# 4 Support Vector Machines (SVM)

**Goal**

Define and understand the classical hard-margin SVM for binary classification. Dual view.

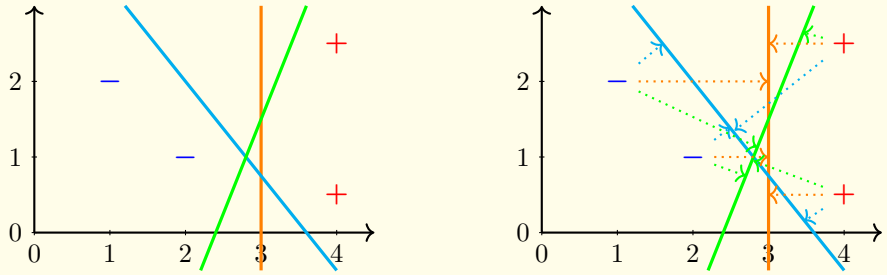
**Alert 4.1: Convention**

Gray boxes are not required hence can be omitted for unenthusiastic readers.

For less mathematical readers, think of the norm $\|\cdot\|$ and its dual norm $\|\cdot\|_\circ$ as the Euclidean $\ell_2$ norm $\|\cdot\|_2$. Treat all distances as the Euclidean distance. All of our pictures are for this special case.

This note is likely to be updated again soon.

**Definition 4.2: SVM as maximizing minimum distance**




Given a (strictly) linearly separable dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathsf{y}_i) \subseteq \mathbb{R}^d \times \{\pm 1\} : i = 1, \dots, n\}$, there exists a separating hyperplane $H_{\mathbf{w}} = \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{x}, \mathbf{w}\rangle + b = 0\}$, namely that

$$\forall i,\ \mathsf{y}_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) > 0.$$

In fact, there exist infinitely many separating hyperplanes: if we perturb $(\mathbf{w}, b)$ *slightly*, the resulting hyperplane would still be separating, thanks to continuity. Is there a particular separating hyperplane that stands out, and be "optimal"?

The answer is yes! Let $H_{\mathbf{w}}$ be any separating hyperplane (w.r.t. the given dataset $\mathcal{D}$). We can compute the distance from each training sample $\mathbf{x}_i$ to the hyperplane $H_{\mathbf{w}}$:

$$\begin{aligned}
\operatorname{dist}(\mathbf{x}_i, H_{\mathbf{w}}) &:= \min_{\mathbf{x}\in H_{\mathbf{w}}} \|\mathbf{x} - \mathbf{x}_i\|_\circ && \text{(e.g., the typical choice } \|\cdot\|_\circ = \|\cdot\| = \|\cdot\|_2) \\
&\geq \left| \frac{\langle \mathbf{x} - \mathbf{x}_i, \mathbf{w}\rangle + b - b}{\|\mathbf{w}\|} \right| && \text{(Cauchy-Schwarz, see Definition 1.25)} \\
&= \frac{|\langle \mathbf{x}_i, \mathbf{w}\rangle + b|}{\|\mathbf{w}\|} && \text{(equality at } \mathbf{x} = \mathbf{x}_i - \tfrac{\mathbf{z}}{\|\mathbf{w}\|^2}(\langle \mathbf{x}_i, \mathbf{w}\rangle + b),\ \underbrace{\langle \mathbf{z}, \mathbf{w}\rangle = \|\mathbf{w}\|^2, \|\mathbf{z}\|_\circ = \|\mathbf{w}\|}_{\mathbf{z}\in\partial\left[\frac{1}{2}\|\mathbf{w}\|^2\right]}) \\
&= \frac{\mathsf{y}_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b)}{\|\mathbf{w}\|} && (\mathsf{y}_i \in \{\pm 1\} \text{ and } H_{\mathbf{w}} \text{ is separating}). \qquad (4.1)
\end{aligned}$$

Here and in the following, we always assume w.l.o.g. that the dataset $\mathcal{D}$ contains at least 1 positive example and 1 negative example, so that $\mathbf{w} = \mathbf{0}$ with any $b$ cannot be a separating hyperplane.

Among all separating hyperplanes, support vector machines (SVM) tries to find one that maximizes the minimum distance (with the typical choice $\|\cdot\| = \|\cdot\|_2$ in mind):

$$\max_{\mathbf{w}:\forall i, \mathsf{y}_i\hat{y}_i > 0} \ \min_{i=1,\dots,n} \frac{\mathsf{y}_i\hat{y}_i}{\|\mathbf{w}\|}, \quad \text{where} \quad \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w}\rangle + b. \qquad (4.2)$$

We remark that the above formulation is scaling-invariant: If $\mathbf{w} = (\mathbf{w}, b)$ is optimal, then so is $\gamma\mathbf{w}$ for any $\gamma > 0$ (the fraction is unchanged and the constraint on $\mathbf{w}$ is not affected). This is not at all surprising, as $\mathbf{w}$ and $\gamma\mathbf{w}$ really represent the same hyperplane: $H_{\mathbf{w}} = H_{\gamma\mathbf{w}}$. Note also that the separating condition $\forall i, \mathsf{y}_i\hat{y}_i > 0$ can be omitted since it is automatically satisfied if the dataset $\mathcal{D}$ is indeed (strictly) linearly separable.

**Exercise 4.3: Alternative: minimizing the maximal distance?**

Use an example to show the difference between minimizing the maximal distance vs. maximizing the minimal distance. Which one do you prefer?

**Alert 4.4: Margin as minimum distance**

We repeat the formula in Definition 4.2:

$$\operatorname{dist}(\mathbf{x}, H_{\mathbf{w}}) := \left[\min_{\mathbf{z}\in H_{\mathbf{w}}} \|\mathbf{z}-\mathbf{x}\|_\circ\right] = \frac{|\langle \mathbf{x}, \mathbf{w}\rangle + b|}{\|\mathbf{w}\|} = \frac{\mathsf{y}(\langle \mathbf{x}, \mathbf{w}\rangle + b)}{\|\mathbf{w}\|} = \frac{\mathsf{y}\hat{y}}{\|\mathbf{w}\|},$$

where the third equality holds if $\mathsf{y}\hat{y} \geq 0$ and $\mathsf{y} \in \{\pm 1\}$. Given any hyperplane $H_{\mathbf{w}}$, we define its margin w.r.t. a data point $(\mathbf{x}, \mathsf{y})$ as:

$$\gamma((\mathbf{x}, \mathsf{y}); H_{\mathbf{w}}) := \frac{\mathsf{y}\hat{y}}{\|\mathbf{w}\|}, \quad \hat{y} = \langle \mathbf{x}, \mathbf{w}\rangle + b.$$

Geometrically, when the hyperplane $H_{\mathbf{w}}$ classifies the data point $(\mathbf{x}, \mathsf{y})$ correctly (i.e., $\mathsf{y}\hat{y} > 0$), this margin is exactly the distance from $\mathbf{x}$ to the hyperplane $H_{\mathbf{w}}$, and the negation of the distance otherwise.

Fixing any hyperplane $H_{\mathbf{w}}$, we can extend the notion of its margin to a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathsf{y}_i) : i = 1, \ldots, n\}$ by taking the (worst-case) minimum:

$$\gamma(\mathcal{D}; H_{\mathbf{w}}) := \left[\min_{i=1,\ldots,n} \gamma((\mathbf{x}_i, \mathsf{y}_i); H_{\mathbf{w}})\right] = \min_i \frac{\mathsf{y}_i\hat{y}_i}{\|\mathbf{w}\|}, \quad \hat{y}_i := \langle \mathbf{x}_i, \mathbf{w}\rangle + b.$$

Again, when the hyperplane $H_{\mathbf{w}}$ (strictly) separates the dataset $\mathcal{D}$, the margin $\gamma(\mathcal{D}; H_{\mathbf{w}}) > 0$ coincides with the minimum distance, as we saw in Definition 4.2. However, when $\mathcal{D}$ is not (strictly) separated by $H_{\mathbf{w}}$, the margin $\gamma(\mathcal{D}; H_{\mathbf{w}}) \leq 0$ is the negation of the maximum distance among all wrongly classified data points.

We can finally define the margin of a dataset $\mathcal{D}$ as the (best-case) maximum among all hyperplanes:

$$\gamma(\mathcal{D}) := \left[\max_{\mathbf{w}} \gamma(\mathcal{D}; H_{\mathbf{w}})\right] = \max_{\mathbf{w}} \min_{i=1,\ldots,n} \frac{\mathsf{y}_i\hat{y}_i}{\|\mathbf{w}\|}. \tag{4.3}$$

Again, when the dataset $\mathcal{D}$ is (strictly) linearly separable, the margin $\gamma(\mathcal{D}) > 0$ reduces to the minimum distance to the SVM hyperplane, in which case the margin definition here coincides with what we saw in Remark 1.30 (with the choice $\|\cdot\|_\circ = \|\cdot\| = \|\cdot\|_2$) and characterizes "how linearly separable" our dataset $\mathcal{D}$ is. On the other hand, when $\mathcal{D}$ is not (strictly) linearly separable, the margin $\gamma(\mathcal{D}) \leq 0$.

To summarize, hard-margin SVM, as defined in Definition 4.2, maximizes the margin among all hyperplanes on a (strictly) linearly separable dataset. Interestingly, with this interpretation, the hard-margin SVM formulation (4.3) continues to make sense even on a linearly inseparable dataset.

In the literature, sometimes people often call the unnormalized quantity $\mathsf{y}\hat{y}$ margin, which is fine as long as the scale $\|\mathbf{w}\|$ is kept constant.

**Definition 4.5: Alternative definition of margin**

We give a slightly different definition of margin here: $\gamma^+$. As the notation suggests, $\gamma^+$ coincides with the definition in Alert 4.4 on a (strictly) linearly separable dataset, and reduces to 0 otherwise.

- Given any hyperplane $H_{\mathbf{w}}$, we define its margin w.r.t. a data point $(\mathbf{x}, \mathsf{y})$ as:

$$\gamma^+((\mathbf{x}, \mathsf{y}); H_{\mathbf{w}}) := \frac{(\mathsf{y}\hat{y})^+}{\|\mathbf{w}\|}, \quad \hat{y} = \langle \mathbf{x}, \mathbf{w} \rangle + b,$$

  where recall $(t)^+ = \max\{t, 0\}$ is the positive part. Geometrically, when the hyperplane $H_{\mathbf{w}}$ classifies the data point $(\mathbf{x}, \mathsf{y})$ correctly (i.e., $\mathsf{y}\hat{y} \geq 0$), this margin is exactly the distance from $\mathbf{x}$ to the hyperplane $H_{\mathbf{w}}$, and 0 otherwise.

- Fixing any hyperplane $H_{\mathbf{w}}$, we can extend the notion of its margin to a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathsf{y}_i) : i = 1, \ldots, n\}$ by taking the (worst-case) minimum:

$$\gamma^+(\mathcal{D}; H_{\mathbf{w}}) := \left[\min_{i=1,\ldots,n} \gamma^+((\mathbf{x}_i, \mathsf{y}_i); H_{\mathbf{w}})\right] = \min_i \frac{(\mathsf{y}_i \hat{y}_i)^+}{\|\mathbf{w}\|}, \quad \hat{y}_i := \langle \mathbf{x}_i, \mathbf{w} \rangle + b.$$

  Again, when the hyperplane $H_{\mathbf{w}}$ (strictly) separates the dataset $\mathcal{D}$, the margin $\gamma^+(\mathcal{D}; H_{\mathbf{w}}) > 0$ coincides with the minimum distance, as we saw in Definition 4.2. However, when $\mathcal{D}$ is not (strictly) separated by $H_{\mathbf{w}}$, the margin $\gamma^+(\mathcal{D}; H_{\mathbf{w}}) = 0$.

- We can finally define the margin of a dataset $\mathcal{D}$ as the (best-case) maximum among all hyperplanes:

$$\gamma^+(\mathcal{D}) := \left[\max_{\mathbf{w}} \gamma^+(\mathcal{D}; H_{\mathbf{w}})\right] = \max_{\mathbf{w}} \min_{i=1,\ldots,n} \frac{[\mathsf{y}_i \hat{y}_i]^+}{\|\mathbf{w}\|}.$$

  Again, when the dataset $\mathcal{D}$ is (strictly) linearly separable, the margin $\gamma^+(\mathcal{D})$ reduces to the minimum distance to the SVM hyperplane. In contrast, when $\mathcal{D}$ is not (strictly) linearly separable, the margin $\gamma^+(\mathcal{D}) = 0$.

**Remark 4.6: Important standardization trick**

A simple *standardization* trick in optimization is to introduce an extra variable so that we can reduce an arbitrary objective function to the canonical linear function. For instance, if we are interested in solving

$$\min_{\mathbf{w}} \quad f(\mathbf{w}),$$

where $f$ can be any complicated nonlinear function. Upon introducing an extra variable $t$, we can reformulate our minimization problem equivalently as:

$$\min_{(\mathbf{w}, t): f(\mathbf{w}) \leq t} \quad t,$$

where the new objective $(\mathbf{0}; 1)^\top (\mathbf{w}; t)$ is a simple linear function of $(\mathbf{w}; t)$. The expense, of course, is that we have to deal with the extra constraint $f(\mathbf{w}) \leq t$ now.

**Remark 4.7: Removing homogeneity by normalizing direction**

To remove the scaling-invariance mentioned in Definition 4.2, we can restrict the direction vector $\mathbf{w}$ to have unit norm, which happened to yield the same formulation as that in Rosen (1965) (see Remark 4.20 below

for more details):

$$\max_{\mathbf{w}:\|\mathbf{w}\|=1} \min_{i=1,\dots,n} \mathsf{y}_i \hat{y}_i. \tag{4.4}$$

Applying the trick in Remark 4.6 (and noting we are maximizing here) yields the reformulation:

$$\max_{(\mathbf{w},\delta):\|\mathbf{w}\|=1} \delta, \quad \text{s.t.} \quad \min_{i=1,\dots,n} \mathsf{y}_i \hat{y}_i \geq \delta \iff \mathsf{y}_i \hat{y}_i \geq \delta, \ \forall i = 1, \dots, n,$$

which is completely equivalent to (4.3) (except by excluding out the trivial solution $\mathbf{w} = 0$).

Observe that on any linearly separable dataset, at optimality we can always achieve $\delta \geq 0$. Thus, we may relax the unit norm constraint on $\mathbf{w}$ slightly:

$$\begin{aligned} \max_{\mathbf{w},\delta} \ & \delta \\ \text{s.t.} \ & \|\mathbf{w}\| \leq 1 \\ & \mathsf{y}_i \hat{y}_i \geq \delta, \ \forall i = 1, \dots, n. \end{aligned} \tag{4.5}$$

It is clear if the dataset $\mathcal{D}$ is indeed linearly separable, at maximum we may choose $\|\mathbf{w}\| = 1$, hence the "relaxation" is in fact equivalent (on any linearly separable dataset that consists of at least 1 positive and 1 negative).

Rosen, J. (1965). "Pattern separation by convex programming". *Journal of Mathematical Analysis and Applications*, vol. 10, no. 1, pp. 123–134.

**Exercise 4.8: Detecting linear separability**

Prove an additional advantage of the "relaxation" (4.5): Its maximum value is always greater than 0, which is attained iff the dataset is not (strictly) linearly separable.

In contrast, prove that the original formulation (4.4) with *exact unit norm constraint*

- is equivalent to (4.5) with strictly positive maximum value, iff the dataset is (strictly) linearly separable;
- is different from (4.5) with strictly negative maximum value, iff the dataset is not (strictly) linearly separable and the intersection of positive and negative convex hulls has nonempty (relative) interior;
- is similar to (4.5) with exactly 0 maximum value, iff the dataset is not (strictly) linearly separable and the intersection of positive and negative convex hulls has empty (relative) interior.

**Remark 4.9: Linear separability, revisited**

Recall our definition of (strict) linear separability of a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathsf{y}_i) \in \mathbb{R}^d \times \{\pm 1\} : i = 1, \dots, n\}$:

$$\exists \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}, s > 0, \text{ such that } \mathsf{y}_i \hat{y}_i \geq s, \ \forall i = 1, \dots, n, \quad \text{where} \quad \hat{y}_i := \langle \mathbf{x}_i, \mathbf{w} \rangle + b.$$

Let us now break the above condition for any positive example $\mathsf{y}_i = 1$ and any negative example $\mathsf{y}_j = -1$:

$$\begin{aligned} \langle \mathbf{x}_i, \mathbf{w} \rangle + b \geq s \geq -s \geq \langle \mathbf{x}_j, \mathbf{w} \rangle + b &\iff \langle \mathbf{x}_i, \mathbf{w} \rangle \geq s - b \geq -s - b \geq \langle \mathbf{x}_j, \mathbf{w} \rangle \\ &\iff \min_{i:\mathsf{y}_i=1} \langle \mathbf{x}_i, \mathbf{w} \rangle > \max_{j:\mathsf{y}_j=-1} \langle \mathbf{x}_j, \mathbf{w} \rangle . \end{aligned}$$

It is clear now that the linear separability condition has nothing to do with the offset term $b$ but the normal vector $\mathbf{w}$.

**Remark 4.10: Removing homogeneity by normalizing offset**

A different way to remove the scaling-invariance mentioned in Definition 4.2 is to perform normalization on the offset so that

$$\min_{i=1,\dots,n} \mathsf{y}_i \hat{y}_i = \delta,$$

where $\delta > 0$ is any fixed constant. When the dataset $\mathcal{D}$ is indeed (strictly) linearly separable, this normalization can always be achieved (simply by scaling $\mathbf{w}$). After normalizing this way, we can simplify (4.2) as:

$$\max_{\mathbf{w}} \ \frac{\delta}{\|\mathbf{w}\|}, \quad \text{s.t.} \quad \min_{i=1,\dots,n} \mathsf{y}_i \hat{y}_i = \delta.$$

We remind again that $\delta$ here is any fixed positive constant and we are *not* optimizing it (in contrast to what we did in Remark 4.7). Applying some elementary transformations (that do not change the minimizer) we arrive at the usual formulation of SVM (due to Boser et al. (1992)):

$$\begin{aligned} \min_{\mathbf{w}} \ & \tfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \ & \mathsf{y}_i \hat{y}_i \geq \delta, \ \forall i = 1, \dots, n. \end{aligned} \tag{4.6}$$

It is clear that the actual value of the positive constant $\delta$ is immaterial. Most often, we simply set $\delta = 1$, which is our default choice for the rest of this note.

The formulation (4.6) only makes sense on (strictly) linearly separable datasets, unlike our original formulation (4.3).

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). "A Training Algorithm for Optimal Margin Classiers". In: *COLT*, pp. 144–152.

**Alert 4.11: Any positive number but not zero**

Note that in the familiar SVM formulation (4.6), we can choose $\delta$ to be any (strictly) positive number (which amounts to a simple change of scale). However, we cannot set $\delta = 0$, for otherwise the solution could be trivially $\mathbf{w} = \mathbf{0}, b = 0$.

**Remark 4.12: Perceptron vs. SVM**

We can formulate perceptron as the following feasibility problem:

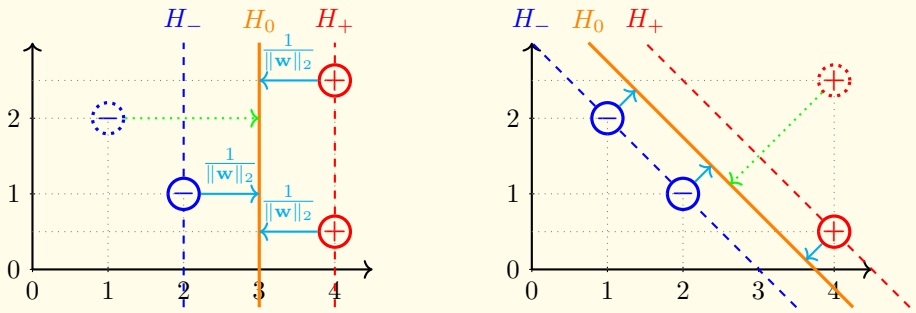
$$\begin{aligned} \min_{\mathbf{w}} \ & 0 \\ \text{s.t.} \ & \mathsf{y}_i \hat{y}_i \geq \delta, \ \forall i = 1, \dots, n, \end{aligned}$$

where as before $\delta > 0$ is any fixed constant.

Unlike SVM, the objective function of perceptron is the trivial constant 0 function, i.e., we are not trying to optimize anything (such as distance/margin) other than satisfying a bunch of constraints (separating the positives from the negatives). Computationally, perceptron belongs to linear programming (LP), i.e., when the objective function and all constraints are linear functions. In contrast, SVM belongs to the slightly more complicated quadratic programming (QP): the objective function is a quadratic function while all constraints are still linear. Needless to say, LP $\subsetneq$ QP.

**Remark 4.13: Three parallel hyperplanes**

Geometrically, we have the following intuitive picture. As an example, the dataset $\mathcal{D}$ consists of 2 positive and 2 negative examples. The left figure shows the SVM solution, and for comparison the right figure depicts a suboptimal solution. We will see momentarily why the left solution is optimal.




To understand the above figure, let us take a closer look at the SVM formulation (4.6), where w.l.o.g. we choose $\delta = 1$. Recall that the dataset $\mathcal{D}$ contains at least 1 positive example and 1 negative example (so that $\mathbf{w} = \mathbf{0}$ is ruled out). Let us breakdown the constraints in (4.6):

$$\left.\begin{array}{ll} \langle \mathbf{x}_i, \mathbf{w}\rangle + b \geq 1, & \mathsf{y}_i = 1 \\ \langle \mathbf{x}_i, \mathbf{w}\rangle + b \leq -1, & \mathsf{y}_i = -1 \end{array}\right\} \iff 1 - \min_{i:\mathsf{y}_i=1} \langle \mathbf{x}_i, \mathbf{w}\rangle \leq b \leq -1 - \max_{i:\mathsf{y}_i=-1} \langle \mathbf{x}_i, \mathbf{w}\rangle .$$

If one of the inequalities is strict, say the left one, then we can decrease $b$ slightly so that both inequalities are strict. But then we can scale down $\mathbf{w}$ and $b$ without violating any constraint while decreasing the objective $\frac{1}{2}\|\mathbf{w}\|^2$ further. Therefore, at minimum, we must have

$$1 - \min_{i:\mathsf{y}_i=1} \langle \mathbf{x}_i, \mathbf{w}\rangle = b = -1 - \max_{i:\mathsf{y}_i=-1} \langle \mathbf{x}_i, \mathbf{w}\rangle , \ \textit{i.e.,}\ \mathsf{y}_i\hat{y}_i = 1 \text{ for at least one } \mathsf{y}_i = 1 \text{ and one } \mathsf{y}_i = -1.$$

Given the SVM solution $(\mathbf{w}, b)$, we can now define three parallel hyperplanes:

$$\begin{aligned} H_0 &:= \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w}\rangle + b = 0\} \\ H_+ &:= \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w}\rangle + b = 1\} && \text{(we choose } \delta = 1) \\ H_- &:= \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w}\rangle + b = -1\}. \end{aligned}$$

The hyperplane $H_0$ is the decision boundary of SVM: any point above or below it is classified as positive or negative, respectively, i.e., $\hat{\mathsf{y}} = \operatorname{sign}(\langle \mathbf{x}, \mathbf{w}\rangle + b)$. The hyperplane $H_+$ is the translate of $H_0$ on which for the first time we pass through some positive examples, and similarly for $H_-$. Note that there are no training examples between $H_-$ and $H_+$ (a dead zone), with $H_0$ at the middle between $H_-$ and $H_+$. More precisely, we can compute the distance between $H_0$ and $H_+$:

$$\begin{aligned} \operatorname{dist}(H_+, H_0) &:= \min_{\mathbf{p}\in H_+} \min_{\mathbf{q}\in H_0} \|\mathbf{p} - \mathbf{q}\|_\circ \\ &= \min_{i:\mathsf{y}_i=1} \operatorname{dist}(\mathbf{x}_i, H_0) && \text{(since } H_+ \text{ first passes through positive examples)} \\ &= \frac{1}{\|\mathbf{w}\|} && \text{(see (4.1))} \\ &= \min_{i:\mathsf{y}_i=-1} \operatorname{dist}(\mathbf{x}_i, H_0) && \text{(since } H_- \text{ first passes through negative examples)} \\ &= \operatorname{dist}(H_-, H_0). \end{aligned}$$

**Exercise 4.14: Uniqueness of w**

For the $\ell_2$ norm, prove the parallelogram equality

$$\|\mathbf{w}_1 + \mathbf{w}_2\|_2^2 + \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2 = 2(\|\mathbf{w}_1\|_2^2 + \|\mathbf{w}_2\|_2^2).$$

(The parallelogram law, in fact, characterizes norms that are induced by an inner product). With this choice $\|\cdot\| = \|\cdot\|_2$, prove

- that the SVM weight vector $\mathbf{w}$ is unique;
- that the SVM offset $b$ is also unique.

**Definition 4.15: Convex set**

A set $C \subseteq \mathbb{R}^d$ is called convex iff for all $\mathbf{x}, \mathbf{z} \in C$ and for all $\alpha \in [0, 1]$ we have

$$(1 - \alpha)\mathbf{x} + \alpha\mathbf{z} \in C,$$

i.e., the line segment connecting any two points in $C$ remains in $C$.

By convention the empty set is convex. Obviously, the universe $\mathbb{R}^d$, being a vector space, is convex.

**Exercise 4.16: Basic properties of convex sets**

Prove the following:

- The intersection $\bigcap_{\gamma\in\Gamma} C_\gamma$ of a collection of convex sets $\{C_\gamma\}_{\gamma\in\Gamma}$ is convex.
- A set in $\mathbb{R}$ (the real line) is convex iff it is an interval (not necessarily bounded or closed).
- The union of two convex sets need not be convex.
- The complement of a convex set need not be convex.
- Hyperplanes $H_0 := \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{x}, \mathbf{w}\rangle + b = 0\}$ are convex.
- Halfspaces $H_\leq := \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{x}, \mathbf{w}\rangle + b \leq 0\}$ are convex.

(In fact, a celebrated result in convex analysis shows that any closed convex set is an intersection of halfspaces.)

**Definition 4.17: Convex hull**

The convex hull $\operatorname{conv}(A)$ of an arbitrary set $A$ is the intersection of all convex supersets of $A$, i.e.,

$$\operatorname{conv}(A) := \bigcap_{\text{convex } C \supseteq A} C.$$

In other words, the convex hull is the "smallest" convex superset.

**Exercise 4.18: Convex hull as convex combination**

We define the convex combination of a finite set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as any point $\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{x}_i$ with

coefficients $\boldsymbol{\alpha} \geq 0, \mathbf{1}^\top \boldsymbol{\alpha} = 1$, i.e., $\boldsymbol{\alpha} \in \Delta_{n-1}$. Prove that for any $A \subseteq \mathbb{R}^d$:

$$\operatorname{conv}(A) = \left\{ \mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i : n \in \mathbb{N}, \boldsymbol{\alpha} \in \Delta_{n-1}, \mathbf{x}_i \in A \right\},$$

i.e., the convex hull is simply the set of all convex combinations of points in $A$.

(The celebrated Carathéodory theorem allows us to restrict $n \leq d+1$, and $n \leq d$ if $A$ is connected.)

**Exercise 4.19: Unit balls of norms are convex**

Recall that the unit ball of the $\ell_p$ "norm" is defined as:

$$\mathsf{B}_p := \{\mathbf{x} : \ \|\mathbf{x}\|_p \leq 1\},$$

which is convex iff $p \geq 1$. The following figure shows the unit ball $\mathsf{B}_p$ for $p = 2, \infty, \frac{1}{2}, 1$.

As shown above:

$$\operatorname{conv}(\mathsf{B}_{\frac{1}{2}}) = \mathsf{B}_1.$$

- For what values of $p$ and $q$ do we have $\operatorname{conv}(\mathsf{B}_p) = \mathsf{B}_q$?
- For what value of $p$ is the sphere $\mathsf{S}_p := \{\mathbf{x} : \ \|\mathbf{x}\|_p = 1\} = \partial \mathsf{B}_p$ convex?

**Remark 4.20: The first dual view of SVM (Rosen 1965)**

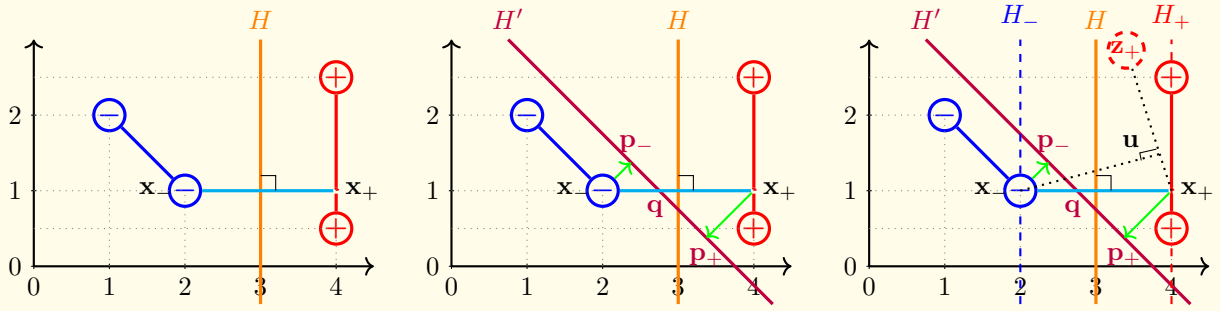
Rosen (1965) was among the first few people who recognized that a dataset $\mathcal{D}$ is (strictly) linearly separable (see Definition 1.24) iff

$$\operatorname{conv}(\mathcal{D}^+) \cap \operatorname{conv}(\mathcal{D}^-) = \emptyset, \quad \text{where} \quad \mathcal{D}^\pm := \{\mathbf{x}_i \in \mathcal{D} : \mathsf{y}_i = \pm 1\}.$$

(Prove the only if part by yourself; to see the if part, note that the convex hull of a compact set (e.g., finite set) is compact, and disjoint compact sets can be strictly separated by a hyperplane, due to the celebrated Hahn-Banach Theorem.)

Rosen, J. (1965). "Pattern separation by convex programming". *Journal of Mathematical Analysis and Applications*, vol. 10, no. 1, pp. 123–134.

**Remark 4.21: Dual view of SVM, as bisector of minimum distance pair**




In Definition 4.2 we defined SVM as maximizing the minimum distance of training examples to the decision boundary $H_0$. We now provide a dual view which geometrically is very appealing.

- We first make a simple observation about a (strict) separating hyperplane $H$:

$$\left.\begin{array}{ll}\langle \mathbf{x}_i, \mathbf{w}\rangle + b > 0, & \text{if } \mathbf{x}_i \in \mathcal{D}^+ := \{\mathbf{x}_j : \mathrm{y}_j = 1\} \\ \langle \mathbf{x}_i, \mathbf{w}\rangle + b < 0, & \text{if } \mathbf{x}_i \in \mathcal{D}^- := \{\mathbf{x}_j : \mathrm{y}_j = -1\}\end{array}\right\} \implies \begin{cases}\langle \mathbf{x}, \mathbf{w}\rangle + b > 0, & \text{if } \mathbf{x} \in \operatorname{conv}(\mathcal{D}^+) \\ \langle \mathbf{x}, \mathbf{w}\rangle + b < 0, & \text{if } \mathbf{x} \in \operatorname{conv}(\mathcal{D}^-)\end{cases},$$

  i.e., $H$ also (strictly) separates the convex hulls of positive examples and negative ones.

- The second observation we make is about the minimum distance of all positive (negative) examples to a separating hyperplane:

$$\min_{\mathbf{x}\in\mathcal{D}^\pm} \operatorname{dist}(\mathbf{x}, H) = \min_{\mathbf{x}\in\mathcal{D}^\pm} \frac{\pm(\langle \mathbf{x}, \mathbf{w}\rangle + b)}{\|\mathbf{w}\|} = \min_{\mathbf{x}\in\operatorname{conv}(\mathcal{D}^\pm)} \frac{\pm(\langle \mathbf{x}, \mathbf{w}\rangle + b)}{\|\mathbf{w}\|} = \min_{\mathbf{x}\in\operatorname{conv}(\mathcal{D}^\pm)} \operatorname{dist}(\mathbf{x}, H),$$

  where the first equality follows from (4.1), the second from linearity, and the third from our observation above. In other words, we could replace the datasets $\mathcal{D}^\pm$ with their convex hulls.

- Based on the second observation, we now find the pair of $\mathbf{x}_+ \in \operatorname{conv}(\mathcal{D}_+)$ and $\mathbf{x}_- \in \operatorname{conv}(\mathcal{D}_-)$ so that $\operatorname{dist}(\mathbf{x}_+, \mathbf{x}_-)$ achieves the minimum distance among all pairs from the two convex hulls. We connect the segment from $\mathbf{x}_+$ to $\mathbf{x}_-$ and find its bisector, a separating hyperplane $H$ that passes the middle point $\frac{1}{2}(\mathbf{x}_+ + \mathbf{x}_-)$ with normal vector proportional to $\partial\left[\frac{1}{2}\|\mathbf{x}_+ - \mathbf{x}_-\|^2\right]$. We claim that

$$\min_{\mathbf{x}\in\mathcal{D}^\pm} \operatorname{dist}(\mathbf{x}, H) = \min_{\mathbf{x}\in\operatorname{conv}(\mathcal{D}^\pm)} \operatorname{dist}(\mathbf{x}, H) = \tfrac{1}{2}\operatorname{dist}(\mathbf{x}_+, \mathbf{x}_-) = \tfrac{1}{2}\operatorname{dist}(\operatorname{conv}(\mathcal{D}^+), \operatorname{conv}(\mathcal{D}^-)).$$

  To see the second equality, we translate $H$ in parallel until it passes $\mathbf{x}_+$ and $\mathbf{x}_-$, and obtain hyperplanes $H_+$ and $H_-$, respectively. Since $H$ is a bisector of the line segment $\mathbf{x}_+\mathbf{x}_-$,

$$\operatorname{dist}(H_+, H) = \operatorname{dist}(H_-, H) = \tfrac{1}{2}\operatorname{dist}(\mathbf{x}_+, \mathbf{x}_-).$$

  We are left to prove there is no point in $\operatorname{conv}(\mathcal{D}^\pm)$ that lies between $H_-$ and $H_+$. Suppose, for the sake of contradiction, there is some $\mathbf{z}_+ \in \operatorname{conv}(\mathcal{D}^+)$ that lies between $H_-$ and $H_+$. The remaining proof for the Euclidean case where $\|\cdot\| = \|\cdot\|_2$ is depicted above: We know the angle $\angle \mathbf{x}_-\mathbf{x}_+\mathbf{z}_+ < 90°$. If we move a point $\mathbf{u}$ on the segment $\mathbf{z}_+\mathbf{x}_+$ from $\mathbf{z}_+$ to $\mathbf{x}_+$, because the angle $\angle \mathbf{u}\mathbf{x}_-\mathbf{x}_+ \to 0°$, so eventually we will have $\angle \mathbf{x}_-\mathbf{u}\mathbf{x}_+ \geq 90°$, in which case we would have $\operatorname{dist}(\mathbf{u}, \mathbf{x}_-) < \operatorname{dist}(\mathbf{x}_+, \mathbf{x}_-)$. Since $\mathbf{u} \in \operatorname{conv}(\mathcal{D}^+)$, we have a contradiction:

$$\operatorname{dist}(\mathbf{u}, \mathbf{x}_-) \geq \operatorname{dist}(\operatorname{conv}(\mathcal{D}^+), \operatorname{conv}(\mathcal{D}^-)) = \operatorname{dist}(\mathbf{x}_+, \mathbf{x}_-) > \operatorname{dist}(\mathbf{u}, \mathbf{x}_-).$$

  The proof for any norm is as follows: Since the line segment $\mathbf{z}_+\mathbf{x}_+ \in \operatorname{conv}(\mathcal{D}^+)$ and by definition $\operatorname{dist}(\mathbf{x}_+, \mathbf{x}_-) = \operatorname{dist}(\operatorname{conv}(\mathcal{D}^+), \operatorname{conv}(\mathcal{D}^-))$, we know for any $\mathbf{u}_\lambda = \lambda\mathbf{z}_+ + (1-\lambda)\mathbf{x}_+$ on the line segment, $f(\lambda) := \operatorname{dist}(\mathbf{u}_\lambda, \mathbf{x}_-) \geq \operatorname{dist}(\mathbf{x}_+, \mathbf{x}_-) = f(0)$, i.e., the minimum of $f(\lambda)$ over the interval $\lambda \in [0, 1]$ is achieved at $\lambda = 0$. Since $f(\lambda)$ is convex its right derivative at $\lambda = 0$, namely $\langle \mathbf{w}, \mathbf{z}_+ - \mathbf{x}_+\rangle$, where $\mathbf{w} \in \partial\|\mathbf{x}_+ - \mathbf{x}_-\|$, must be positive. But we know the hyperplane $H_+ = \{\mathbf{x} : \mathbf{w}^\top(\mathbf{x} - \mathbf{x}_+) = 0\}$ and the middle point $\frac{1}{2}(\mathbf{x}_+ + \mathbf{x}_-)$ is on the left side of $H_+$, hence $\mathbf{z}_+$ is on the right side of $H_+$, contradiction.

- We can finally claim that $H$ is the SVM solution, i.e., $H$ maximizes the minimum distance to every training examples in $\mathcal{D}$. Indeed, let $H'$ be any other separating hyperplane. According to our first observation above, $H'$ intersects with the line segment $\mathbf{x}_+\mathbf{x}_-$ at some point $\mathbf{q}$ (due to separability). Define $\mathbf{p}_\pm$ as the projection of $\mathbf{x}_\pm$ onto the hyperplane $H'$, and since $\mathbf{q} \in H'$,

$$\operatorname{dist}(\mathbf{x}_\pm, \mathbf{p}_\pm) = \operatorname{dist}(\mathbf{x}_\pm, H') \le \operatorname{dist}(\mathbf{x}_\pm, \mathbf{q}).$$

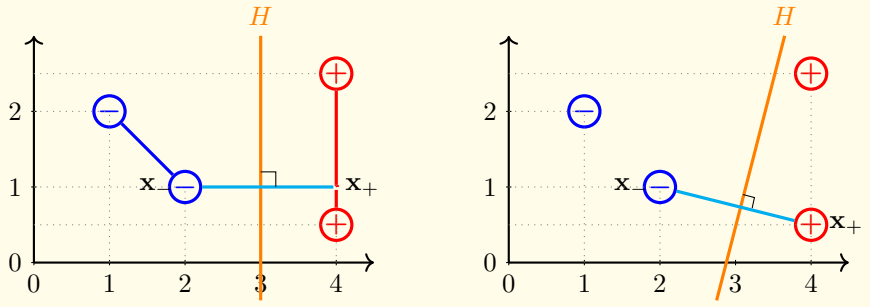
Therefore, using our second and third observations above:

$$\begin{aligned}
\min_{\mathbf{x}\in\mathcal{D}^\pm} \operatorname{dist}(\mathbf{x}, H') = \min_{\mathbf{x}\in\operatorname{conv}(\mathcal{D}^\pm)} \operatorname{dist}(\mathbf{x}, H') &\le \operatorname{dist}(\mathbf{x}_+, \mathbf{p}_+) \wedge \operatorname{dist}(\mathbf{x}_-, \mathbf{p}_-) \\
&\le \tfrac{1}{2}[\operatorname{dist}(\mathbf{x}_+, \mathbf{p}_+) + \operatorname{dist}(\mathbf{x}_-, \mathbf{p}_-)] \\
&\le \tfrac{1}{2}[\operatorname{dist}(\mathbf{x}_+, \mathbf{q}) + \operatorname{dist}(\mathbf{x}_-, \mathbf{q})] \\
&= \tfrac{1}{2}\operatorname{dist}(\mathbf{x}_+, \mathbf{x}_-) \\
&= \min_{\mathbf{x}\in\operatorname{conv}(\mathcal{D}^\pm)} \operatorname{dist}(\mathbf{x}, H) = \min_{\mathbf{x}\in\mathcal{D}^\pm} \operatorname{dist}(\mathbf{x}, H).
\end{aligned}$$

**Exercise 4.22: Necessity of convex hull**

In Remark 4.21, we picked the pair $\mathbf{x}_+$ and $\mathbf{x}_-$ from the two convex hulls $\mathcal{D}^\pm$ of the positive and negative examples, respectively. Prove the following:

- One of $\mathbf{x}_+$ and $\mathbf{x}_-$ can be chosen from the original datasets $\mathcal{D}^\pm$.
- Not both of $\mathbf{x}_+$ and $\mathbf{x}_-$ may be chosen from the original datasets $\mathcal{D}^\pm$.
- What observation(s) in Remark 4.21 might fail if we insist in picking both $\mathbf{x}_+$ and $\mathbf{x}_-$ from the original datasets $\mathcal{D}^\pm$?




**Remark 4.23: SVM dual, from geometry to algebra**

We complement the geometric dual view of SVM in Remark 4.21 with a "simpler" algebraic view. Applying scaling we may assume the weight vector $\mathbf{w}$ of a separating hyperplane $H_\mathbf{w}$ is normalized. Then, we maximize the minimum distance as follows:

$$\begin{aligned}
\max_{\|\mathbf{w}\|=1,b} \operatorname{dist}(\mathcal{D}^+, H_\mathbf{w}) \wedge \operatorname{dist}(\mathcal{D}^-, H_\mathbf{w}) &= \max_{\|\mathbf{w}\|=1,b} \left[\min_{\mathbf{x}_+\in\mathcal{D}^+} (\langle \mathbf{x}_+, \mathbf{w}\rangle + b) \wedge \min_{\mathbf{x}_-\in\mathcal{D}^-} -(\langle \mathbf{x}_-, \mathbf{w}\rangle + b)\right] \\
&= \max_{\|\mathbf{w}\|=1,b} \left[\min_{\mathbf{x}_\pm\in\mathcal{D}^\pm, t\in[0,1]} t(\langle \mathbf{x}_+, \mathbf{w}\rangle + b) + (1-t)(-\langle \mathbf{x}_-, \mathbf{w}\rangle - b)\right] \\
&= \max_{\|\mathbf{w}\|\le 1,b} \left[\min_{\mathbf{x}_+\in t\operatorname{conv}(\mathcal{D}^+), \mathbf{x}_-\in(1-t)\operatorname{conv}(\mathcal{D}_-), t\in[0,1]} \langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w}\rangle + b(2t-1)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \min_{\mathbf{x}_+ \in t\,\mathrm{conv}(\mathcal{D}^+), \mathbf{x}_- \in (1-t)\,\mathrm{conv}(\mathcal{D}_-), t\in[0,1]} \max_{\|\mathbf{w}\|\le 1, b} [\langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w}\rangle + b(2t-1)] \\
&= \min_{\mathbf{x}_+ \in \frac{1}{2}\,\mathrm{conv}(\mathcal{D}^+), \mathbf{x}_- \in \frac{1}{2}\,\mathrm{conv}(\mathcal{D}_-)} \max_{\|\mathbf{w}\|\le 1} \langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w}\rangle \\
&= \min_{\mathbf{x}_+ \in \frac{1}{2}\,\mathrm{conv}(\mathcal{D}^+), \mathbf{x}_- \in \frac{1}{2}\,\mathrm{conv}(\mathcal{D}_-)} \|\mathbf{x}_+ - \mathbf{x}_-\|_\circ \\
&= \tfrac{1}{2}\mathrm{dist}(\mathrm{conv}(\mathcal{D}^+), \mathrm{conv}(\mathcal{D}^-)),
\end{aligned}
$$

where in the third equality we used linearity to replace with convex hulls, which then allowed us to apply the minimax theorem to swap max with min. The sixth equality follows from Cauchy-Schwarz and is attained when $\mathbf{w} \propto \mathbf{x}_+ - \mathbf{x}_-$, i.e., when $H_{\mathbf{w}}$ is a bisector.