

Support Vector Machine



Two different approaches to regression/classification

$$x \sim P(x), \quad y = x^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Generative

- Assume something about $P(x,y)$ $p(y/x)$
- Find f which maximizes likelihood of training data | assumption
 - Often reformulated as minimizing loss

Versus

discriminative approach

- Pick a loss function $l: l_2, 0/1, \dots$
- Pick a set of hypotheses H *all linear classifiers*
- Pick f from H which minimizes loss on training data

$f(x) = w^T x$
 if $f(x) > 0$ predict 1
 if $f(x) < 0$ predict 0

$$\underset{f \in H}{\text{arg min}} \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)$$

Our description of logistic regression was the former

- **Learn:** $f: X \rightarrow Y$

- X – features

- Y – target classes

$$Y \in \{-1, 1\}$$

- **Expected loss of f :**

$$\mathbb{E}_{XY}[\mathbf{1}\{f(X) \neq Y\}] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x]]$$

$$\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x] = 1 - P(Y = f(x)|X = x)$$

- **Bayes optimal classifier:**

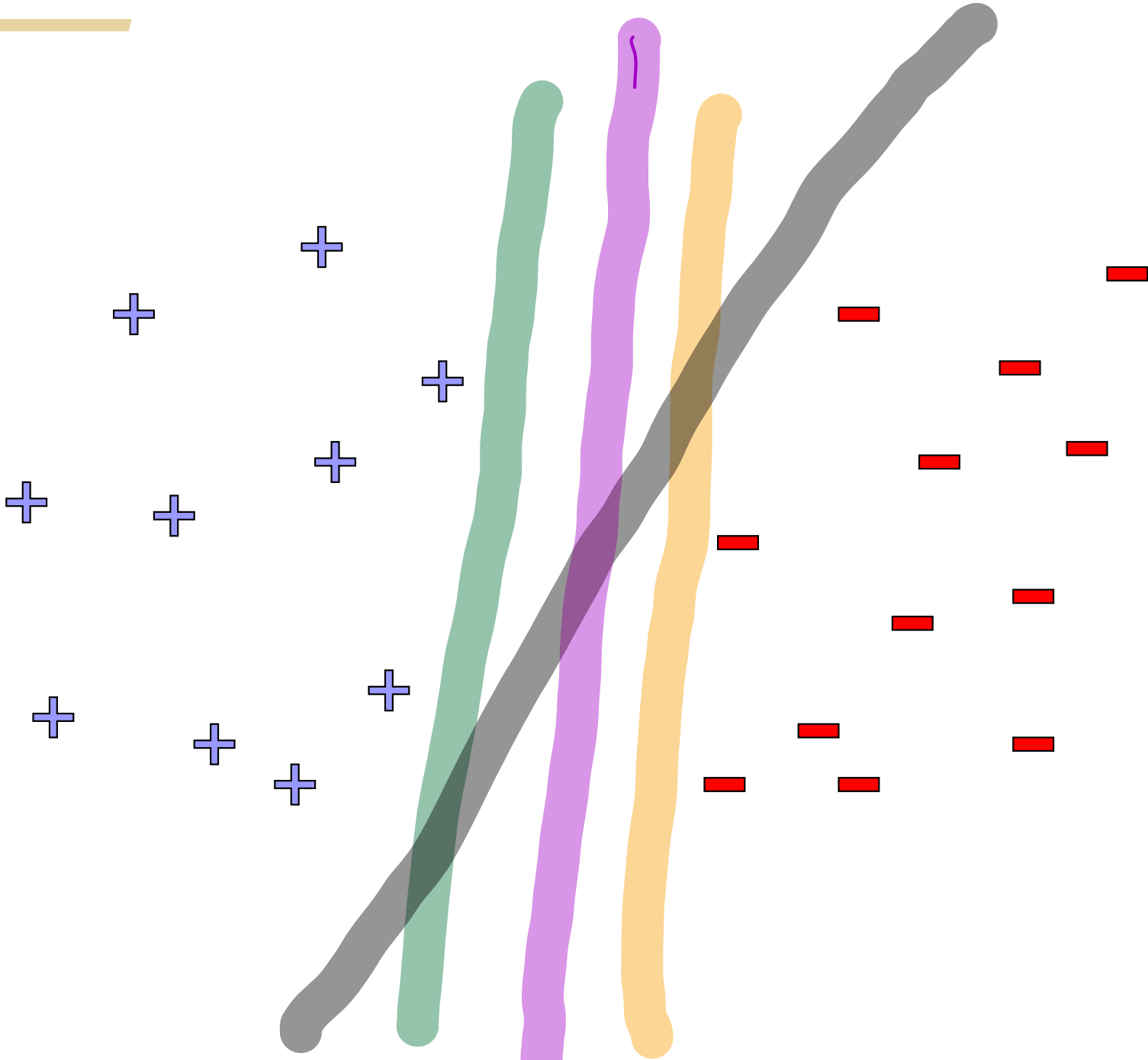
$$f(x) = \arg \max_y \mathbb{P}(Y = y|X = x)$$

- **Model of logistic regression:**

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y w^T x)}$$

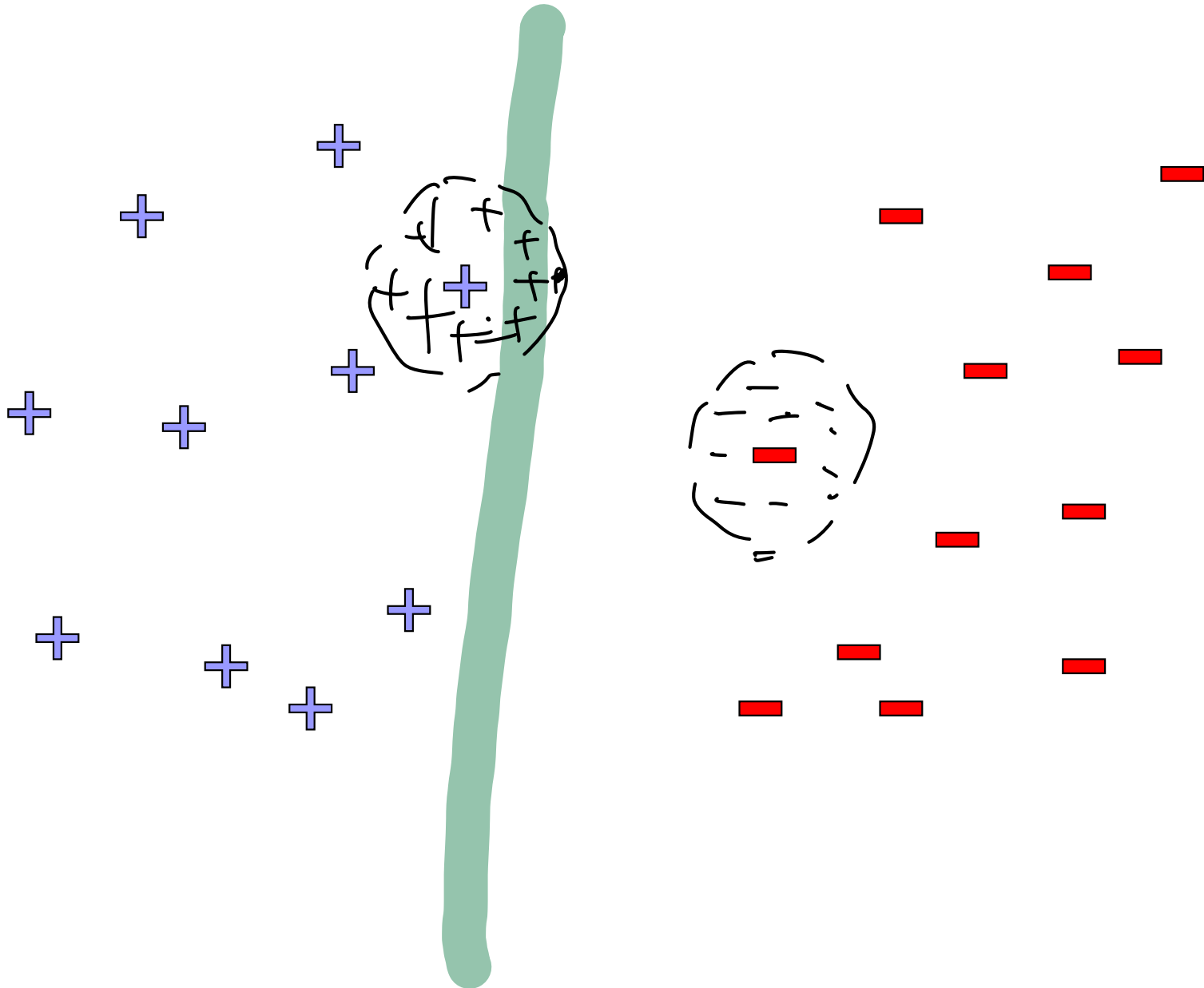
What if the model is wrong? What other ways can we pick linear decision rules?

Linear classifiers – Which line is better?

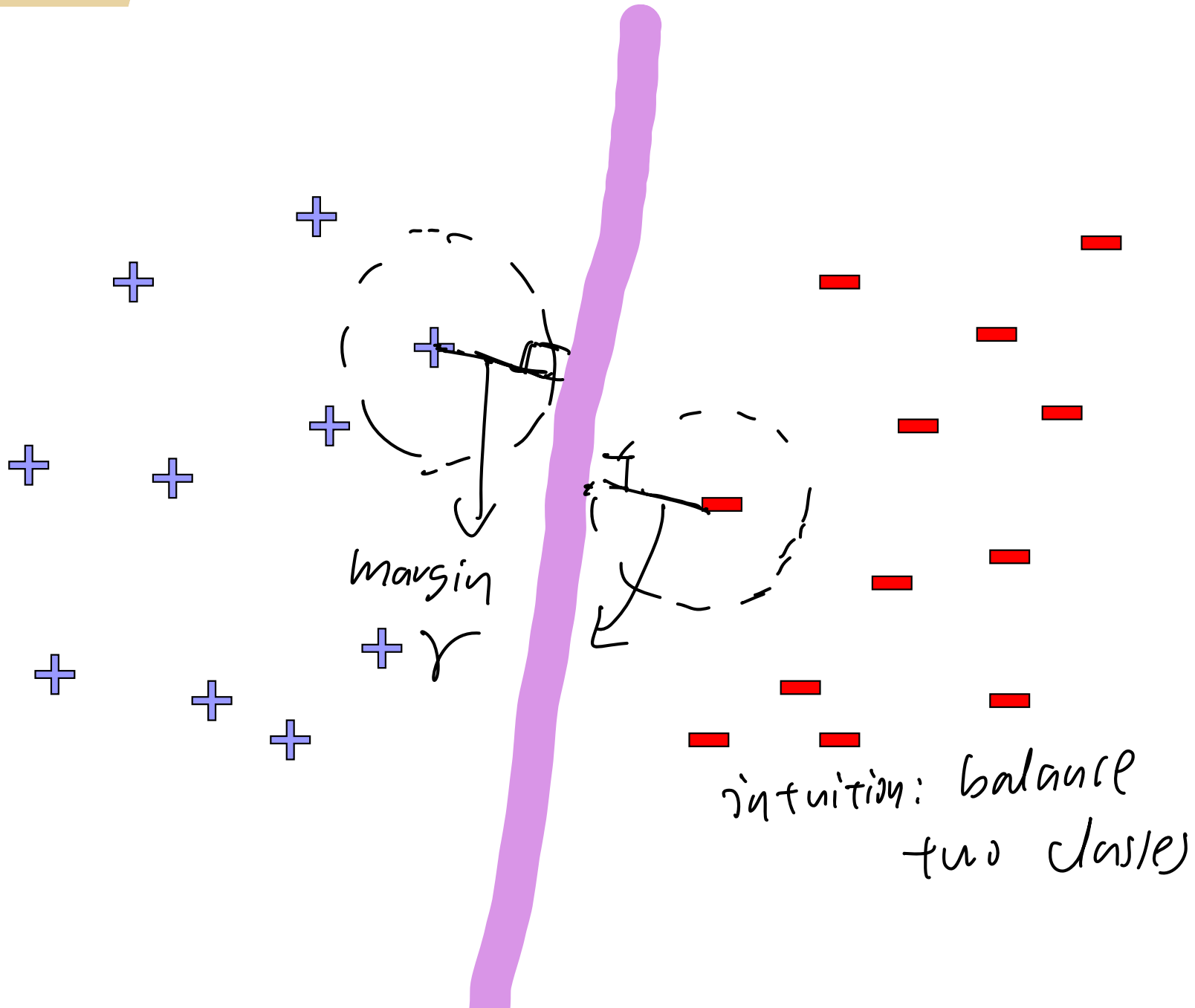


Linear classifiers – Which line is better?

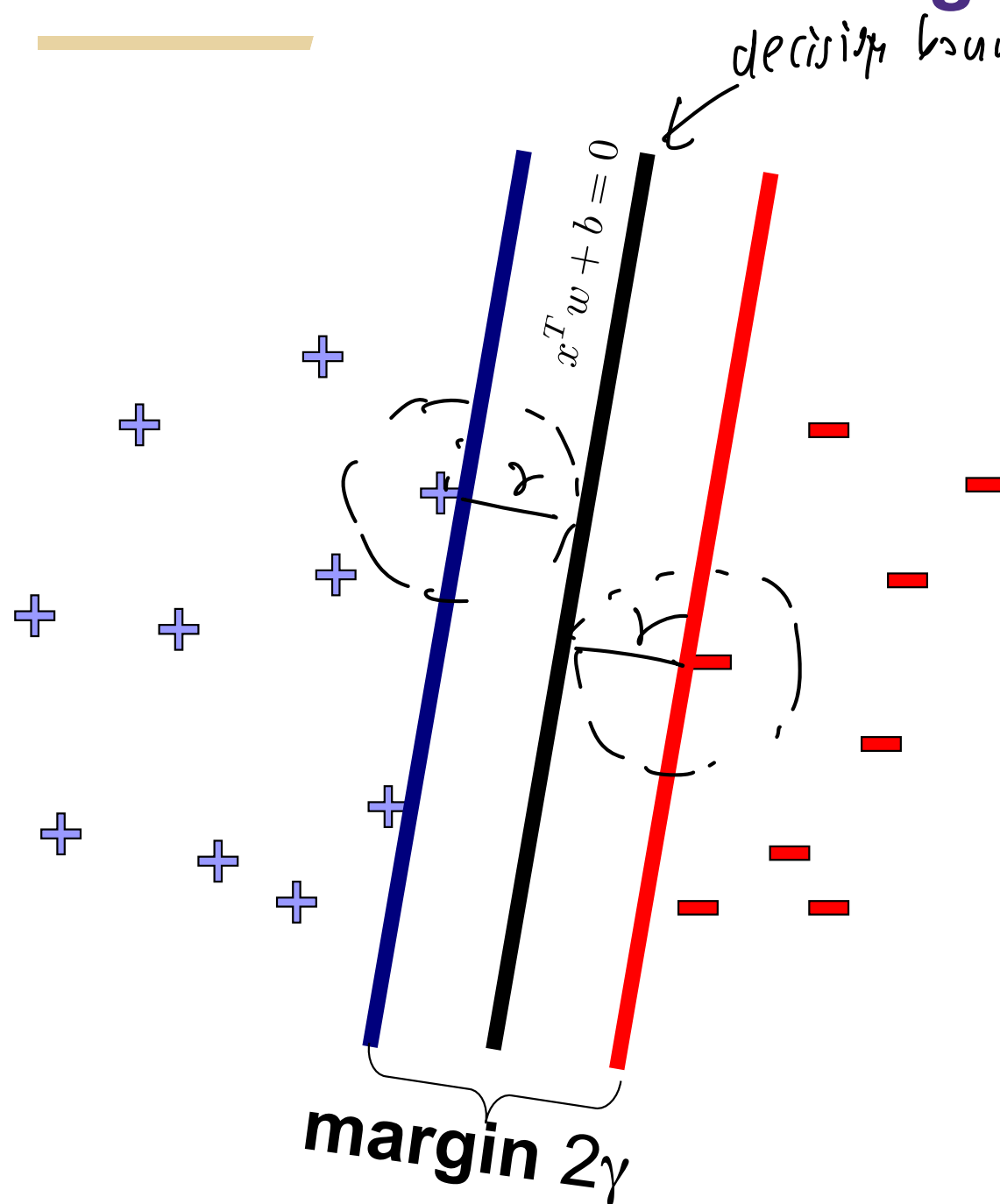
robustness to perturbation



Linear classifiers – Which line is better?



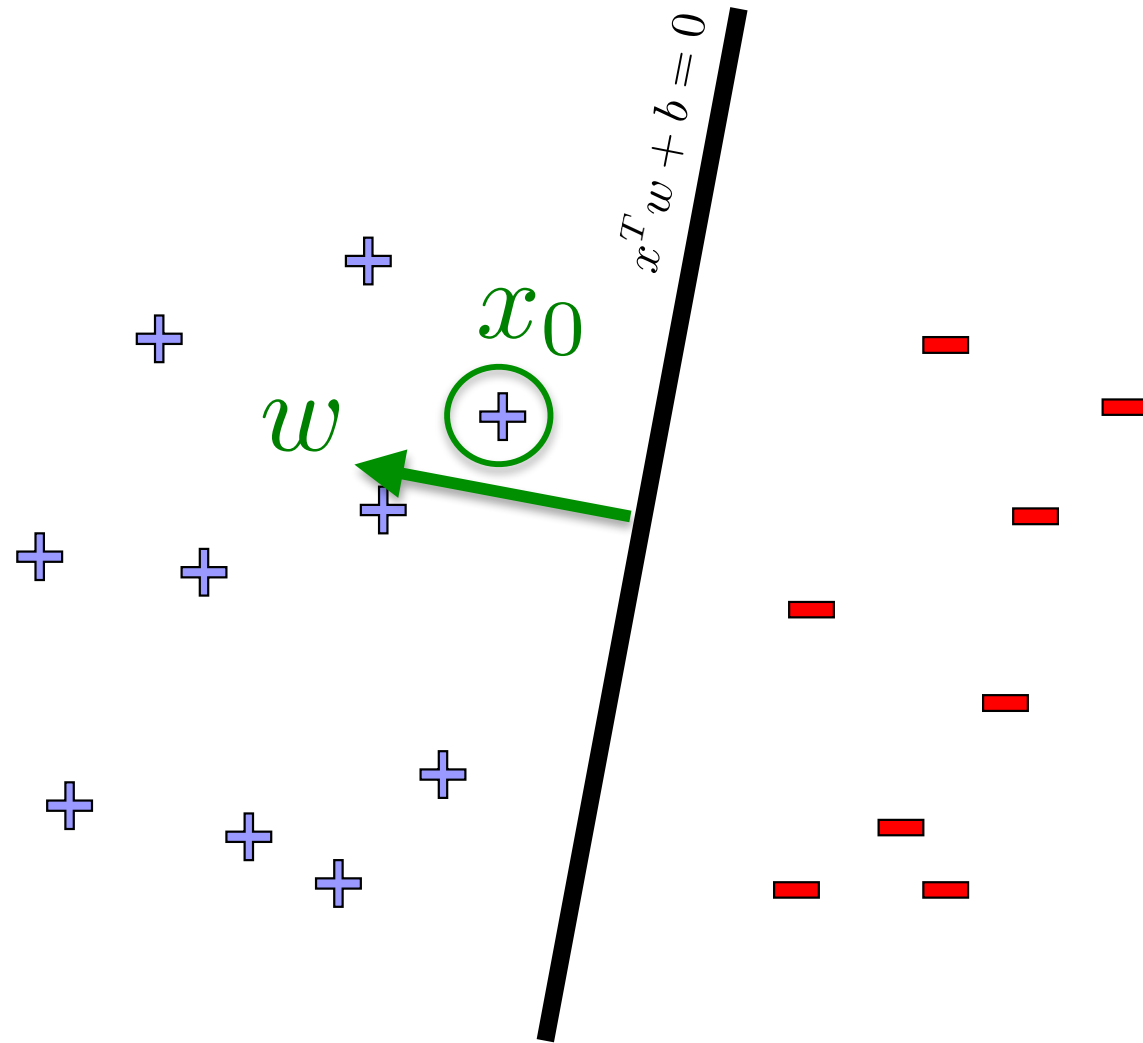
Pick the one with the largest margin!



γ : margin
the minimum distance
to decision boundary
of training data

Goal: find (w, b)
maximize γ

Pick the one with the largest margin!



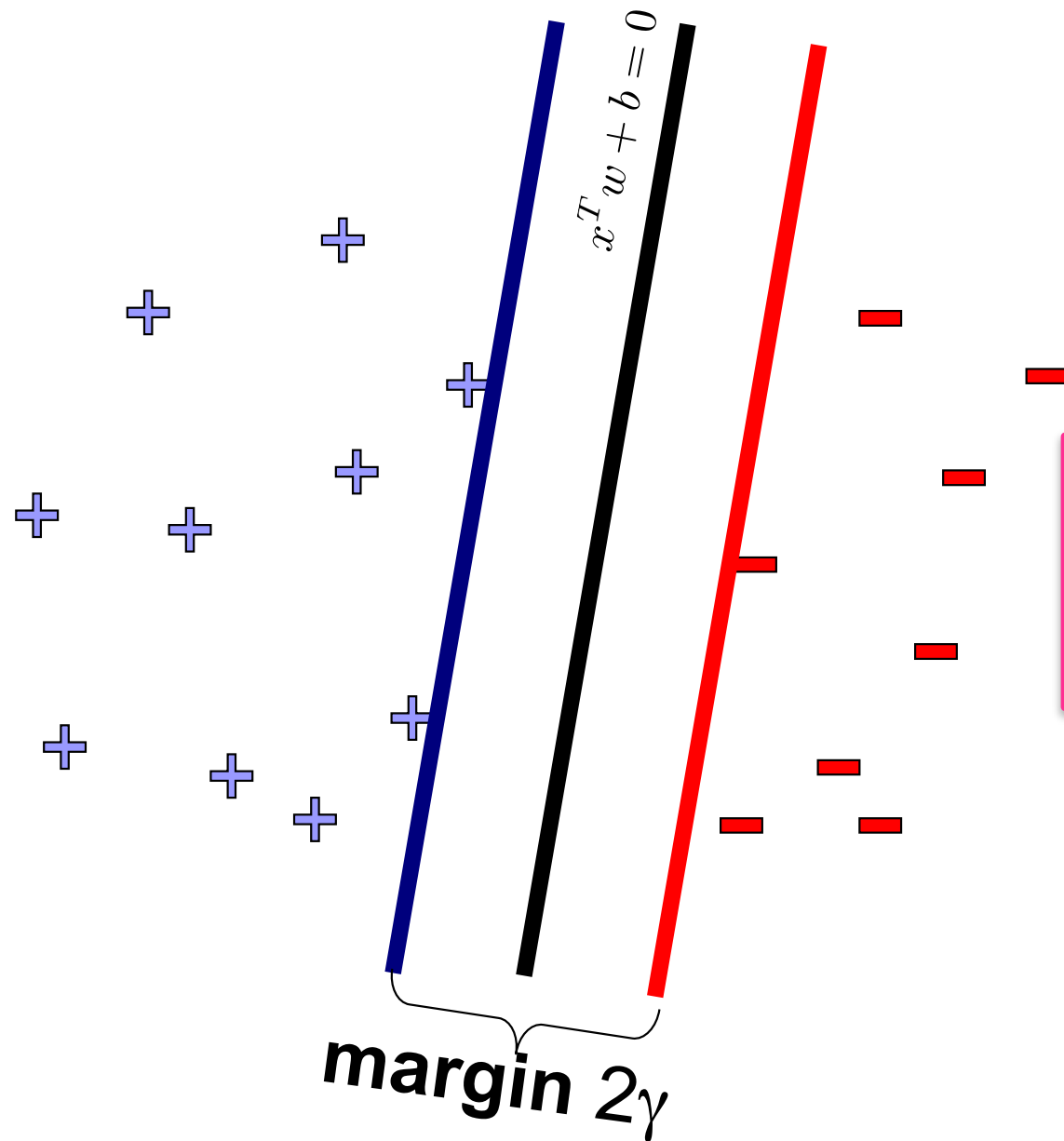
Distance from x_0 to hyperplane defined by $x^T w + b = 0$?

If \tilde{x}_0 is the projection of x_0 onto the hyperplane then
 $\|x_0 - \tilde{x}_0\|_2 = |(x_0^T - \tilde{x}_0^T) \frac{w}{\|w\|_2}|$

$$= \frac{1}{\|w\|_2} |x_0^T w - \tilde{x}_0^T w|$$

$$= \frac{1}{\|w\|_2} |x_0^T w + b|$$

Pick the one with the largest margin!



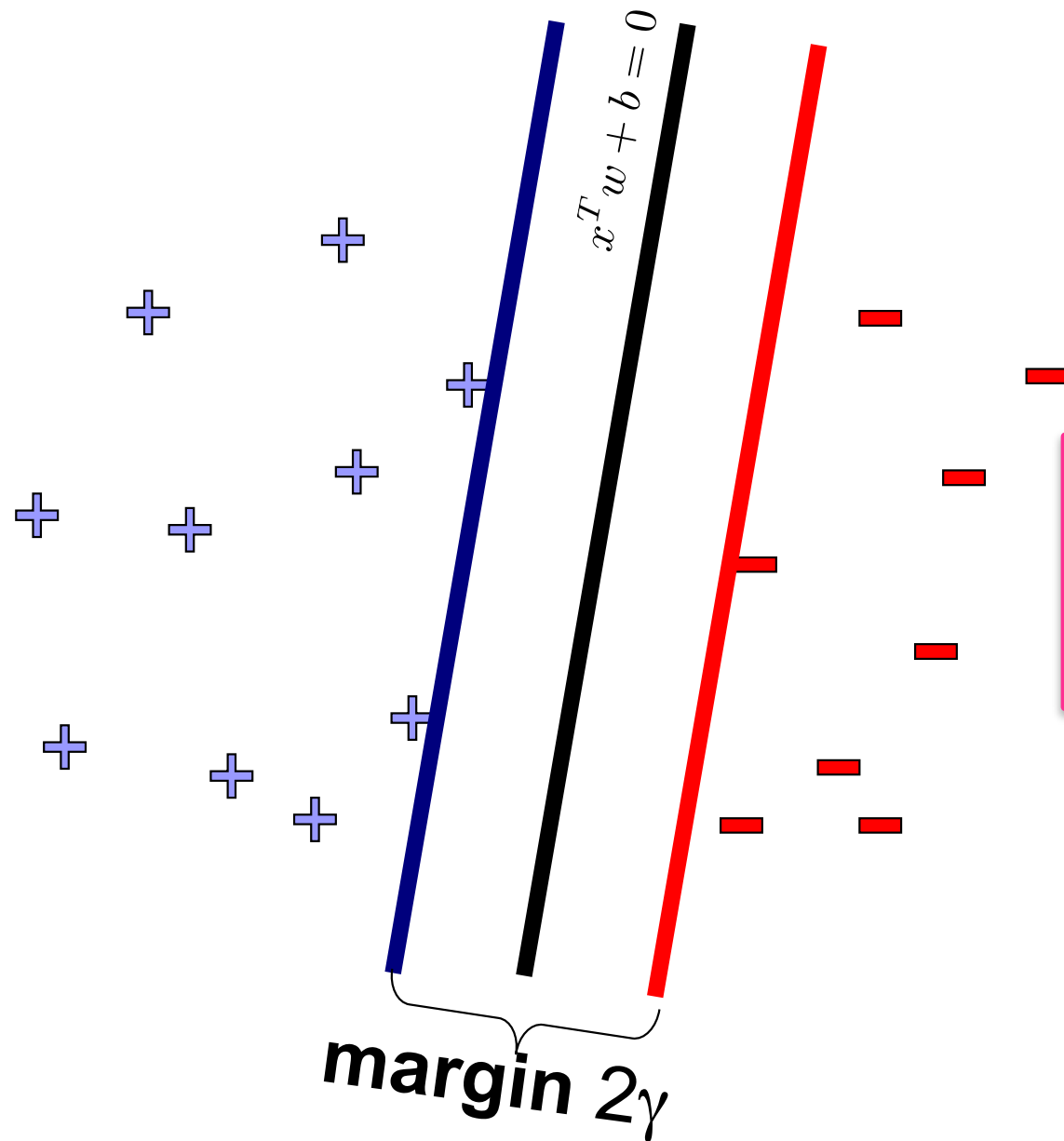
Distance of x_0 from
hyperplane $x^T w + b$:

$$\frac{1}{\|w\|_2} (x_0^T w + b)$$

Optimal Hyperplane



Pick the one with the largest margin!



Distance of x_0 from
hyperplane $x^T w + b$:

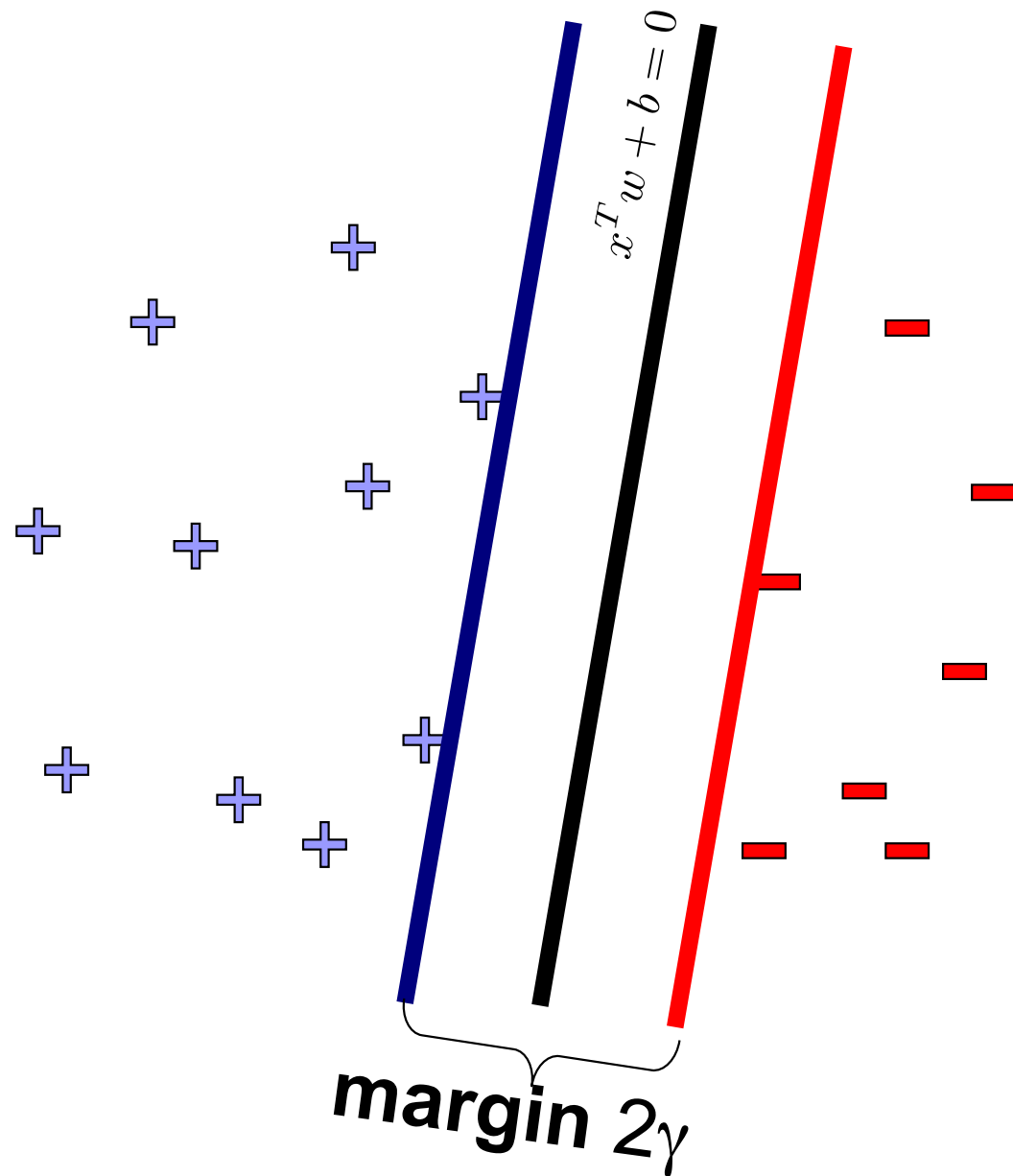
$$\frac{1}{\|w\|_2} (x_0^T w + b)$$

Optimal Hyperplane

$$\max_{w,b} \gamma$$

$$\text{subject to } \frac{1}{\|w\|_2} y_i (x_i^T w + b) \geq \gamma \quad \forall i$$

Pick the one with the largest margin!



Distance of x_0 from hyperplane $x^T w + b$:

$$\frac{1}{\|w\|_2} (x_0^T w + b)$$

Optimal Hyperplane

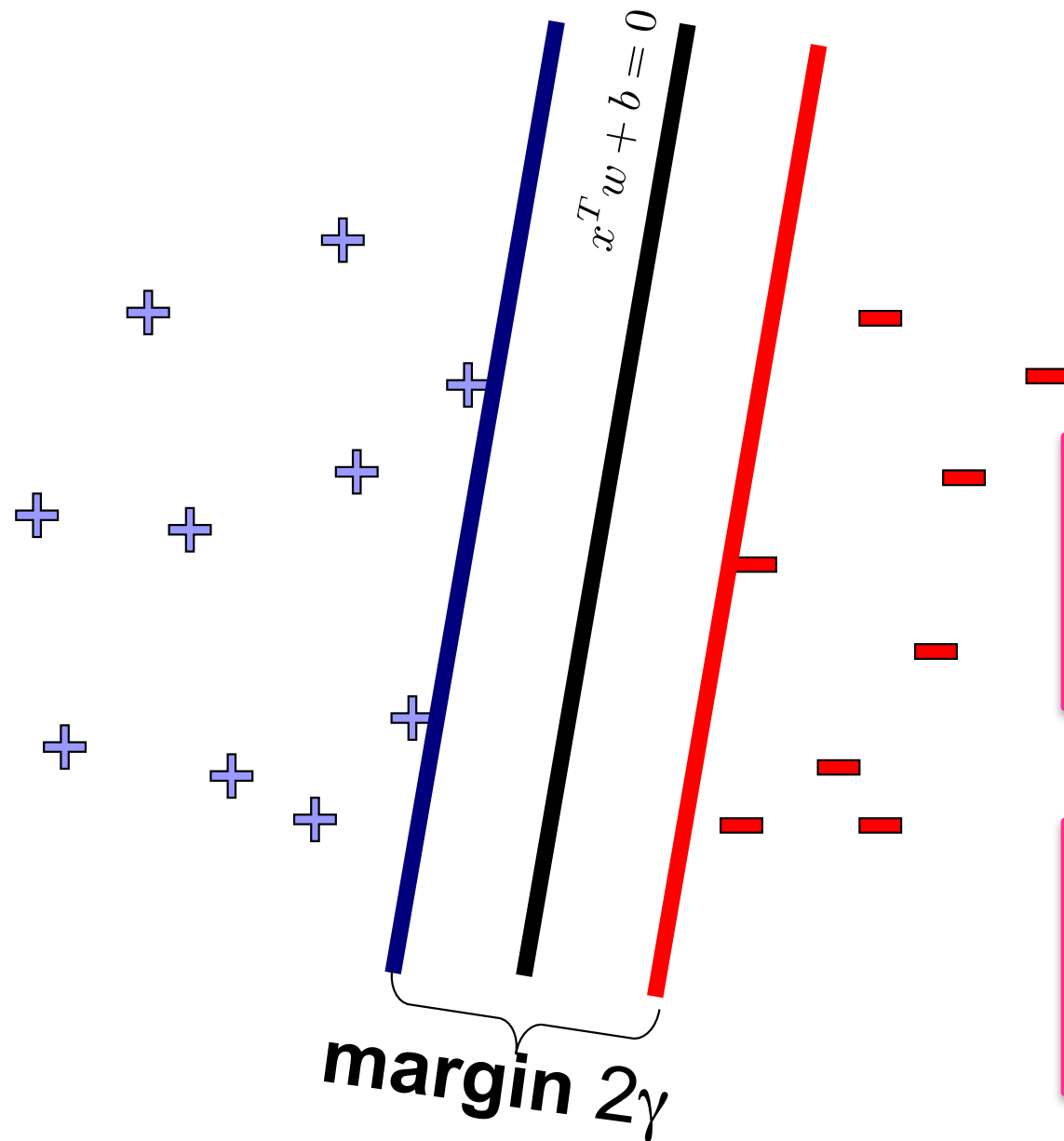
$$\max_{w,b} \gamma$$

$$\text{subject to } \frac{1}{\|w\|_2} y_i (x_i^T w + b) \geq \gamma \quad \forall i$$

Optimal Hyperplane (reparameterized)



Pick the one with the largest margin!



Distance of x_0 from
hyperplane $x^T w + b$:

$$\frac{1}{\|w\|_2} (x_0^T w + b)$$

Optimal Hyperplane

$$\max_{w,b} \gamma$$

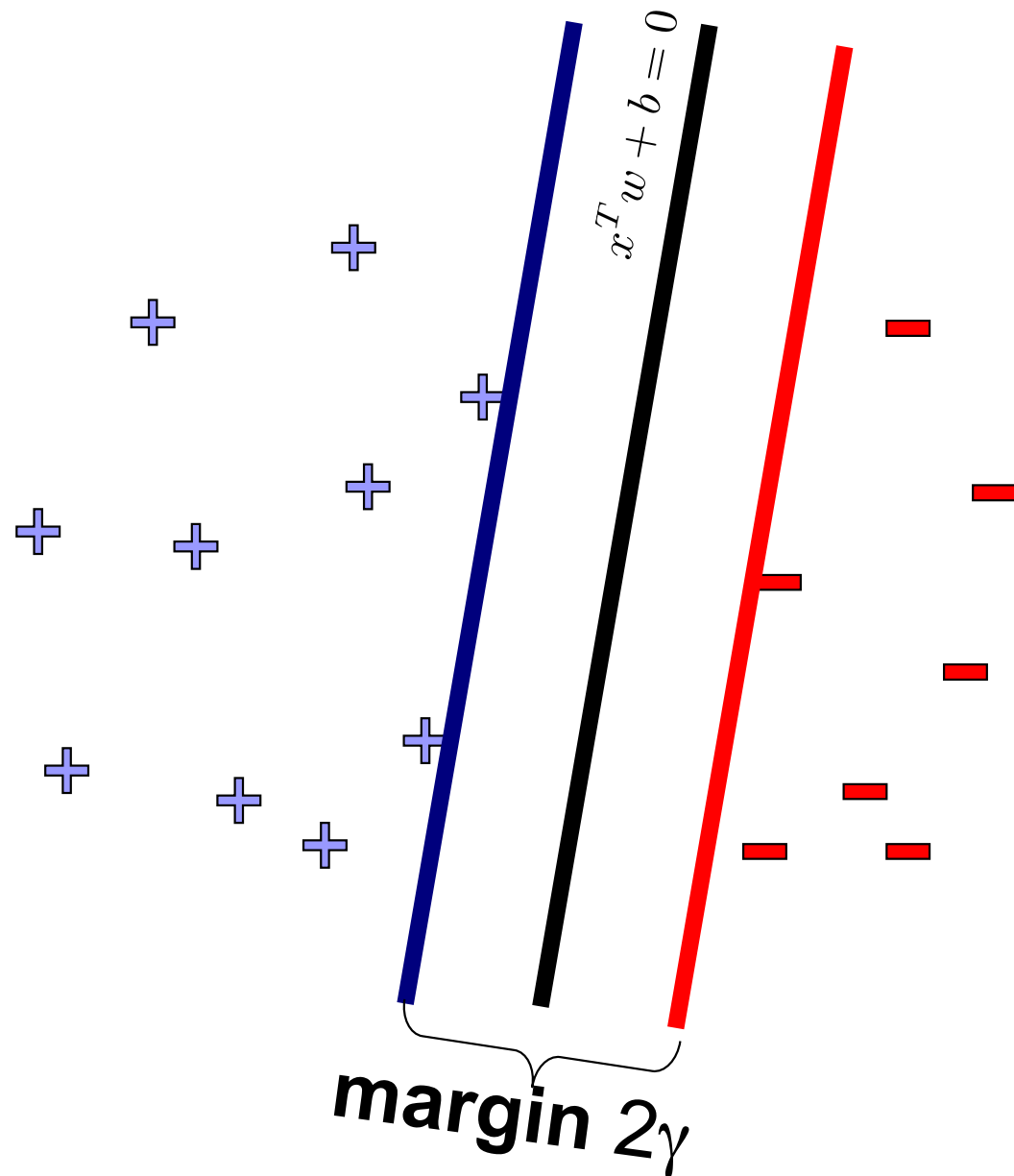
$$\text{subject to } \frac{1}{\|w\|_2} y_i (x_i^T w + b) \geq \gamma \quad \forall i$$

Optimal Hyperplane (reparameterized)

$$\min_{w,b} \|w\|_2^2$$

$$\text{subject to } y_i (x_i^T w + b) \geq 1 \quad \forall i$$

Pick the one with the largest margin!



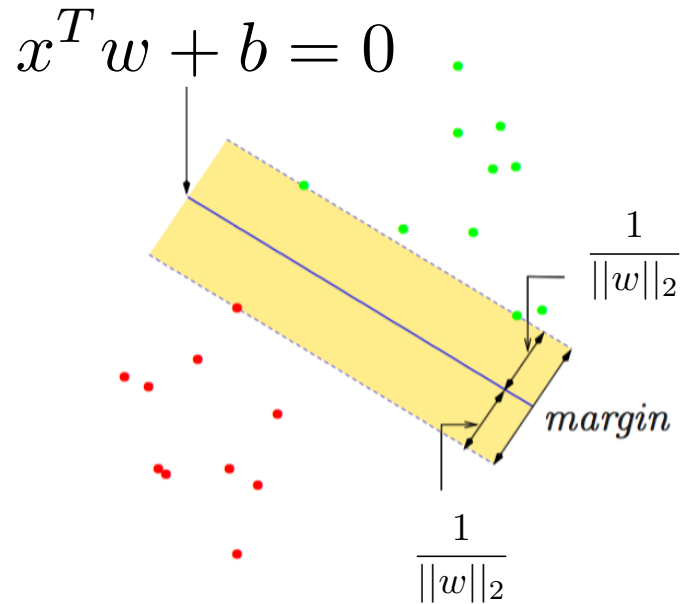
- Solve efficiently by many methods, e.g.,
 - quadratic programming (QP)
 - Well-studied solution algorithms
 - Stochastic gradient descent
 - Coordinate descent (in the dual)

Optimal Hyperplane (reparameterized)

$$\min_{w,b} \|w\|_2^2$$

$$\text{subject to } y_i(x_i^T w + b) \geq 1 \quad \forall i$$

What are support vectors



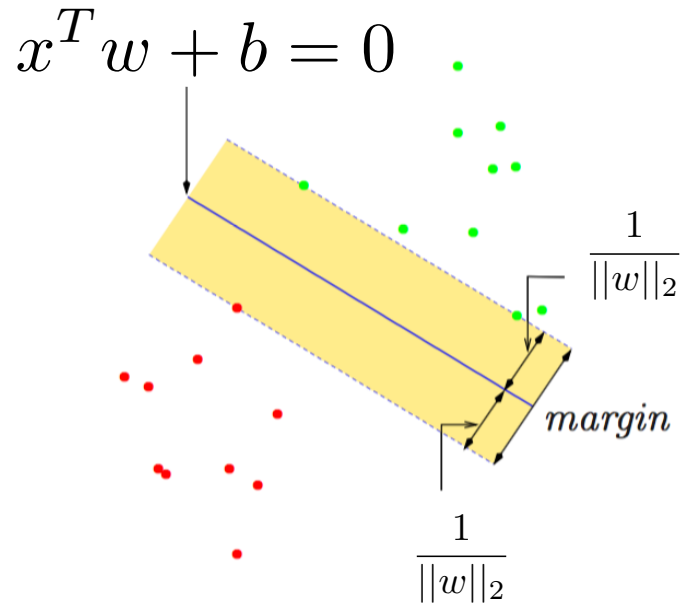
If data is linearly separable

$$\min_{w,b} \|w\|_2^2$$

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

Note: the solution of this can be written in terms of very few of the training points. These points are known as support vectors.

What if the data is not linearly separable?



If data is linearly separable

$$\min_{w,b} \|w\|_2^2$$

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

If data is not linearly separable, some points don't satisfy margin constraint:

Two options:

1. Introduce slack to this optimization problem
2. Lift to higher dimensional space

What if the data is not linearly separable?

If data is linearly separable:

$$\min_{w,b} \|w\|_2^2$$

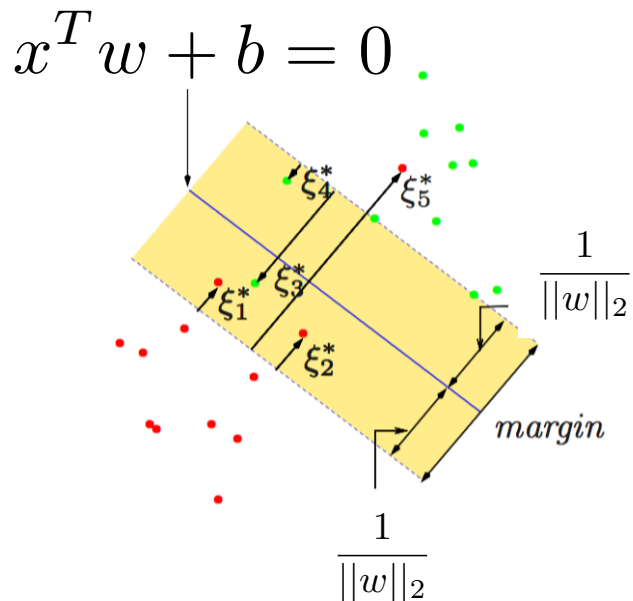
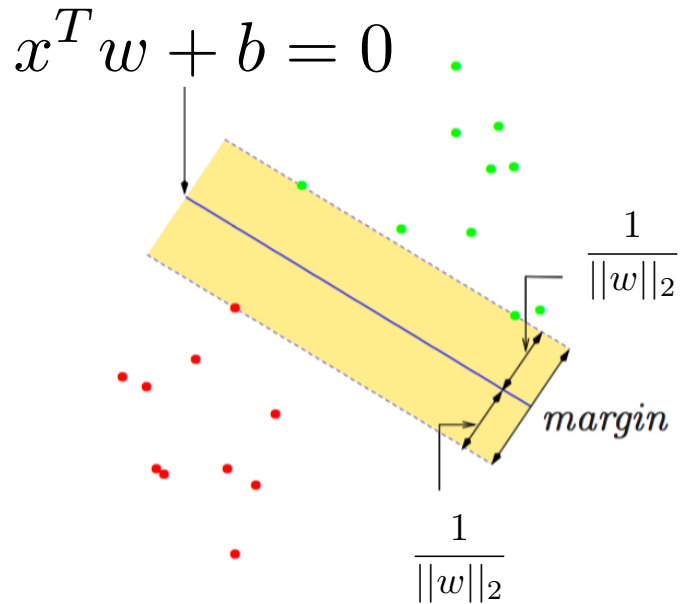
$$y_i(x_i^T w + b) \geq 1 \quad \forall i$$

If data is not linearly separable,
some points don't satisfy margin constraint:

$$\min_{w,b} \|w\|_2^2$$

$$y_i(x_i^T w + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0, \sum_{j=1}^n \xi_j \leq \nu$$



SVM as penalization method

- Original quadratic program with linear constraints:

$$\min_{w,b} \|w\|_2^2$$

$$y_i(x_i^T w + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0, \sum_{j=1}^n \xi_j \leq \nu$$

SVM as penalization method

- Original quadratic program with linear constraints:

$$\min_{w,b} \|w\|_2^2$$

$$y_i(x_i^T w + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0, \sum_{j=1}^n \xi_j \leq \nu$$

- Using same constrained convex optimization trick as for lasso:
For any $\nu \geq 0$ there exists a $\lambda \geq 0$ such that the solution the following solution is equivalent:

$$\sum_{i=1}^n \max\{0, 1 - y_i(b + x_i^T w)\} + \lambda \|w\|_2^2$$

SVMs: optimizing what?

SVM objective:

$$\sum_{i=1}^n \max\{0, 1 - y_i(b + x_i^T w)\} + \lambda \|w\|_2^2 = \sum_{i=1}^n \ell_i(w, b)$$

$$\nabla_w \ell_i(w, b) = \begin{cases} -x_i y_i + \frac{2\lambda}{n} w & \text{if } y_i(b + x_i^T w) < 1 \\ \frac{2\lambda}{n} & \text{otherwise} \end{cases}$$

$$\nabla_b \ell_i(w, b) = \begin{cases} -y_i & \text{if } y_i(b + x_i^T w) < 1 \\ 0 & \text{otherwise} \end{cases}$$