\[ D(\alpha) = \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n \geq 0} L(w, \xi, \alpha) \quad \text{(Dual)}, \]

\[ = \max_{0 \leq \alpha_i \leq C} \left( \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 \right). \]

Remarks.

- Dual solution \( \bar{\alpha} \) still gives primal solution \( \bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i \).
- Can take \( C \to \infty \) to recover hard-margin case.
- Dual is still a constrained concave quadratic (used in many solvers).
- Some presentations include bias in primal \( x_i^T w + b \);
  this introduces a constraint \( \sum_{i=1}^{n} y_i \alpha_i = 0 \) in dual.
Nonlinear SVM: feature mapping annoying in the primal?

SVM hard-margin primal, with a feature mapping $\phi$:

$$\min_{w \in \mathbb{R}^p} \frac{1}{2} \|w\|_2^2, \quad \forall i \phi(x_i)^T w \geq 1.$$  

Now the search space has $p$ dimensions; potentially $p \gg d$.

Can we do better?
Nonlinear SVM: feature mapping annoying in the primal?

SVM hard-margin primal, with a feature mapping \( \phi : \mathbb{R}^d \to \mathbb{R}^p \):

\[
\min \left\{ \frac{1}{2} \|w\|_2^2 : \ w \in \mathbb{R}^p, \ \forall \ i \ . \ \phi(x_i)^T w \geq 1 \right\}.
\]

Now the search space has \( p \) dimensions; potentially \( p \gg d \).

Can we do better?
Feature mapping in the dual

\[ \max_{\alpha_1, \alpha_2, \ldots, \alpha_n \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j). \]

Given dual optimum \( \bar{\alpha} \), since \( \bar{\omega} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(x_i) \), we can predict on future \( x \) with

\[ \bar{\omega}^T \phi(x). \]

▶ Dual form never needs \( \phi(x) \in \mathbb{R}^p \), only \( \phi(x)^T \phi(x_i) \in \mathbb{R} \).

▶ Kernel trick: replace every \( \phi(x)^T \phi(x') \) with kernel evaluation \( k(x, x') \).

Sometimes \( k(\cdot, \cdot) \) is much cheaper than \( \phi(x)^T \phi(x') \).

▶ This idea started with SVM, but appears in many other places.

▶ Downside: implementations usually store Gram matrix \( G \in \mathbb{R}^{n \times n} \) where \( G_{ij} := k(x_i, x_j) \).
Feature mapping in the dual

SVM hard-margin dual, with a feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$:

$$\max_{\alpha_1, \alpha_2, \ldots, \alpha_n \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j).$$

Given dual optimum $\bar{\alpha}$, since $\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(x_i)$, we can predict on future $x$ with

$$x \mapsto \phi(x)^T \bar{w} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \phi(x)^T \phi(x_i).$$
Feature mapping in the dual

SVM hard-margin dual, with a feature mapping \( \phi : \mathbb{R}^d \to \mathbb{R}^p \):

\[
\max_{\alpha_1, \alpha_2, \ldots, \alpha_n \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j).
\]

Given dual optimum \( \bar{\alpha} \), since \( \bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(x_i) \), we can predict on future \( x \) with

\[
x \mapsto \phi(x)^T \bar{w} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \phi(x)^T \phi(x_i).
\]

- **Dual form never needs** \( \phi(x) \in \mathbb{R}^p \), only \( \phi(x)^T \phi(x_i) \in \mathbb{R} \).
- **Kernel trick:** replace every \( \phi(x)^T \phi(x') \) with kernel evaluation \( k(x, x') \).
  
  Sometimes \( k(\cdot, \cdot) \) is much cheaper than \( \phi(x)^T \phi(x') \).
- This idea started with SVM, but appears in many other places.
- **Downside:** implementations usually store Gram matrix \( G \in \mathbb{R}^{n \times n} \) where \( G_{ij} := k(x_i, x_j) \).
Affine features: $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{1+d}$, where

$$\phi(x) = (1, x_1, \ldots, x_d).$$

Kernel form:

$$\phi(x)^T \phi(x') = 1 + x^T x'.$$
Kernel example: quadratic features

Consider re-normalized quadratic features
\[ \phi: \mathbb{R}^d \rightarrow \mathbb{R}^{1+2d+\binom{d}{2}}, \text{ where} \]
\[ \phi(x) = \left( 1, \sqrt{2}x_1, \ldots, \sqrt{2}x_d, x_1^2, \ldots, x_d^2, \right. \]
\[ \left. \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_1x_d, \ldots, \sqrt{2}x_{d-1}x_d \right). \]

Just writing this down takes time \( O(d^2) \).
Meanwhile,
\[ \phi(x)^\top \phi(x') = (1 + x^\top x')^2, \]

time \( O(d) \).
Consider re-normalized quadratic features
\[ \phi: \mathbb{R}^d \to \mathbb{R}^{1+2d+\binom{d}{2}}, \]
where
\[ \phi(x) = \left( 1, \sqrt{2}x_1, \ldots, \sqrt{2}x_d, x_1^2, \ldots, x_d^2, \right. \]
\[ \left. \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_1x_d, \ldots, \sqrt{2}x_{d-1}x_d \right). \]

Just writing this down takes time \( O(d^2) \).
Meanwhile,
\[ \phi(x)^\top \phi(x') = (1 + x^\top x')^2, \]
time \( O(d) \).

**Tweaks:**
- What if we change exponent “2”?
- What if we replace additive “1” with 0?
For any $\sigma > 0$, there is an infinite feature expansion $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^\infty$ such that

$$
\phi(x)^T \phi(x') = \exp \left( -\frac{\|x - x'\|_2^2}{2\sigma^2} \right),
$$

which can be computed in $O(d)$ time.

This is called a Gaussian kernel or RBF kernel. It has some similarities to nearest neighbor methods (later lecture).

$\phi$ maps to an infinite-dimensional space, but there's no reason to know that.
A (positive definite) kernel function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric function so that for any $n$ and any data examples $(x_i)_{i=1}^n$, the corresponding Gram matrix $G \in \mathbb{R}^{n \times n}$ with $G_{ij} := k(x_i, x_j)$ is positive semi-definite.
A (positive definite) **kernel function** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric function so that for any $n$ and any data examples $(x_i)_{i=1}^n$, the corresponding Gram matrix $G \in \mathbb{R}^{n \times n}$ with $G_{ij} := k(x_i, x_j)$ is positive semi-definite.

- There is a ton of theory about this formalism; e.g., keywords RKHS, representer theorem, Mercer’s theorem.
- Given any such $k$, there always exists a corresponding $\phi$.
- This definition ensures the SVM dual is still concave.
Source data.

Quadratic SVM.

RBF SVM ($\sigma = 1$).

RBF SVM ($\sigma = 0.1$).
Summary for SVM

- Hard-margin SVM.
- Soft-margin SVM.
- SVM duality.
- Nonlinear SVM: kernels
(Appendix.)
Equivalence of two hard-margin formulations.

**Proof.** First note that both have unique solutions (when feasible). For the first formulation, suppose we have two solutions \( u \) and \( u' \), and define another vector \( u'' := \frac{u + u'}{2} \). Then \( u'' \) achieves the same margin value as \( u \) and \( u' \), but if \( u \neq u' \), then \( \|u''\| < 1 \), which means \( u''/\|u''\| \) achieves a larger margin value than the purported optima \( u \) and \( u' \), a contradiction. For the second formulation, it suffices to note that the objective is strictly convex.

Now consider solution \( u \) to the first, with margin \( \gamma \). Then \( v := u/\gamma \) is feasible for second, with optimal value \( 1/(2\gamma^2) \). So the optimal value is at most this; if it is exactly the optimal value, we are done, otherwise suppose the optimum \( \bar{v} \) has \( \|\bar{v}\|^2/2 = 1/(2\rho^2) < 1/(2\gamma^2) \). Then \( \bar{u} := \rho \bar{v} \) is a unit vector, and moreover has \( \min_i y_i x_i^T \bar{u} = \rho > \gamma \), a contradiction since this is better than the supposed optimum \( u \).
Soft-margin dual derivation

Let’s derive the final dual form:

\[ L(w, \xi, \alpha) = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i x_i^T w) \]

\[ D(\alpha) = \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n \geq 0} L(w, \xi, \alpha) = \max_{\alpha \in \mathbb{R}^n} \left[ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 \right]. \]

Given \( \alpha \) and \( \xi \), the minimizing \( w \) is still \( w = \sum_{i=1}^{n} \alpha_i y_i x_i \); plugging in,

\[ D(\alpha) = \min_{\xi \in \mathbb{R}^n \geq 0} L \left( \sum_{i=1}^{n} \alpha_i y_i x_i, \xi, \alpha \right) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 + \sum_{i=1}^{n} \xi_i (C - \alpha_i). \]

The goal is to maximize \( D \); if \( \alpha_i > C \), then \( \xi_i \uparrow \infty \) gives \( D(\alpha) = -\infty \). Otherwise, minimized at \( \xi_i = 0 \). Therefore the dual problem is

\[ \max_{\alpha \in \mathbb{R}^n} \left[ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 \right]. \]
Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi : \mathbb{R} \to \mathbb{R}^\infty$.
What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?
Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi : \mathbb{R} \to \mathbb{R}^\infty$.

What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2}$$
Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi : \mathbb{R} \to \mathbb{R}^\infty$.

What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2}$$

$$= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left( \frac{xy}{\sigma^2} \right)^k$$
Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi : \mathbb{R} \rightarrow \mathbb{R}^{\infty}$.

What $\phi$ has $\phi(x) \phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2}$$

$$= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{xy}{\sigma^2}\right)^k$$

So let

$$\phi(x) := e^{-x^2/(2\sigma^2)} \left(1, \frac{x}{\sigma}, \frac{1}{\sqrt{2!}} \left(\frac{x}{\sigma}\right)^2, \frac{1}{\sqrt{3!}} \left(\frac{x}{\sigma}\right)^3, \ldots \right).$$
Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi : \mathbb{R} \to \mathbb{R}^\infty$.

What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

\[
e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{xy}{\sigma^2}\right)^k
\]

So let

\[
\phi(x) := e^{-x^2/(2\sigma^2)} \left(1, \frac{x}{\sigma}, \frac{1}{\sqrt{2!}} \left(\frac{x}{\sigma}\right)^2, \frac{1}{\sqrt{3!}} \left(\frac{x}{\sigma}\right)^3, \ldots \right).
\]

How to handle $d > 1$?
First consider $d = 1$, meaning $\phi: \mathbb{R} \to \mathbb{R}^\infty$.

What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left( \frac{xy}{\sigma^2} \right)^k$$

So let

$$\phi(x) := e^{-x^2/(2\sigma^2)} \left( 1, \frac{x}{\sigma}, \frac{x^2}{2!}, \frac{x^3}{3!}, \ldots \right).$$

How to handle $d > 1$?

$$e^{-\|x-y\|^2/(2\sigma^2)} = e^{-\|x\|^2/(2\sigma^2)} \cdot e^{-\|y\|^2/(2\sigma^2)} \cdot e^{x^T y/\sigma^2}$$

$$= e^{-\|x\|^2/(2\sigma^2)} \cdot e^{-\|y\|^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left( \frac{x^T y}{\sigma^2} \right)^k.$$
Consider $\phi : \mathbb{R}^d \to \mathbb{R}^{2^d}$, where

$$\phi(x) = \left( \prod_{i \in S} x_i \right)_{S \subseteq \{1, 2, \ldots, d\}}$$

Time $O(2^d)$ just to write down.

Kernel evaluation takes time $O(d)$:

$$\phi(x)^T \phi(x') = \prod_{i=1}^{d} (1 + x_i x'_i).$$
Other kernels

Suppose $k_1$ and $k_2$ are positive definite kernel functions.
Other kernels

Suppose \( k_1 \) and \( k_2 \) are positive definite kernel functions.

1. \( k(x, y) := k_1(x, y) + k_2(x, y) \) define a positive definite kernel?
Other kernels

Suppose $k_1$ and $k_2$ are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?
2. $k(x, y) := c \cdot k_1(x, y) \ (\text{for } c \geq 0)$ define a positive definite kernel?
Suppose $k_1$ and $k_2$ are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?
2. $k(x, y) := c \cdot k_1(x, y)$ (for $c \geq 0$) define a positive definite kernel?
3. $k(x, y) := k_1(x, y) \cdot k_2(x, y)$ define a positive definite kernel?
Suppose $k_1$ and $k_2$ are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?
2. $k(x, y) := c \cdot k_1(x, y)$ (for $c \geq 0$) define a positive definite kernel?
3. $k(x, y) := k_1(x, y) \cdot k_2(x, y)$ define a positive definite kernel?

Another approach: random features $k(x, x') = \mathbb{E}_w F(w, x')^T F(w, x')$ for some $F$; we will revisit this with deep networks and the neural tangent kernel (NTK).
Kernel ridge regression:

\[
\min_w \frac{1}{2n} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2.
\]

Solution:

\[
\hat{w} = (X^TX + \lambda n I)^{-1} X^T y.
\]

Linear algebra fact:

\[
X^T(XX^T + \lambda n I)^{-1} y.
\]

Therefore predict with

\[
x \mapsto x^T \hat{w} = (X x)^T (X X^T + \lambda n I)^{-1} y = \sum_{i=1}^n (x_i^T x)^T \left[(X X^T + \lambda n I)^{-1} y\right]_i.
\]

Kernel approach:

- Compute \( \alpha := (G + \lambda n I)^{-1} \), where \( G \in \mathbb{R}^{n \times n} \) is the Gram matrix: \( G_{ij} = k(x_i, x_j) \).
- Predict with \( x \mapsto \sum_{i=1}^n \alpha_i x_i^T x \).
There are a few ways; one is to use one-against-all as in lecture.

Many researchers have proposed various multiclass SVM methods, but some of them can be shown to fail in trivial cases.
Supplemental reading

- Shalev-Shwartz/Ben-David: chapter 15.
- Murphy: chapter 14.