



An intentionally blank page

1) The data and the goal

You have a labelled training set of n examples:

$$(x_i, y_i), \quad i = 1, \dots, n$$

where:

- $x_i \in \mathbb{R}^p$ is the feature vector for example i . Think of x_i as a list of p numbers (for example, height and weight for a person would be $p = 2$).
- $y_i \in \{+1, -1\}$ is the class label (two classes: +1 or -1).

Goal: find a straight decision boundary (a hyperplane) that separates the +1 points from the -1 points **with the largest possible gap** between the two classes. This gap is called the **margin**.

2) What is a hyperplane? — $w^T x + b = 0$

A hyperplane in p -dimensional space is the set of points x satisfying

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

where:

- $w \in \mathbb{R}^p$ is the **normal vector** (it points perpendicular to the hyperplane).
- $b \in \mathbb{R}$ is the **bias** or offset (it moves the hyperplane away from the origin).

Interpretation:

- $w^T x$ is the dot product (sum of elementwise products). It measures how much x aligns with w .
- The hyperplane splits space: points with $w^T x + b > 0$ lie on one side, and those with $w^T x + b < 0$ lie on the other.

3) Distance from a point to the hyperplane

The perpendicular (shortest) distance from a point x_0 to the hyperplane $w^T x + b = 0$ is

$$\text{dist}(x_0, H) = \frac{|w^T x_0 + b|}{\|w\|},$$

where $\|w\| = \sqrt{w_1^2 + \dots + w_p^2}$ is the Euclidean length (norm) of w .

Why that formula? Intuition: divide the signed value $w^T x_0 + b$ by the length of w to convert the algebraic expression into a geometric length (you can think of projecting x_0 onto the unit normal $w / \|w\|$).

4) The margin and the canonical scaling (important correction)

We place two parallel supporting hyperplanes so that they touch the closest points of each class:

$$w^T x + b = +1 \quad \text{and} \quad w^T x + b = -1.$$

Those two lines (or hyperplanes) are parallel to the decision boundary $w^T x + b = 0$. The **margin** γ is the distance between these two supporting hyperplanes.

Correct formula (fixing the typo in your text):

$$\gamma = \frac{2}{\|w\|}$$

So the margin equals 2 divided by the length of w . (Your note had $\gamma = 2 \|w\|$ which is backwards — the margin shrinks when $\|w\|$ grows.)

How we get that: distance between hyperplanes $w^T x + b = c_1$ and $w^T x + b = c_2$ is $\frac{|c_1 - c_2|}{\|w\|}$. Here $c_1 = 1$ and $c_2 = -1$ so distance = $\frac{2}{\|w\|}$.

Also, the distance from the center hyperplane $w^T x + b = 0$ to either supporting hyperplane is $1/\|w\|$.

Why set the support hyperplanes at ± 1 ? Because we can scale w and b by any positive constant without changing the decision boundary: $w^T x + b = 0$ is the same hyperplane as $(\alpha w)^T x + (\alpha b) = 0$ for $\alpha \neq 0$. So we pick the scaling so the closest points satisfy

$$y_i (w^T x_i + b) = 1 \quad \text{for support vectors.}$$

This is called **canonical scaling** and it makes the math neat.

5) Maximizing margin \rightarrow optimization problem

Because $\gamma = 2/\|w\|$, maximizing γ is the same as minimizing $\|w\|$. For convenience (and because it is mathematically smooth), we minimize $\frac{1}{2} \|w\|^2$. The constrained optimization (for perfectly separable data) becomes:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ \text{subject to} \quad & y_i (w^T x_i + b) \geq 1, \quad i = 1, \dots, n. \end{aligned}$$

Explanation of constraint $y_i(w^\top x_i + b) \geq 1$:

- If $y_i = +1$ this says $w^\top x_i + b \geq 1$ (point is on or beyond the +1 supporting hyperplane).
- If $y_i = -1$ this says $w^\top x_i + b \leq -1$.

So every point is on the correct side and at least at distance $1 / \|w\|$ from the decision boundary.

Why $\frac{1}{2} \|w\|^2$? The half just makes derivatives cleaner (the factor 1/2 cancels when differentiating); the squared norm is convex and easier to optimize.

6) Support vectors — what actually defines the classifier

Only the points that lie exactly on the supporting hyperplanes (those with $y_i(w^\top x_i + b) = 1$) are called **support vectors**. They "support" the optimal margin — if you move a support vector slightly the optimal hyperplane often changes. Points farther away (with strict > 1) do not affect the optimal w, b directly.

In the solution (from the Lagrangian method) you get:

$$w = \sum_{i=1}^n \alpha_i y_i x_i,$$

where the multipliers $\alpha_i \geq 0$ and $\alpha_i > 0$ only for support vectors. So the final w is a weighted sum of the support vectors.

7) Soft margin (short note) — when data is not perfectly separable

If the classes overlap (no perfect separating hyperplane), introduce slack variables $\xi_i \geq 0$ and a penalty $C > 0$ that trades off margin size versus misclassification:

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

Large $C \rightarrow$ try hard to avoid errors (may get small margin). Small $C \rightarrow$ allow more errors to get bigger margin.

8) Small 2D numeric example (very simple)

Take $p = 2$. Let $w = (0, 1)$ and $b = 0$. Then hyperplane is $0 \cdot x_1 + 1 \cdot x_2 + 0 = 0 \Rightarrow x_2 = 0$ (the x-axis).

- Norm $\|w\| = \sqrt{0^2 + 1^2} = \sqrt{1} = 1$.
- Supporting hyperplanes: $w^\top x + b = 1 \Rightarrow x_2 = 1$ and $= -1 \Rightarrow x_2 = -1$.
- Margin $\gamma = 2 / \|w\| = 2/1 = 2$. So the distance between $x_2 = 1$ and $x_2 = -1$ is 2, and the nearest distance from the decision boundary to either margin line is $1 / \|w\| = 1$.

If a point $(0, 0.8)$ has label $+1$, check the constraint: $y(w^\top x + b) = (+1) \cdot (0 \cdot 0 + 1 \cdot 0.8 + 0) = 0.8$. That is less than 1, so $(0, 0.8)$ lies inside the margin and would either be a slack point (in soft margin) or disallowed (in hard-margin separable case).

9) Plain-English summary


- x_i : a data point (list of features). y_i : class (+1 or -1).
- w : direction perpendicular to the decision boundary; b shifts it.
- We scale w, b so the closest points satisfy $y_i(w^\top x_i + b) = 1$. Then the margin (gap between the classes) equals $2 / \|w\|$.
- To get the widest margin we minimize $\frac{1}{2} \|w\|^2$ subject to those constraints. The points that “touch” the margin are support vectors — they determine the classifier.

An intentionally blank page



An intentionally blank page

Hard-Margin SVM Explanation



An intentionally blank page

1. The Problem SVM Tries to Solve

Imagine you have two types of points on a paper:

- Red points (Class -1)
- Blue points (Class +1)

They are **linearly separable**, meaning you can draw a **straight line** that separates all red points from all blue points **without mistakes**.

But... there are **infinite possible lines** you could draw. Which one should we pick?

SVM chooses the line that has the largest possible margin between the two classes.

The *margin* is the distance from the line to the nearest data point from either class.

2. What is "Hard-Margin" SVM?

- "Hard margin" means *no mistakes allowed* — every point must be correctly classified and lie outside (or exactly on) the margin boundary.
 - Works **only** when data is **perfectly separable**.
 - If points overlap, hard-margin SVM won't work — we need "soft-margin" instead.
-

3. The Geometry

We want:

- A separating line (in 2D) or hyperplane (in higher dimensions).
- Two parallel lines (margins) that are as far apart as possible, touching the nearest points of each class.

Mathematically:

$$\text{Decision boundary: } w^T x + b = 0$$

$$\text{Margin boundaries: } w^T x + b = +1 \quad \text{and} \quad w^T x + b = -1$$

Here:

- w = **vector** that defines the orientation of the line/hyperplane.
 - b = bias (shifts the hyperplane up/down).
 - x = input point.
-

4. Why Maximize the Margin?

A bigger margin means:

- The classifier is more confident.
- The model is less likely to overfit.
- Even small noise won't make the classifier flip its decision.

So we want the **widest possible gap** between the two margin lines.

5. The Math Formulation (Primal Problem)

The optimization problem is:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

subject to:

$$y_i(w^T x_i + b) \geq 1, \quad i = 1, 2, \dots, n$$

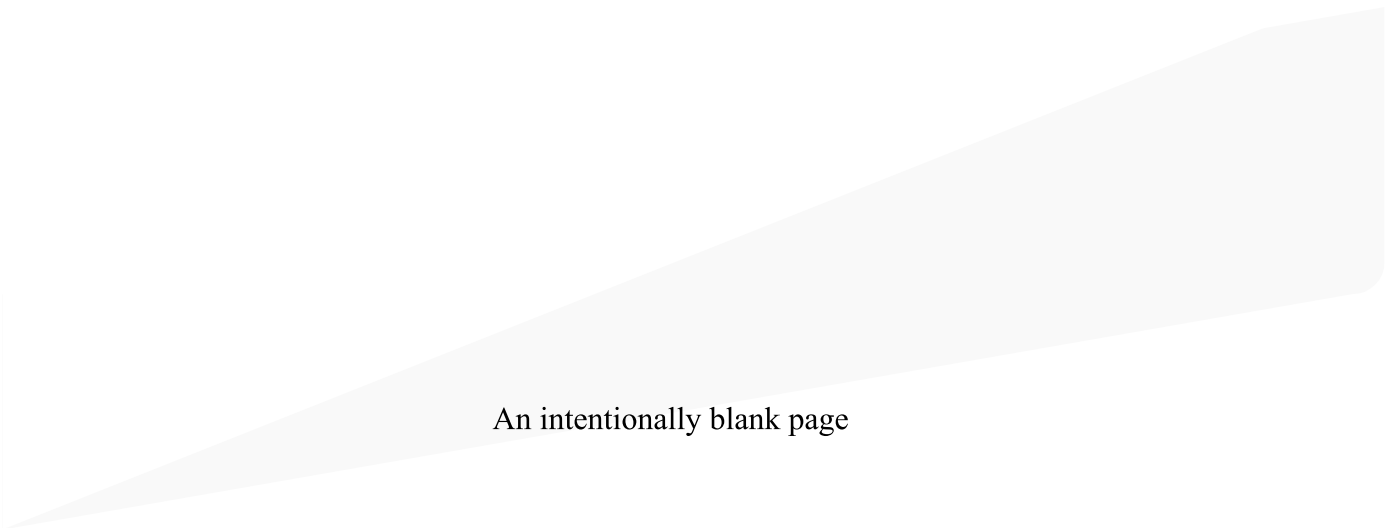
What does this mean?

1. $\frac{1}{2} \|w\|^2 \rightarrow$ This is what we minimize.
 - $\|w\|$ means the length of vector w .
 - Smaller $\|w\| =$ larger margin.

2. **Constraint:**

$$y_i(w^T x_i + b) \geq 1$$

- y_i = class label (+1 or -1).
 - If $y_i = +1$, the condition becomes $w^T x_i + b \geq 1 \rightarrow$ point is on or beyond the positive margin line.
 - If $y_i = -1$, the condition becomes $w^T x_i + b \leq -1 \rightarrow$ point is on or beyond the negative margin line.
-



An intentionally blank page

Lagrange multipliers in SVM

1) What problem are we solving?

Sometimes we must **optimize** (maximize or minimize) something subject to a rule (a constraint).

Example: “Find the highest point on a hill, but you must stay on a circular path.”

The hill height is the thing to optimize; the circle is the constraint.

You can’t just climb straight up the hill (take the ordinary derivative) because you must stay on the circle. Lagrange multipliers are a clever tool that let us find optima while respecting the constraint.

2) Intuition for a Lagrange multiplier

Think of a contour map of the hill (lines of equal height). If you’re restricted to the circle, the highest point on that circle happens where a contour line just touches the circle — they are tangent. At that touching point the direction that increases the hill the fastest (the **gradient** of the hill) is perpendicular to the circle, and the direction perpendicular to the circle is the gradient of the constraint. So these two gradients point the same way (are parallel).

The Lagrange multiplier λ is just a number that says “how much the constraint’s gradient must be scaled so it matches the hill’s gradient.” Algebraically we write:

$$\nabla f(x,y) = \lambda \nabla g(x,y)$$

where f is the thing we optimize and g is the constraint (e.g., $g(x,y)=0$ describes the circle). Solving this plus the constraint gives the answer.

3) Short algebraic recipe (no heavy calculus required)

To find extrema of $f(x,y)$ subject to $g(x,y)=0$:

1. Build the **Lagrangian**: $L(x,y,\lambda) = f(x,y) - \lambda \cdot (g(x,y))$
2. Solve the system of equations: $\partial L / \partial x = 0$, $\partial L / \partial y = 0$, and $g(x,y)=0$.

This gives candidate points; pick the one that fits.

4) Now: SVM (Support Vector Machine) — what is optimized?

SVM wants a straight line (in 2D) or hyperplane (higher D) that separates two classes and keeps the classes as far away from the line as possible. That “far away” is called the **margin**. For a *hard-margin* linear SVM the optimization is:

Minimize:

$$\frac{1}{2} \|w\|^2 \text{ (this is the same as maximizing the margin)}$$

Subject to (for every training point i):

$$y_i(w \cdot x_i + b) \geq 1$$

Here:

- w is the normal vector to the separating hyperplane,
- b shifts the plane,
- y_i is +1 or -1 (class label),
- x_i is the point.

These are many constraints (one per training point). You cannot just set derivative = 0 because of those constraints — you must account for them.

5) Why use Lagrange multipliers in SVM?

We turn the constrained minimization into an easier problem using Lagrange multipliers α_i (one non-negative multiplier for each constraint). Form the Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i (y_i(w \cdot x_i + b) - 1), \quad \alpha_i \geq 0.$$

Steps we do next (conceptual):

1. Take partial derivatives of L w.r.t w and b and set them to zero (this enforces optimality while accounting for constraints).
 - From $\partial L / \partial w = 0$ we get: $w = \sum_i \alpha_i y_i x_i$.

- From $\partial L / \partial b = 0$ we get: $\sum_i \alpha_i y_i = 0$.
2. Substitute w back into L to get the **dual problem**. The dual becomes a maximization problem in the variables α_i only.

6) Why this helps — the big benefits

1. **Only some points matter (support vectors):** By a condition called complementary slackness, for each training point either the constraint is not tight (point far away) and $\alpha_i = 0$, or the constraint is tight (point lies on the margin) and $\alpha_i > 0$. So only the points with $\alpha_i > 0$ — the **support vectors** — determine the final hyperplane. That's very efficient and intuitive: only the closest points to the boundary count.
2. **Kernel trick becomes possible:** The dual problem depends on dot products $x_i \cdot x_j$. By replacing dot products with kernel functions, SVM can learn non-linear boundaries without working in high-dimensional space explicitly. This huge flexibility comes from using the Lagrange dual form.
3. **Solving is easier / standard:** The dual is a convex quadratic programming problem in α_i with simple constraints ($\alpha_i \geq 0, \sum \alpha_i y_i = 0$). Convex problems have a single global optimum — good for reliability.

7) A friendly analogy

Think of each constraint (each training point) as a person who “complains” if the separating line treats them unfairly. The Lagrange multiplier α_i is how loudly that person complains. Most people are quiet ($\alpha=0$) because they are safely far away from the boundary. The loud ones ($\alpha>0$) are the ones right on the margin — they force the final decision boundary to be where it is. The SVM solution balances these complaints while also trying to make the boundary as “simple” as possible (small $|w|$).

8) Key takeaways (quick)

- Lagrange multipliers let you solve optimization problems with constraints by turning them into equations you can solve.
- In SVMs they let us include the constraints $y_i(w \cdot x_i + b) \geq 1$ directly and produce a dual that depends only on dot products.
- The multipliers α_i tell us which training points are important (support vectors) and enable kernels for non-linear SVMs.
- The math stays clean and gives a unique, global best solution because the problem is convex.

An intentionally blank page

Support Vector Machine Proofs

1. Intuition and goal

Given a labelled training set (x_i, y_i) for $i = 1, \dots, n$, where $x_i \in \mathbb{R}^p$ and labels $y_i \in \{+1, -1\}$, the linear SVM seeks a hyperplane that separates the two classes with the **maximum margin**. A hyperplane is

$$H : w^\top x + b = 0,$$

where $w \in \mathbb{R}^p$ (normal vector) and $b \in \mathbb{R}$ (bias).

We scale w, b so that the closest points satisfy $y_i(w^\top x_i + b) = 1$. With that canonical scaling, the margin γ (distance between the two parallel supporting hyperplanes) is

$$\gamma = \frac{2}{\|w\|},$$

and the distance from the hyperplane $w^\top x + b = 0$ to the closest point is $1/\|w\|$. So maximizing the margin is equivalent to minimizing $\|w\|$ (or $\frac{1}{2}\|w\|^2$) under certain constraints.

2. Hard-margin SVM (linearly separable data)

Primal optimization problem (convex quadratic program):

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

This enforces that all training points lie on or outside the two margin planes $w^\top x + b = \pm 1$.

2.1 Lagrangian (primal \rightarrow dual)

Introduce Lagrange multipliers $\alpha_i \geq 0$ for the inequality constraints. The (primal) Lagrangian is

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

Stationarity conditions (set partial derivatives to zero):

$$\begin{aligned} 1. \quad \frac{\partial L}{\partial w} = 0 &\Rightarrow w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^n \alpha_i y_i x_i. \\ 2. \quad \frac{\partial L}{\partial b} = 0 &\Rightarrow - \sum_{i=1}^n \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^n \alpha_i y_i = 0. \end{aligned}$$

Plugging w back into L yields the dual objective (maximize w.r.t. α subject to $\alpha_i \geq 0$ and $\sum \alpha_i y_i = 0$):

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

This is a concave quadratic maximization (equivalently convex QP in standard form).

2.2 Support vectors and classification rule

From solution α^* :

- Only those training points with $\alpha_i^* > 0$ appear in w . These are the **support vectors**.
- For any support vector x_k that lies exactly on the margin ($y_k(w^\top x_k + b) = 1$), we can compute b as

$$b = y_k - w^\top x_k.$$

In practice use the average over support vectors.

Decision function for a new x :

$$f(x) = \text{sign}(w^\top x + b) = \text{sign}\left(\sum_{i \in S} \alpha_i^* y_i x_i^\top x + b\right),$$

where S is the set of indices with $\alpha_i^* > 0$.

2.3 Relationship of $\|w\|$ and margin

With canonical scaling we enforced $y_i(w^\top x_i + b) \geq 1$. The perpendicular distance from hyperplane to the nearest point is $\min_i \frac{y_i(w^\top x_i + b)}{\|w\|} = \frac{1}{\|w\|}$. Thus margin between the two support hyperplanes is $\frac{2}{\|w\|}$. So minimizing $\|w\|^2/2$ maximizes the margin.

3. Karush–Kuhn–Tucker (KKT) conditions (hard-margin)

KKT conditions give necessary and sufficient optimality conditions (convex QP implies sufficiency). They are:

1. Primal feasibility: $y_i(w^\top x_i + b) \geq 1$.
2. Dual feasibility: $\alpha_i \geq 0$.
3. Stationarity: $w = \sum_i \alpha_i y_i x_i$, $\sum_i \alpha_i y_i = 0$.
4. Complementary slackness: $\alpha_i [y_i(w^\top x_i + b) - 1] = 0$ for each i .

From complementary slackness:

- If $\alpha_i > 0$, then $y_i(w^\top x_i + b) - 1 = 0$: the point lies on the margin (support vector).
- If $y_i(w^\top x_i + b) > 1$, then $\alpha_i = 0$: non-support interior point.

4. Soft-margin SVM (non-separable or noisy data)

Allow slack variables $\xi_i \geq 0$ to tolerate violations. Primal problem:

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

$C > 0$ trades off margin size and penalty for margin violation. Larger $C \rightarrow$ fewer violations but smaller margin (more strongly penalized errors).

4.1 Soft-margin dual

Form Lagrangian with multipliers $\alpha_i \geq 0$ for main constraints and $\mu_i \geq 0$ for $\xi_i \geq 0$:

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

Stationarity:

- $\partial L / \partial w = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i.$
- $\partial L / \partial b = 0 \Rightarrow \sum_i \alpha_i y_i = 0.$
- $\partial L / \partial \xi_i = 0 \Rightarrow C - \alpha_i - \mu_i = 0 \Rightarrow 0 \leq \alpha_i \leq C.$

Dual becomes:

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

So only difference is box constraint $0 \leq \alpha_i \leq C$. Points with $0 < \alpha_i < C$ lie exactly on the margin, points with $\alpha_i = C$ are margin-violating/error points, and $\alpha_i = 0$ are interior non-support points.

KKT complementarity for slack:

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

5. Kernel trick — nonlinear separation

Replace inner products $x_i^\top x_j$ with kernel function $K(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$ where ϕ maps inputs to a (possibly high-dimensional) feature space. Dual problem becomes:

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

SS ROY

Decision function:

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

Common kernels: linear $K(x, z) = x^\top z$, polynomial $(x^\top z + c)^d$, RBF/Gaussian $K(x, z) = \exp(-\|x - z\|^2 / (2\sigma^2))$.

6. Full derivation of the dual (step-by-step)

Starting primal hard-margin:

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

Form Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2} w^\top w - \sum_i \alpha_i [y_i(w^\top x_i + b) - 1].$$

Stationary conditions:

- $\nabla_w L = w - \sum_i \alpha_i y_i x_i = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i$.
- $\partial L / \partial b = - \sum_i \alpha_i y_i = 0$.

Plugging w back:

$$\begin{aligned} L(w, b, \alpha) &= \frac{1}{2} \left(\sum_i \alpha_i y_i x_i \right)^\top \left(\sum_j \alpha_j y_j x_j \right) - \sum_i \alpha_i (y_i ((\sum_j \alpha_j y_j x_j)^\top x_i + b) - 1) \\ &= \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_i \alpha_i y_i \sum_j \alpha_j y_j x_j^\top x_i - b \sum_i \alpha_i y_i + \sum_i \alpha_i \\ &= -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_i \alpha_i, \end{aligned}$$

using $\sum_i \alpha_i y_i = 0$ and symmetry $x_i^\top x_j = x_j^\top x_i$. So dual objective:

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j,$$

subject to $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$. That proves the dual formulation.

7. Connection to hinge loss and primal unconstrained form

The soft-margin primal can be written as unconstrained minimization:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i(w^\top x_i + b)),$$

where $\max(0, 1 - z)$ is the hinge loss.

8. Practical observations

- Support vectors completely determine w and classification complexity often depends on number of support vectors.
- Solve the dual QP via specialized solvers (e.g., SMO) — especially efficient when n is modest and p is large after kernels.
- Scaling features (zero mean, unit variance) is important for performance.
- For large datasets, primal methods or linear SVM solvers (e.g., LIBLINEAR) are used.

9. Short proofs / reminders (compact)

Margin formula

If the hyperplane is $w^\top x + b = 0$. Distance from point x to hyperplane is $|w^\top x + b| / \|w\|$. Under canonical scaling the closest points satisfy $|w^\top x_i + b| = 1$. So distance to margin plane = $1 / \|w\|$, margin between classes = $2 / \|w\|$.

Stationarity \rightarrow representation

From $\nabla_w L = 0$, $w = \sum_i \alpha_i y_i x_i$. This proves that the optimal w is a linear combination of training points (support vectors).

Dual convexity

Dual is concave in α (negative definite quadratic term), constraints are linear \rightarrow global maximum is found by convex QP solvers.

10. Example (small numeric)

Given two 2-D points $x_1 = (1, 1), y_1 = +1$ and $x_2 = (2, 2), y_2 = -1$ they are not separable by margin > 0 obviously, but you can use soft-margin with C chosen

appropriately.

