Support Vector Machines (SVM): Intuition, Math, Primal–Dual, Lagrange Multipliers, Kernels, KKT, and Simulation

Pournami P N

October 30, 2025

Learning Goals

- SVM - Early Foundations
- Build intuition for SVMs: hyperplanes, margins, support vectors
- Understand **primal** and **dual** formulations
- Learn **Lagrange multipliers** and their role
- Derive and interpret the **KKT** conditions
- See the **kernel trick** and decision function
- Practice-ready guidance + **simulation code**

What is an SVM?

- Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression.
- It finds the **optimal separating hyperplane** that maximizes the **margin** between classes.
- Originally designed for **linear separation**, later extended using the **kernel trick** for nonlinear data.

Core idea :

$$\max_{\text{margin}} \Rightarrow \min \tfrac{1}{2}\|w\|^2$$

1960s–1970s: Theoretical Foundations

- **Vladimir Vapnik** and **Alexey Chervonenkis** (Moscow, USSR) developed the foundations of **statistical learning theory**.
- Introduced:
  - **VC dimension** — a measure of model complexity.
  - **Optimal hyperplane** for linearly separable data.
- Laid groundwork for modern SVM theory.

Vapnik & Chervonenkis (1963): “A note on one class of perceptrons.”

1980s: Computational Era Begins

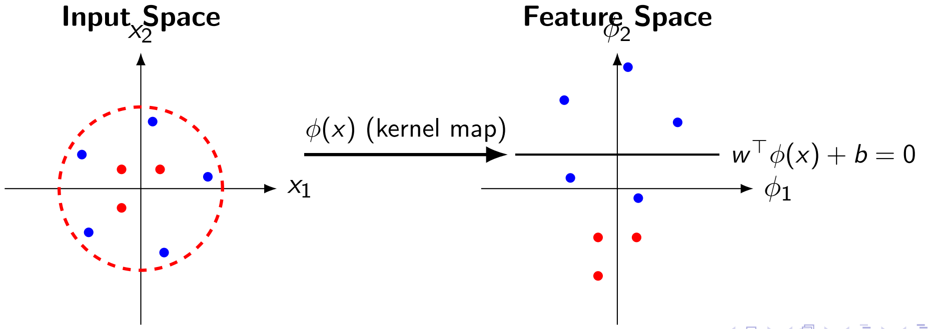
- Advances in optimization made it feasible to compute separating hyperplanes.
- Early algorithms could now handle small datasets using quadratic programming.
- Focus: improving **generalization** and introducing **soft margins**.

1992: The Kernel Trick is Born

- **Boser, Guyon, and Vapnik (1992)**: “A Training Algorithm for Optimal Margin Classifiers.”
- Introduced the **kernel trick**:

$$K(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$$

- Enabled nonlinear classification by mapping data into higher dimensions implicitly.



1995: Soft-Margin SVM

- **Cortes & Vapnik (1995):** "Support-Vector Networks."
- Introduced:
  - **Soft margins** – allow misclassifications via slack variables ξ_i.
  - Regularization parameter C to balance margin width and error.
- Made SVMs practical for noisy, real-world data.

Optimization problem :

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \quad \text{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i$$

Late 1990s: SVMs in Action

- Widely adopted for:
  - Handwritten digit recognition (MNIST)
  - Text and document classification
  - Bioinformatics (gene expression data)
  - Image recognition
- Popular kernels:
  - Polynomial kernel
  - RBF (Gaussian) kernel
  - Sigmoid kernel

2000s: Competing with Neural Networks

- Early 2000s: SVMs became the gold standard for small-to-medium datasets.
- Outperformed neural networks in many tasks due to:
  - Convex optimization (no local minima)
  - Strong generalization guarantees
- Still used heavily in text mining, genomics, and medical data.

Major Contributors to SVM Development

Year	Contributors	Contribution
1963	Vapnik & Chervonenkis	Optimal hyperplane, VC theory
1979	Vapnik	Statistical learning theory
1992	Boser, Guyon, Vapnik	Kernel trick introduced
1995	Cortes & Vapnik	Soft-margin SVMs
1998–2000	Scholkopf & Smola	Kernel methods, SVM theory

SVM in the Deep Learning Era

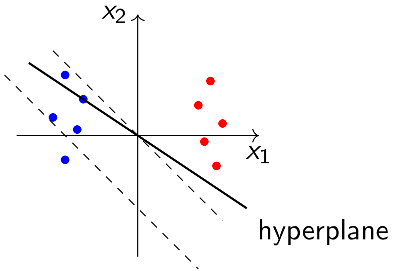
- Deep neural networks now dominate large-scale image and speech tasks.
- SVMs remain powerful for:
  - Small datasets
  - High-dimensional data (e.g., text, bioinformatics)
  - Outlier and novelty detection (one-class SVM)
- The SVM's theoretical legacy continues in modern margin-based learning.

What is an SVM?

Find a separating **hyperplane** that maximizes the **margin** between classes.

Decision function (binary classification):

$$f(x) = \operatorname{sign}\left(w^\top x + b\right)$$



- w controls the orientation of the hyperplane.
- b shifts it.

Margin and Support Vectors

- Margin = distance from hyperplane to nearest points of either class.
- For a hyperplane $w^\top x + b = 0$, the (signed) distance of x is $\dfrac{w^\top x + b}{\|w\|}$.
- The closest points that *touch* the margin are the **support vectors**.

Maximize margin $\iff$ minimize $\frac{1}{2}\|w\|^2$ under appropriate constraints.

Hard-Margin SVM (Separable Data)

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i\left(w^\top x_i + b\right) \geq 1, \quad \forall i. \end{aligned}$$

- All points must be correctly classified and lie **outside** the margin.

Soft-Margin SVM (Real Data)

Introduce slacks $\xi_i \geq 0$:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0. \end{aligned}$$

- C controls trade-off between large margin and training violations.
- Hinge-loss view: $\max\{0, 1 - y_i(w^\top x_i + b)\}$.

Lagrange Multipliers (Idea)

- For constrained optimization, build the **Lagrangian** to merge objective and constraints.
- Equality case: $L(x, \lambda) = f(x) - \lambda g(x)$; optimality requires stationarity.
- Inequalities use KKT conditions (next).

SVM Lagrangian (Soft-Margin)

Let $\alpha_i \geq 0$ for margin constraints and $\mu_i \geq 0$ for $\xi_i \geq 0$:

$$\begin{aligned} L(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C \sum_i \xi_i \\ - \sum_i \alpha_i \left[y_i(w^\top x_i + b) - 1 + \xi_i \right] - \sum_i \mu_i \xi_i . \end{aligned}$$

Stationarity gives

$$w = \sum_i \alpha_i y_i x_i, \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i + \mu_i = C.$$

Dual Problem (Soft-Margin)

Eliminating w, b, ξ yields the dual:

$$\begin{aligned} \max_{\alpha} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \, (x_i^\top x_j) \\ \text{s.t.} \quad & 0 \le \alpha_i \le C, \qquad \sum_i \alpha_i y_i = 0. \end{aligned}$$

- Only samples with $\alpha_i > 0$ are **support vectors**.
- Recover $w^* = \sum_i \alpha_i^* y_i x_i$; find b^* from any margin SV.

KKT Conditions at Optimum

- **Primal feasibility:** $y_i(w^\top x_i + b) - 1 + \xi_i \geq 0,\ \xi_i \geq 0.$
- **Dual feasibility:** $\alpha_i \geq 0,\ \mu_i \geq 0.$
- **Stationarity:** $w = \sum_i \alpha_i y_i x_i.$
- **Complementary slackness:** $\alpha_i \left[y_i(w^\top x_i + b) - 1 + \xi_i \right] = 0,\ \mu_i \xi_i = 0.$

- $0 < \alpha_i < C$: points on the margin ⇒ support vectors.
- $\alpha_i = 0$: points strictly outside the margin (no influence).
- $\alpha_i = C$: violations (inside margin / misclassified).

Kernel Trick

Replace dot products with kernels $K(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$:

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

$$\text{s.t.} \quad 0 \le \alpha_i \le C, \ \sum_i \alpha_i y_i = 0.$$

Common kernels:

- Linear: $K(x, z) = x^\top z$
- Polynomial: $K(x, z) = (x^\top z + c)^d$
- RBF (Gaussian): $K(x, z) = \exp\big(-\gamma \|x - z\|^2\big)$
- Sigmoid: $K(x, z) = \tanh\big(\alpha x^\top z + c\big)$

Decision Function

$$f(x) = \operatorname{sign}\Big(\sum_{i \in \mathrm{SV}} \alpha_i y_i K(x_i, x) + b \Big).$$

- **Sparsity:** the sum involves only SVs.
- **Confidence:** distance to hyperplane $\propto \frac{w^\top x + b}{\|w\|}$ (linear case).

Training Recipe (Classification)

1. **Scale** features (zero-mean, unit variance).
2. Choose kernel: start with **linear**; try **RBF** for nonlinearity.
3. Tune hyperparameters via cross-validation:
   - Linear: C
   - RBF: C and γ
4. Handle class imbalance with weights or resampling.
5. Watch % of support vectors as a sanity check for overfitting.

Multiclass and Complexity

- **Multiclass**: One-vs-Rest (OvR) or One-vs-One (OvO).
- **Complexity**: Kernel SVM training scales roughly between $O(N^2)$ and $O(N^3)$; linear SVM scales much better for large N and sparse features.

Tiny 1D Example (Hard-Margin)

Data: $-2, -1$ labeled -1 and $1, 2$ labeled $+1$.
A separating hyperplane at $x = 0$ (i.e., $w = 1, b = 0$) satisfies $y_i(wx_i + b) \geq 1$.
Margin width $= \dfrac{2}{\|w\|} = 2$, nearest points at ± 1 are the SVs.

Key Equations (Summary)

Concept	Equation
Boundary	$w^\top x + b = 0$
Margin width	$2/\|w\|$
Hard-margin primal	$\min \frac{1}{2}\|w\|^2$ s.t. $y_i(w^\top x_i + b) \geq 1$
Soft-margin primal	$\min \frac{1}{2}\|w\|^2 + C\sum \xi_i$
Dual objective	$\max_\alpha \sum \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i\alpha_j y_i y_j K_{ij}$
Decision function	$\mathrm{sign}\left(\sum_{i\in SV} \alpha_i y_i K(x_i, x) + b\right)$

The SVM Objective

Goal: Find a hyperplane that separates data with the largest possible margin.

Hyperplane equation

$$w^T x + b = 0$$

Margin: the distance between the support vectors (the nearest points) and the separating hyperplane.

$$\text{Margin width} = \frac{2}{\|w\|}$$

We want to maximize this margin.

$$\max \frac{2}{\|w\|} \quad \Longleftrightarrow \quad \min \|w\|$$

To make differentiation easier

$$\boxed{\min \frac{1}{2}\|w\|^2}$$

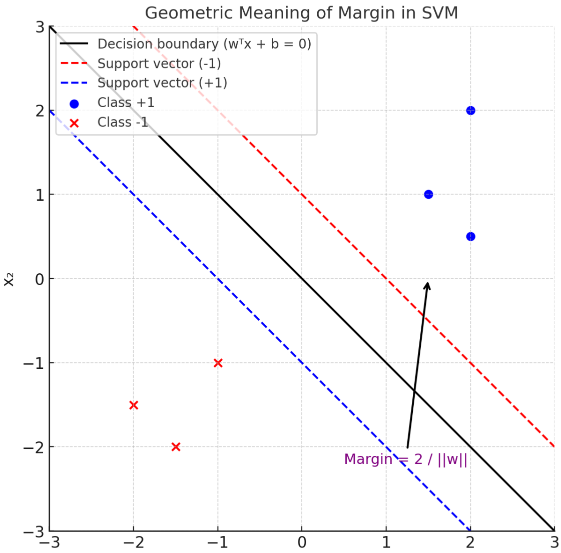
Why Minimizing $\frac{1}{2}\|w\|^2$ Maximizes the Margin ?

- The SVM tries to make the separating hyperplane as “flat” as possible.
- The vector w is perpendicular to the hyperplane.
- Smaller $\|w\|$ means a gentler slope — hence a **wider margin**.
- The margin width is inversely proportional to $\|w\|$:

$$\text{Margin} = \frac{2}{\|w\|}$$

- So minimizing $\|w\|$ (or equivalently $\frac{1}{2}\|w\|^2$) directly maximizes the margin.

Geometric Meaning of Margin in SVM



Hard-Margin Primal Problem (Separable Data)

Given (x_i, y_i) with $y_i \in \{\pm 1\}$,

$$\begin{aligned} \min_{w,b} \quad & \tfrac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i\,(w^\top x_i + b) \;\geq\; 1, \qquad i = 1, \ldots, n. \end{aligned}$$

Intuition: enforce perfect separation while making $\|w\|$ small $\Rightarrow$ large margin.

Soft-Margin Primal Problem (Non-separable Data)

Introduce slacks $\xi_i \geq 0$ and penalty $C > 0$:

$$\begin{aligned} \min_{w,b,\xi} \quad & \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i\left(w^\top x_i + b\right) \;\geq\; 1 - \xi_i, \quad \xi_i \geq 0. \end{aligned}$$

Trade-off: large $C \Rightarrow$ fewer violations; small $C \Rightarrow$ wider margin but more slack.

Lagrangian and KKT Conditions (Soft-Margin)

Lagrangian with multipliers $\alpha_i \geq 0$ and $\mu_i \geq 0$:

$$\mathcal{L}(w, b, \xi, \alpha, \mu) = \tfrac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \Big[y_i(w^\top x_i + b) - 1 + \xi_i\Big] - \sum_i \mu_i \xi_i.$$

Stationarity:

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow \boxed{w = \sum_i \alpha_i y_i x_i}, \quad \frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_i \alpha_i y_i = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow \alpha_i + \mu_i = C \;\Rightarrow\; 0 \leq \alpha_i \leq C.$$

Complementary slackness:

$$\alpha_i \left[y_i(w^\top x_i + b) - 1 + \xi_i\right] = 0, \quad \mu_i \, \xi_i = 0.$$

Dual Quadratic Program (Soft-Margin)

Eliminate w, b, ξ using stationarity:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \tfrac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, x_i^\top x_j \\ \text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad \boxed{0 \le \alpha_i \le C}. \end{aligned}$$

Hard-margin is the special case $C \to \infty$ (effectively $\alpha_i \ge 0$ without upper bound).

Kernel Trick and Decision Function

Replace dot products with a kernel $K(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$:

$$\max_{\alpha} \quad \sum_i \alpha_i - \tfrac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \, K(x_i, x_j)$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C.$$

Classifier:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \, K(x_i, x) + b \right).$$

Recovering b: choose any support vector x_s with $0 < \alpha_s < C$,

$$b = y_s - \sum_{i=1}^{n} \alpha_i y_i \, K(x_i, x_s).$$

Strong Duality & Support Vectors

- The primal is convex with linear constraints $\Rightarrow$ **strong duality** holds.
- Only points with $\alpha_i > 0$ contribute to w (or the decision function): these are the **support vectors**.
- If $0 < \alpha_i < C$ then point i lies exactly on the margin ($y_i(w^\top x_i + b) = 1$).
- If $\alpha_i = C$ the point either violates the margin or is misclassified (active slack).

Primal vs Dual: A Quick Map

Aspect	**Primal**	**Dual**
Objective	min $\frac{1}{2}\|w\|^2 + C\sum \xi_i$	max $\sum \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j K_{ij}$
Vars	w, b, ξ	α
Constraints	$y_i(w^\top x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0$	$\sum_i \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C$
Kernelization	Implicit via w	Natural via $K(x_i, x_j)$
Sparsity	Not explicit	Explicit: only $\alpha_i > 0$ matter

Algorithmic Note (Practice)

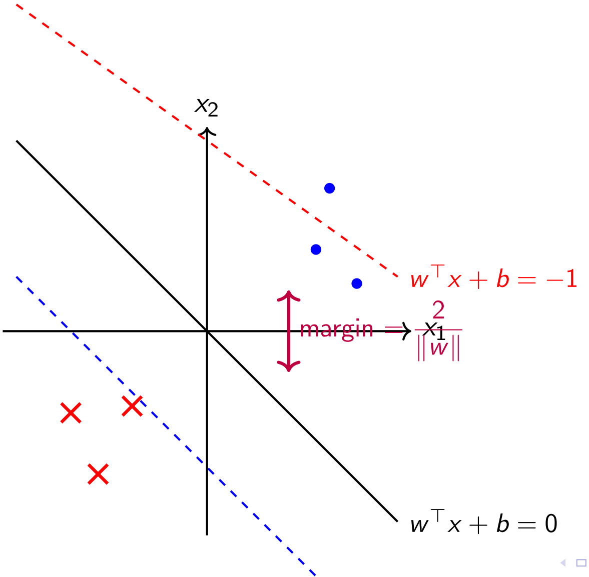
- Dual QP often solved by **SMO** (Sequential Minimal Optimization) or modern QP solvers.
- Common kernels: linear, polynomial, RBF, sigmoid.
- Hyperparameters: C (soft-margin), kernel params (e.g., γ in RBF).

Goal of an SVM (in plain words)

- We have points labeled $+1$ and -1.
- We want a line/plane (a **hyperplane**) that separates them.
- Not just any separator — we want the one with the **largest gap** (the **margin**).

$$\text{Hyperplane: } w^\top x + b = 0$$

Margin picture (2D)



Hard-margin Primal (perfectly separable data)

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i\,(w^\top x_i + b) \ge 1, \qquad i = 1, \dots, n. \end{aligned}$$

- Minimizing $\frac{1}{2}\|w\|^2 \iff$ **maximizing the margin** $\left(\frac{2}{\|w\|}\right)$.
- Constraints keep points on the correct side of the margin.

Soft-margin Primal (real data is messy)

Allow some violations using slacks $\xi_i \geq 0$ and a trade-off $C > 0$:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i\,(w^\top x_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0. \end{aligned}$$

- C large ⇒ fewer mistakes, possibly smaller margin.
- C small ⇒ wider margin, more tolerance to mistakes.

Why build a Dual?

- Handling constraints directly can be harder.
- The **Dual** becomes a nice Quadratic Program (QP).
- In the Dual, dot products $x_i^\top x_j$ allow the **kernel trick** for non-linear boundaries.

Lagrangian (soft-margin form)

Introduce multipliers $\alpha_i \geq 0$ (for margin constraints) and $\mu_i \geq 0$ (for slacks):

$$\mathcal{L}(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i\left[y_i(w^\top x_i + b) - 1 + \xi_i\right] - \sum_i \mu_i \xi_i.$$

Stationarity gives

$$w = \sum_i \alpha_i y_i x_i, \qquad \sum_i \alpha_i y_i = 0, \qquad 0 \leq \alpha_i \leq C.$$

Dual QP (soft-margin)

Eliminating w, b, ξ:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, x_i^\top x_j$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad 0 \le \alpha_i \le C.$$

Classifier after solving:

$$f(x) = \operatorname{sign}\left(\sum_{i} \alpha_i y_i \, x_i^\top x + b \right).$$

(Replace $x_i^\top x$ by kernel $K(x_i, x)$ for non-linear SVM.)

Two views of the same goal

Aspect	**Primal**	**Dual**
Objective	$\min \frac{1}{2}\|w\|^2 + C\sum \xi_i$	$\max \sum \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K_{ij}$
Variables	w, b, ξ	α
Constraints	$y_i(w^\top x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0$	$\sum_i \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C$
Kernelization	Implicit via w	Natural via $K(x_i, x_j)$
Sparsity	Not explicit	Only $\alpha_i > 0$ (support vectors) matter

Setup: two 2D points, one per class

Data (hard-margin, linearly separable):

$$x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; y_1 = +1, \qquad x_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \; y_2 = -1.$$

- Symmetric about the origin along x_1-axis.
- The separating hyperplane should be the vertical line through the origin.

Dual formulation for this toy data

With two points and hard margin, the Dual becomes:

$$\max_{\alpha_1,\alpha_2} \ \alpha_1 + \alpha_2 - \frac{1}{2}\Big(\alpha_1^2 y_1^2 \left\|x_1\right\|^2 + \alpha_2^2 y_2^2 \left\|x_2\right\|^2 + 2\alpha_1\alpha_2 y_1 y_2 \, x_1^\top x_2\Big)$$

Numbers:

$$\|x_1\|^2 = \|x_2\|^2 = 1, \quad x_1^\top x_2 = -1, \quad y_1 = +1, \ y_2 = -1.$$

Constraint $\sum_i \alpha_i y_i = 0 \Rightarrow \alpha_1 = \alpha_2 \, (= a)$.

$$\Rightarrow \ \max_a \ \ 2a - \frac{1}{2}\big(a^2 + a^2 - 2a^2(-1)\big) = 2a - \frac{1}{2}(4a^2) = 2a - 2a^2.$$

$$\frac{d}{da}(2a - 2a^2) = 2 - 4a = 0 \ \Rightarrow \ \boxed{a = \tfrac{1}{2}}.$$

So $\boxed{\alpha_1 = \alpha_2 = \tfrac{1}{2}}$.

Recovering w and b (KKT stationarity)

From $w = \sum_i \alpha_i y_i x_i$:

$$w = \frac{1}{2} \cdot (+1) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \frac{1}{2} \cdot (-1) \begin{bmatrix} -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Use a support vector with $0 < \alpha_i$ to get b:

$$y_1(w^\top x_1 + b) = 1 \Rightarrow 1 \cdot (1 \cdot 1 + b) = 1 \Rightarrow \boxed{b = 0}.$$

Decision function:

$$f(x) = \operatorname{sign}(w^\top x + b) = \operatorname{sign}(x_1).$$

So the boundary is $x_1 = 0$ (vertical line through the origin).

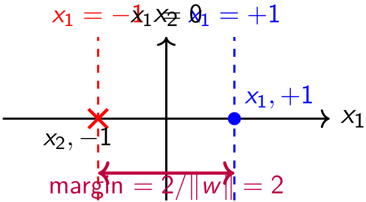
Margin check (does it match theory?)

$$\|w\| = \left\| \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\| = 1 \quad \Rightarrow \quad \text{margin} = \frac{2}{\|w\|} = 2.$$

Support hyperplanes:

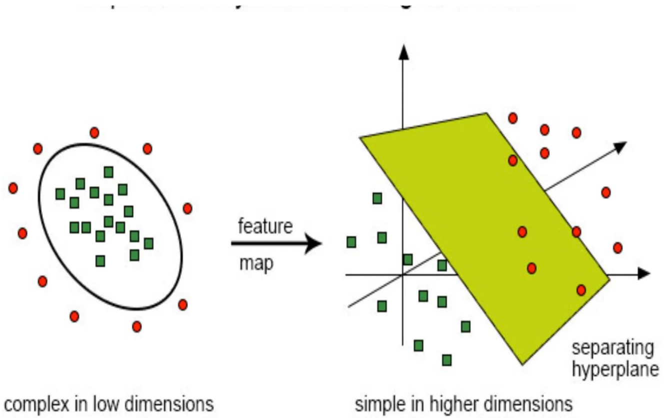
$$w^\top x + b = \pm 1 \ \Rightarrow \ x_1 = \pm 1,$$

whose distance along the x_1-axis is 2, matching $\frac{2}{\|w\|}$.



Kernel Trick: From Input Space to Feature Space

- The key idea behind kernel methods is to transform data into a higher-dimensional space using a kernel function, making it easier to perform linear separation in this new space.
- This approach is particularly useful for dealing with non-linear data.



Key ideas to remember

- SVM $=$ max-margin $+$ convex optimization $+$ kernels.
- Lagrange multipliers α_i link constraints to solution; nonzero α_i mark SVs.
- KKT ties primal and dual; kernel trick enables nonlinear boundaries.
- **Primal:** “Find the flattest separator that keeps points on the right side.”
- **Dual:** “Find weights α_i on data so that their influence defines the separator.”
- **KKT:** Glue between the two; gives $w = \sum_i \alpha_i y_i x_i$.
- **Support vectors:** Only points with $\alpha_i > 0$ matter.
- **Kernel trick:** Replace dot products with $K(x_i, x_j)$ for curved boundaries.

Questions?