7 Support Vector Machines (SVM)

Goal

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

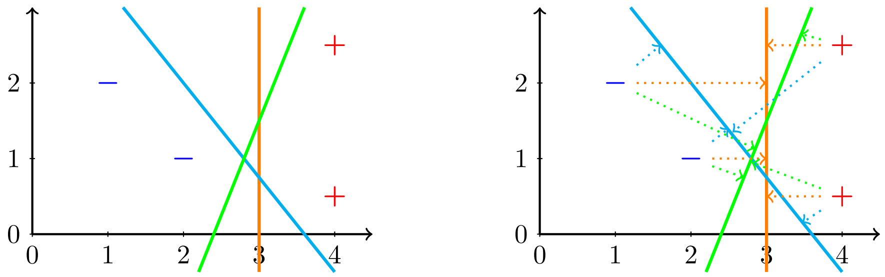
Alert 7.1: Convention

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 ℓ_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 7.2: SVM as maximizing minimum distance



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

$$\forall i, \ y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0.$$

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 $H_\mathbf{w}$:

$$\begin{aligned}
\operatorname{dist}(\mathbf{x}_i, H_\mathbf{w}) &:= \min_{\mathbf{x} \in H_\mathbf{w}} \|\mathbf{x} - \mathbf{x}_i\|_\circ && \text{(e.g., the typical choice } \|\cdot\|_\circ = \|\cdot\| = \|\cdot\|_2) \\
&\geq \left| \frac{\mathbf{w}^\top(\mathbf{x} - \mathbf{x}_i) + b - b}{\|\mathbf{w}\|} \right| && \text{(Cauchy-Schwarz, cf. Definition 1.23)} \\
&= \frac{|\mathbf{w}^\top \mathbf{x}_i + b|}{\|\mathbf{w}\|} && \text{(equality at } \mathbf{x} = \mathbf{x}_i - \tfrac{\mathbf{z}}{\|\mathbf{w}\|^2}(b + \mathbf{w}^\top \mathbf{x}_i), \ \underbrace{\mathbf{z}^\top \mathbf{w} = \|\mathbf{w}\|^2, \|\mathbf{z}\|_\circ = \|\mathbf{w}\|}_{\mathbf{z} \in \partial\left[\frac{1}{2}\|\mathbf{w}\|^2\right]}) \\
&= \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|} && (y_i \in \{\pm 1\} \text{ and } H_\mathbf{w} \text{ is separating}). \qquad (7.1)
\end{aligned}$$

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):

$$\max_{\mathbf{w}: \forall i, y_i \hat{y}_i > 0} \ \min_{i=1,\ldots,n} \frac{y_i \hat{y}_i}{\|\mathbf{w}\|}, \quad \text{where} \quad \hat{y}_i = \mathbf{w}^\top \mathbf{x}_i + b. \qquad (7.2)$$

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

Alert 7.3: Margin as minimum distance

We repeat the formula in Definition 7.2:

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

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

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

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

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

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

Again, when the hyperplane H_{w} (strictly) separates the dataset $\mathcal{D}$, the margin $\gamma(\mathcal{D}; H_{\mathsf{w}}) > 0$ coincides with the minimum distance, as we saw in Definition 7.2. However, when $\mathcal{D}$ is not (strictly) separated by H_{w}, the margin $\gamma(\mathcal{D}; H_{\mathsf{w}}) \leq 0$ 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:

$$\gamma(\mathcal{D}) := \left[\max_{\mathsf{w}} \gamma(\mathcal{D}; H_{\mathsf{w}})\right] = \max_{\mathsf{w}} \min_{i=1,\ldots,n} \frac{y_i\hat{y}_i}{\|\mathbf{w}\|}. \tag{7.3}$$

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 1.28 (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 7.2, maximizes the margin among all hyperplanes on a (strictly) linearly separable dataset. Interestingly, with this interpretation, the hard-margin SVM formulation (7.3) continues to make sense even on a linearly inseparable dataset.

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

Definition 7.4: Alternative definition of margin

We give a slightly different definition of margin here: γ^+. As the notation suggests, γ^+ coincides with the definition in Alert 7.3 on a (strictly) linearly separable dataset, and reduces to 0 otherwise.

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

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

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

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

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

  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 7.2. 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:

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

  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 7.5: 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

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

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

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

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 $f(\mathbf{w}) \le t$ now.

Remark 7.6: Removing homogeneity by normalizing direction

To remove the scaling-invariance mentioned in Definition 7.2, 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 7.23 below for more details):

$$\max_{\mathbf{w}:\|\mathbf{w}\|=1} \ \min_{i=1,\ldots,n} y_i \hat{\mathsf{y}}_i. \tag{7.4}$$

Applying the trick in Remark 7.5 (and noting we are maximizing here) yields the reformulation:

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

which is completely equivalent to (7.3) (except by excluding out the trivial solution $\mathbf{w} = 0$).

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

$$\begin{aligned} \max_{\mathbf{w},\delta} \ & \delta \\ \text{s.t.} \ & \|\mathbf{w}\| \le 1 \\ & y_i \hat{\mathsf{y}}_i \ge \delta, \ \forall i = 1, \ldots, n. \end{aligned} \tag{7.5}$$

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).

Note that (7.5) is exactly the bound we got for the perceptron algorithm, cf. Remark 1.28. Thus, SVM could have been derived by optimizing the convergence bound of perceptron. So, theory does inspire new algorithms, although this was not what really happened in history (Vapnik and Chervonenkis (1964) did not seem to be aware of Theorem 1.26 at the time, in fact they wrongly claimed that the perceptron algorithm may not find a solution even when one exists...).

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

Vapnik, Vladimir N. and A. Ya. Chervonenkis (1964). "On a class of perceptrons". *Automation and Remote Control*, vol. 25, no. 1, pp. 112–120.

Exercise 7.7: Detecting linear separability

Prove an additional advantage of the "relaxation" (7.5): 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 (7.4) with *exact unit norm constraint*

- is equivalent to (7.5) with strictly positive maximum value, iff the dataset is (strictly) linearly separable;
- is different from (7.5) 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 (7.5) 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 7.8: History of SVM

In this box we summarize the first SVM paper due to Vapnik and Lerner (1963). Our terminology and notation are different from the somewhat obscure original.

Let our universe be $\mathcal{X}$ and $\mathcal{D} \subseteq \mathcal{X}$ a training set. Vapnik and Lerner (1963) considered essentially the unsupervised setting, where labels are not provided even at training. Let $\varphi : \mathcal{X} \to \mathcal{S} \subseteq \mathcal{H}$ be a mapping that turns the original input $\mathbf{x} \in \mathcal{X}$ into a point $\mathbf{z} := \varphi(\mathbf{x})$ in the unit sphere $\mathcal{S} := \{\mathbf{z} \in \mathcal{H} : \|\mathbf{z}\|_2 = 1\}$ of a Hilbert space $\mathcal{H}$ (with induced norm $\|\cdot\|_2$). Our goal is to divide the data into c (disjoint) categories $\mathcal{C}_1, \ldots, \mathcal{C}_\mathsf{c}$, each of which is represented by a center $\mathbf{c}_k \in \mathcal{S}$ so that

$$\max_{i \in \mathcal{C}_k} \|\mathbf{z}_i - \mathbf{c}_k\|_2^2 < \min_{j \notin \mathcal{C}_k} \|\mathbf{z}_j - \mathbf{c}_k\|_2^2.$$

In other words, if we circumscribe the training examples in each category $\mathcal{C}_k$ by the smallest ball with center in the unit sphere $\mathcal{S}$, then these balls only contain training examples in the same category (in fact this could be how we define categories). (However, these balls may still intersect.) Since both $\mathbf{z}$ and $\mathbf{c}$ have unit norm, equivalently we may require

$$\min_{i \in \mathcal{C}_k} \langle \mathbf{z}_i, \mathbf{c}_k \rangle > \max_{j \notin \mathcal{C}_k} \langle \mathbf{z}_j, \mathbf{c}_k \rangle . \tag{7.6}$$

In other words, there exists a hyperplane with direction $\mathbf{c}_k$ that (strictly) separates the category $\mathcal{C}_k$ from the rest categories $\mathcal{C}_{\neg k} := \bigcup_{l \neq k} \mathcal{C}_l$. Let us define

$$r_k = \max_{i \in \mathcal{C}_k} \|\mathbf{z}_i - \mathbf{c}_k\|_2, \quad k = 1, \ldots, \mathsf{c},$$

so that we may declare any point $\mathbf{z} \in \mathsf{B}(\mathbf{c}_k, r_k) := \{\mathbf{z} \in \mathcal{H} : \|\mathbf{z} - \mathbf{c}_k\|_2 \le r_k\}$, or equivalently any $\mathbf{z} \in H_k^+ := \{\mathbf{z} : \langle \mathbf{w}_k, \mathbf{z} \rangle \ge 1\}$ where $\mathbf{w}_k := \frac{\mathbf{c}_k}{1 - \frac{1}{2} r_k^2}$, is in category k.

Vapnik and Lerner (1963) considered two scenarios:

- transductive learning (distinction): each (test) example is known to belong to one and only one ball $\mathsf{B}_k := \mathsf{B}(\mathbf{c}_k, r_k)$. In this case we may identify the category for any $\mathbf{x} \in \mathcal{X}$:

  $$\mathbf{x} \in \mathcal{C}_k \iff \mathbf{z} := \varphi(\mathbf{x}) \in \mathsf{B}_k \setminus \mathsf{B}_{\neg k} := \bigcup_{l \neq k} \mathsf{B}_l, \quad \text{or simply} \quad \mathbf{z} \in \mathsf{B}_k.$$

  In other words, $\|\mathbf{z} - \mathbf{c}_k\|_2 \leq r_k$ and $\|\mathbf{z} - \mathbf{c}_l\|_2 > r_l$ for all $l \neq k$, or equivalently

  $$\forall l \neq k, \quad \langle \mathbf{z}, \mathbf{w}_k \rangle \geq 1 > \langle \mathbf{z}, \mathbf{w}_l \rangle, \quad \text{where} \quad \mathbf{w}_k := \frac{\mathbf{c}_k}{1 - \frac{1}{2} r_k^2}.$$

  Therefore, we may use the simple rule to predict the category $c(\mathbf{x})$ of $\mathbf{x}$ (under transformation φ):

  $$c(\mathbf{x}) = \operatorname*{argmax}_{k=1,\ldots,\mathsf{c}} \langle \varphi(\mathbf{x}), \mathbf{w}_k \rangle. \tag{7.7}$$

- inductive learning (recognition): each (test) example may be in several balls or may not be in any ball at all. We may still use the same prediction rule (7.7), but declare "failure"
  - if $\max_{k=1,\ldots,\mathsf{c}} \langle \varphi(\mathbf{x}), \mathbf{w}_k \rangle < 1$: $\mathbf{x}$ does not belong to any existing category;
  - on the other hand, if $|c(\mathbf{x})| > 1$: $\mathbf{x}$ belongs to at least two existing categories. Ambiguous.

Vapnik and Lerner (1963) ended with an announcement of the main result in Remark 7.11, namely how to find the centers $\mathbf{c}_k$ (or equivalently the weights $\mathbf{w}_k$) using (7.6).

Vapnik, Vladimir N. and A. Ya. Lerner (1963). "Pattern Recognition using Generalized Portraits". *Automation and Remote Control*, vol. 24, no. 6, pp. 709–715.

Remark 7.9: More on (Vapnik and Lerner 1963)

Vapnik and Lerner (1963) defined a dataset as indefinite, if there exists some test example $\mathbf{x}$ so that

$$\forall k, \ \mathbf{z} \in \mathsf{B}(\mathbf{c}_k, r_{\neg k}) \setminus \mathsf{B}(\mathbf{c}_k, r_k), \quad \text{where} \quad r_{\neg k} := \max_{l \neq k} r_l,$$

i.e., $\mathbf{z}$ is not in category k but would be if we increase its radius r_k to that of the best alternative $r_{\neg k}$.

Vapnik and Lerner (1963) proposed to use the (positive) quantity

$$I = I(\mathcal{D}) = 1 - \max_{k \neq l} \langle \mathbf{c}_k, \mathbf{c}_l \rangle$$

to measure the distinguishability of our dataset: the bigger I is, the more spread (orthogonal) the centers are hence the easier to distinguish the categories. Using I as the evaluation metric one can sequentially refine the feature transformation φ so that the resulting distinguishability is steadily increased.

Vapnik and Lerner (1963) also noted the product space trick: Let $\{\varphi_j : \mathcal{X} \to \mathcal{H}_j, j = 1, \ldots, m\}$ be a set of feature transformations. Then, w.l.o.g., we can assemble them into a single transformation $\varphi : \mathcal{X} \to \mathcal{H} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_m$.

Vapnik, Vladimir N. and A. Ya. Lerner (1963). "Pattern Recognition using Generalized Portraits". *Automation and Remote Control*, vol. 24, no. 6, pp. 709–715.

Remark 7.10: Linear separability, revisited

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

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

Let us now break the above condition for any positive example $y_i = 1$ and any negative example $y_j = -1$:

$$\mathbf{w}^\top \mathbf{x}_i + b \geq s \geq -s \geq \mathbf{w}^\top \mathbf{x}_j + b \iff \mathbf{w}^\top \mathbf{x}_i \geq s - b \geq -s - b \geq \mathbf{w}^\top \mathbf{x}_j$$

$$\iff \min_{i:y_i=1} \mathbf{w}^\top \mathbf{x}_i > \max_{j:y_j=-1} \mathbf{w}^\top \mathbf{x}_j.$$

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 7.11: History of SVM, continued

In this box we summarize the main result in (Vapnik and Chervonenkis 1964).

Inspired by (7.6), Vapnik and Chervonenkis (1964) essentially applied the one-vs-all reduction (cf. Remark 1.34) and arrived at what they called the *optimal approximation*:

$$\begin{aligned} \max_{\|\mathbf{w}\|=1} \quad & \min_{i:y_i=1} \mathbf{w}^\top \mathbf{x}_i \\ \text{s.t.} \quad & \min_{i:y_i=1} \mathbf{w}^\top \mathbf{x}_i \geq \max_{j:y_j=-1} \mathbf{w}^\top \mathbf{x}_j. \end{aligned} \tag{7.8}$$

According to Remark 7.10, the last condition is equivalent to requiring the dataset to be linearly separable (strictly or not). Reintroducing the offset b and applying the standardization trick in Remark 7.5:

$$\begin{aligned} \max_{\mathbf{w},b,\delta} \quad & \delta - b \\ \text{s.t.} \quad & \|\mathbf{w}\| = 1 \quad \xrightarrow{\text{assuming obj} > 0} \quad \|\mathbf{w}\| \leq 1 \\ & y_i \hat{y}_i \geq \delta \geq 0, \quad \hat{y}_i := \mathbf{w}^\top \mathbf{x}_i + b, \quad i = 1, \ldots, n. \end{aligned} \tag{7.9}$$

The above problem obviously admits a solution iff the dataset is linearly separable. Moreover, if we assume the maximum objective is strictly positive (which is different from strict linear separability: take say $\mathcal{D} = \{(-1,+), (-2,-)\}$), then we can relax the unit norm constraint, in which case the optimal $\mathbf{w}$ is unique.

This formulation (7.9) of Vapnik and Chervonenkis differs from our previous one (7.5) mainly in the objective function: $\delta - b$ vs. δ, i.e. Vapnik and Chervonenkis always subtract the offset from the minimum margin. This difference can be significant though, see Example 7.12 below for an illustration.

The main result in (Vapnik and Chervonenkis 1964), aside from the formulation in (7.8), is the derivation of its Lagrangian dual, which is quite a routine derivation nowadays. Nevertheless, we reproduce the original argument of Vapnik and Chervonenkis for historical interests.

Define $C(\mathbf{w}) = \min_{i:y_i=1} \mathbf{w}^\top \mathbf{x}_i$, and define the set of support vectors $S(\mathbf{w}) := \{y_i \mathbf{x}_i : \mathbf{w}^\top \mathbf{x}_i = C(\mathbf{w})\}$. By definition there is always a positive support vector while there may not be any negative support vector (consider say $\mathcal{D} = \{(1,-), (2,+)\}$). Recall that Vapnik and Chervonenkis assumed the optimal objective $C^\star = C(\mathbf{w}^\star)$ in (7.6) is strictly positive, hence the uniqueness of the optimal solution $\mathbf{w}^\star$. By restricting the norm $\|\cdot\| = \|\cdot\|_2$, Vapnik and Chervonenkis made the following observations:

- $\mathbf{w}^\star$ is a conic combination of support vectors $S = S(\mathbf{w}^\star)$. Suppose not, let P denote the ℓ_2 projector onto the conic hull of support vectors. Let $\mathbf{w}_\eta = (I - P)\mathbf{w}^\star + \eta P \mathbf{w}^\star$. For any support vector $y\mathbf{x} \in S$:

  $$y[\langle \mathbf{w}_\eta, \mathbf{x} \rangle - \eta \langle \mathbf{w}^\star, \mathbf{x} \rangle] = (1-\eta) \langle \mathbf{w}^\star - P\mathbf{w}^\star, y\mathbf{x} \rangle = (1-\eta) \langle \mathbf{w}^\star - P\mathbf{w}^\star, (y\mathbf{x} + P\mathbf{w}^\star) - P\mathbf{w}^\star \rangle \geq 0,$$

  since P is the projector onto the conic hull and $\eta \geq 1$. Since $C(\mathbf{w}^\star) > 0$ by assumption, slightly increase η from 1 to $1 + \epsilon$ will maintain all constraints but increase the objective in (7.6), contradiction to the optimality of $\mathbf{w}^\star$.

- If we normalize the weight vector $\mathbf{w}^* = \mathbf{w}^\star / C(\mathbf{w}^\star)$, the constraint in (7.6) is not affected but the objective $C(\mathbf{w}^*)$ now becomes unit. Of course, $\mathbf{w}^*$ remains to be a conic combination of support vectors:

  $$\mathbf{w}^* = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i, \quad \alpha_i \geq 0, \quad \alpha_i(\mathbf{x}_i^\top \mathbf{w}^* - 1) = 0, \quad y_i(\mathbf{x}_i^\top \mathbf{w}^* - 1) \geq 0, \quad i = 1, \ldots, n, \tag{7.10}$$

  where the conditions follow from our definition of support vectors and is known as the KKT condition.

- Vapnik and Chervonenkis mentioned the following differential system:

$$\frac{\mathrm{d}\boldsymbol{\alpha}}{\mathrm{d}t} = -\epsilon\boldsymbol{\alpha} + [\mathbf{y} - (K \odot \mathbf{y}\mathbf{y}^\top)\boldsymbol{\alpha}]^+, \quad K_{ij} := \langle \mathbf{x}_i, \mathbf{x}_j \rangle,$$

  whose equilibrium will approach the KKT condition hence the solution in (7.10) as $\epsilon \to 0$. Vapnik and Chervonenkis concluded that to compute $\boldsymbol{\alpha}$, we need only the dot product K. Moreover, to reconstruct $\mathbf{w}$, only the support vectors from the training set are needed.

- To perform testing, we compare $\mathbf{x}^\top\mathbf{w}^*$ with threshold 1 (cf. the last condition in (7.10)). Or equivalently, using uniqueness we can recover $\mathbf{w}^\star = \mathbf{w}^*/\|\mathbf{w}^*\|_2$, where

$$\|\mathbf{w}^*\|_2^2 = \langle \mathbf{w}^*, \mathbf{w}^* \rangle = \left\langle \mathbf{w}^*, \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right\rangle = \sum_{i=1}^n \alpha_i y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle = \boldsymbol{\alpha}^\top \mathbf{y}.$$

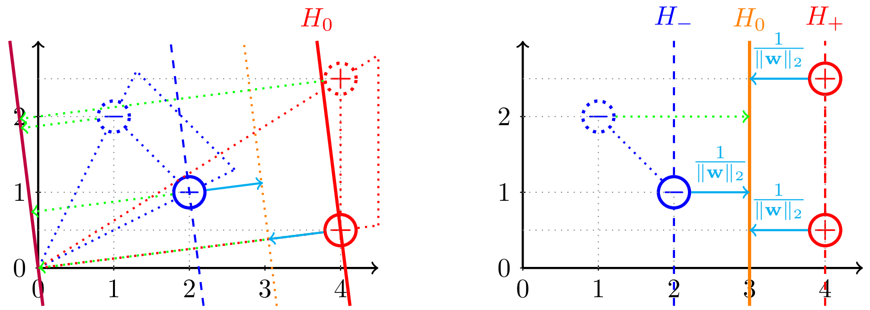
  Thus, we may also compare $\mathbf{x}^\top\mathbf{w}^\star$ with the threshold $\frac{1}{\sqrt{\boldsymbol{\alpha}^\top\mathbf{y}}}$. Usually we prefer to use $\mathbf{w}^*$ because its threshold is normalized: in case there are multiple classes, we can then use the argmax rule: $\hat{y} = \mathrm{argmax}_k\, \mathbf{w}_k^\top\mathbf{x}$. However, note that this decision boundary corresponds to the hyperplane that first passes through a positive example!

- Vapnik and Chervonenkis mentioned the possibility to dynamically update the dual variables $\boldsymbol{\alpha}$: upon making a mistake on an example $(\mathbf{x}, y)$, we augment the support vectors S with $(\mathbf{x}, y)$ and recompute the dual variables on the "effective training set" S. Again, only dot products are needed.

Vapnik, Vladimir N. and A. Ya. Chervonenkis (1964). "On a class of perceptrons". *Automation and Remote Control*, vol. 25, no. 1, pp. 112–120.

Example 7.12: SVM, old and new

As illustrated in the following figure, the optimal approximation in Remark 7.11 can behave very differently from the familiar SVM formulation (7.11) in Remark 7.13. To verify that the left purple solid line in the left plot is indeed optimal, simply note that the minimum distance from positive examples to any hyperplane passing through the origin (e.g. the objective in (7.8)) is at most the minimum distance from positive examples to the origin, which the left purple line achieves.



Remark 7.13: Removing homogeneity by normalizing offset

A different way to remove the scaling-invariance mentioned in Definition 7.2 is to perform normalization on the offset so that

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

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 (7.2) as:

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

We remind again that δ here is any fixed positive constant and we are *not* optimizing it (in contrast to what we did in Remark 7.6). Applying the trick in Exercise 3.21 we arrive at the usual formulation of SVM (due to Boser et al. (1992)):

$$\begin{aligned} \min_{\mathbf{w}} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i \hat{\mathsf{y}}_i \geq \delta, \ \forall i = 1, \ldots, n. \end{aligned} \tag{7.11}$$

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

The formulation (7.11) only makes sense on (strictly) linearly separable datasets, unlike our original formulation (7.3).

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik (1992). "A Training Algorithm for Optimal Margin Classifiers". In: *COLT*, pp. 144–152.

Alert 7.14: Any positive number but not zero

Note that in the familiar SVM formulation (7.11), we can choose δ 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 7.15: Perceptron vs. SVM

We can formulate perceptron as the following feasibility problem:

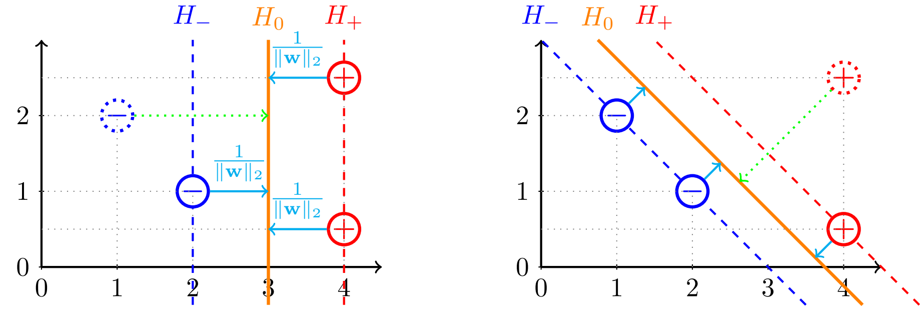
$$\begin{aligned} \min_{\mathbf{w}} \quad & 0 \\ \text{s.t.} \quad & y_i \hat{\mathsf{y}}_i \geq \delta, \ \forall i = 1, \ldots, n, \end{aligned}$$

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 7.16: 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.


H_-
H_0
H_+
$\frac{1}{\|\mathbf{w}\|_2}$
$\frac{1}{\|\mathbf{w}\|_2}$
$\frac{1}{\|\mathbf{w}\|_2}$
0
1
2
0
1
2
3
4
H_-
H_0
H_+
0
1
2
0
1
2
3
4

To understand the above figure, let us take a closer look at the SVM formulation (7.11), 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 (7.11):

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

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

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

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

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

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. $y = \mathrm{sign}(\mathbf{w}^\top \mathbf{x} + b)$. 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_+:

$$\begin{aligned} \mathrm{dist}(H_+, H_0) &:= \min_{\mathbf{p} \in H_+} \min_{\mathbf{q} \in H_0} \|\mathbf{p} - \mathbf{q}\|_\circ \\ &= \min_{i:y_i=1} \mathrm{dist}(\mathbf{x}_i, H_0) && \text{(since } H_+ \text{ first passes through positive examples)} \\ &= \frac{1}{\|\mathbf{w}\|} && \text{(cf. (7.1))} \\ &= \min_{i:y_i=-1} \mathrm{dist}(\mathbf{x}_i, H_0) && \text{(since } H_- \text{ first passes through negative examples)} \\ &= \mathrm{dist}(H_-, H_0). \end{aligned}$$

Exercise 7.17: Uniqueness of $\mathbf{w}$

For the ℓ_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 7.18: 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 7.19: 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 $H_0 := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^\top\mathbf{x} + b = 0\}$ are convex.
- Halfspaces $H_\leq := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^\top\mathbf{x} + b \leq 0\}$ are convex.

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

Definition 7.20: Convex hull

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

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

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

Exercise 7.21: 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 $A \subseteq \mathbb{R}^d$:

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

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

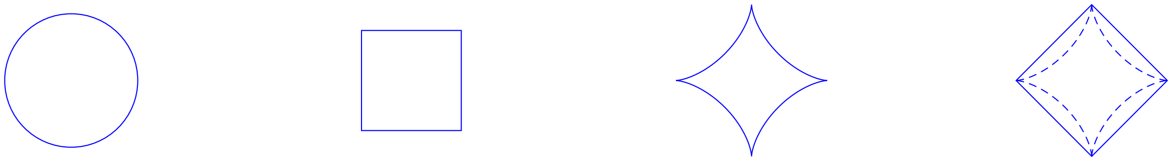
(The celebrated Carathéodory theorem allows us to restrict $n \leq d+1$, and $n \leq d$ if A is connected.)

Exercise 7.22: Unit balls of norms are convex

Recall that the unit ball of the ℓ_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 B_p for $p = 2, \infty, \frac{1}{2}, 1$.



As shown above:

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

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

Remark 7.23: 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 (cf. Definition 1.22) iff

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

(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.)

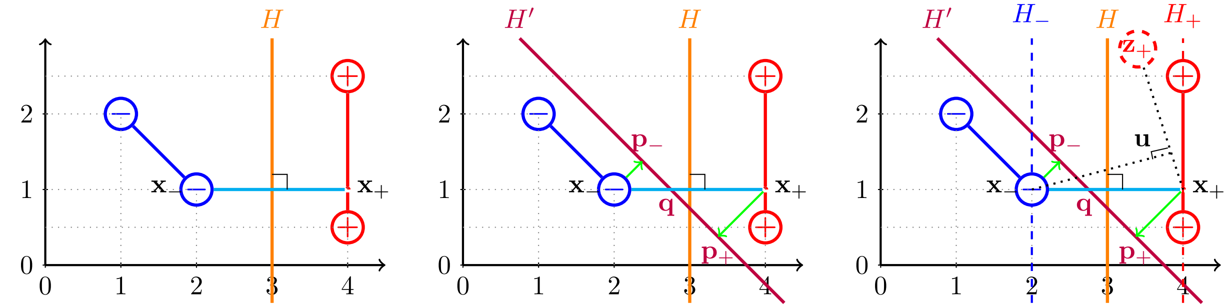
To test if a given dataset $\mathcal{D}$ is (strictly) linearly separable, Rosen's idea was to compute the minimum (Euclidean) distance between the (convex hulls of the) two classes. In his Eq (2.5), after applying the standardization trick, cf. Remark 7.5, Rosen proposed exactly (up to a constant $\frac{1}{2}$) the hard-margin SVM formulation (7.4). Then, to get an equivalent *convex* formulation, Rosen (1965) did some simple algebraic manipulations to arrive at the familiar hard-margin SVM in (7.11) (his Eq (2.6), again, up to a constant $\frac{1}{2}$). Rosen (1965) proved the uniqueness of the hard-margin SVM solution, and he further proved that the number of support vectors can be bounded by $d + 1$. Rosen (1965) also discussed how to separate more than two classes, using basically the one-vs-one and the one-vs-all reductions (cf. Remark 1.34). It seems appropriate to attribute our hard-margin SVM formulations in (7.4) and (7.11) to Rosen (1965).

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

Remark 7.24: More on linear separability detection (Mangasarian 1965)

Mangasarian, O. L. (1965). "Linear and Nonlinear Separation of Patterns by Linear Programming". *Operations Research*, vol. 13, no. 3, pp. 444–452.

Remark 7.25: Dual view of SVM, as bisector of minimum distance pair



In Definition 7.2 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:

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

  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:

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

  where the first equality follows from (7.1), 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 $\mathbf{q} \in H'$,

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

  Therefore, using our second and third observations above:

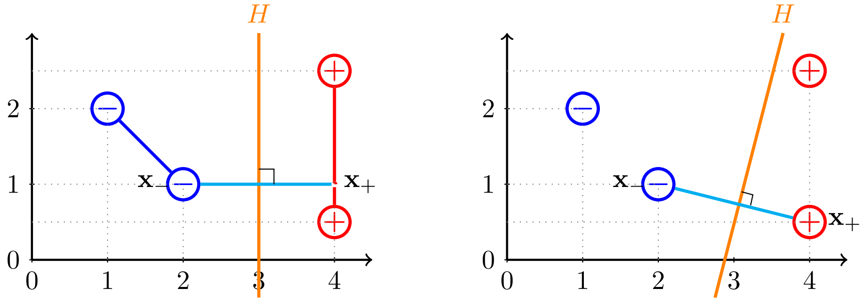
$$\min_{\mathbf{x}\in\mathcal{D}^\pm} \operatorname{dist}(\mathbf{x}, H') = \min_{\mathbf{x}\in\operatorname{conv}(\mathcal{D}^\pm)} \operatorname{dist}(\mathbf{x}, H') \leq \operatorname{dist}(\mathbf{x}_+, \mathbf{p}_+) \wedge \operatorname{dist}(\mathbf{x}_-, \mathbf{p}_-)$$

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

Exercise 7.26: Necessity of convex hull

In Remark 7.25, 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 7.25 might fail if we insist in picking both $\mathbf{x}_+$ and $\mathbf{x}_-$ from the original datasets $\mathcal{D}^\pm$?



Remark 7.27: SVM dual, from geometry to algebra

We complement the geometric dual view of SVM in Remark 7.25 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:

$$
\begin{aligned}
\max_{\|\mathbf{w}\|=1,b} \operatorname{dist}(\mathcal{D}^+, H_{\mathbf{w}}) \wedge \operatorname{dist}(\mathcal{D}^-, H_{\mathbf{w}}) &= \max_{\|\mathbf{w}\|=1,b} \left[\min_{\mathbf{x}_+\in\mathcal{D}^+} (\mathbf{w}^\top\mathbf{x}_+ + b) \wedge \min_{\mathbf{x}_-\in\mathcal{D}^-} -(\mathbf{w}^\top\mathbf{x}_- + b)\right] \\
&= \max_{\|\mathbf{w}\|=1,b} \left[\min_{\mathbf{x}_\pm\in\mathcal{D}^\pm, t\in[0,1]} t(\mathbf{w}^\top\mathbf{x}_+ + b) + (1-t)(-\mathbf{w}^\top\mathbf{x}_- - b)\right] \\
&= \max_{\|\mathbf{w}\|\le1,b} \left[\min_{\mathbf{x}_+\in t\operatorname{conv}(\mathcal{D}^+), \mathbf{x}_-\in(1-t)\operatorname{conv}(\mathcal{D}_-), t\in[0,1]} \mathbf{w}^\top(\mathbf{x}_+ - \mathbf{x}_-) + b(2t-1)\right] \\
&= \min_{\mathbf{x}_+\in t\operatorname{conv}(\mathcal{D}^+), \mathbf{x}_-\in(1-t)\operatorname{conv}(\mathcal{D}_-), t\in[0,1]} \max_{\|\mathbf{w}\|\le1,b} \left[\mathbf{w}^\top(\mathbf{x}_+ - \mathbf{x}_-) + b(2t-1)\right] \\
&= \min_{\mathbf{x}_+\in\frac{1}{2}\operatorname{conv}(\mathcal{D}^+), \mathbf{x}_-\in\frac{1}{2}\operatorname{conv}(\mathcal{D}_-)} \max_{\|\mathbf{w}\|\le1} \mathbf{w}^\top(\mathbf{x}_+ - \mathbf{x}_-) \\
&= \min_{\mathbf{x}_+\in\frac{1}{2}\operatorname{conv}(\mathcal{D}^+), \mathbf{x}_-\in\frac{1}{2}\operatorname{conv}(\mathcal{D}_-)} \|\mathbf{x}_+ - \mathbf{x}_-\|_\circ \\
&= \tfrac{1}{2}\operatorname{dist}(\operatorname{conv}(\mathcal{D}^+), \operatorname{conv}(\mathcal{D}^-)),
\end{aligned}
$$

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.