Today

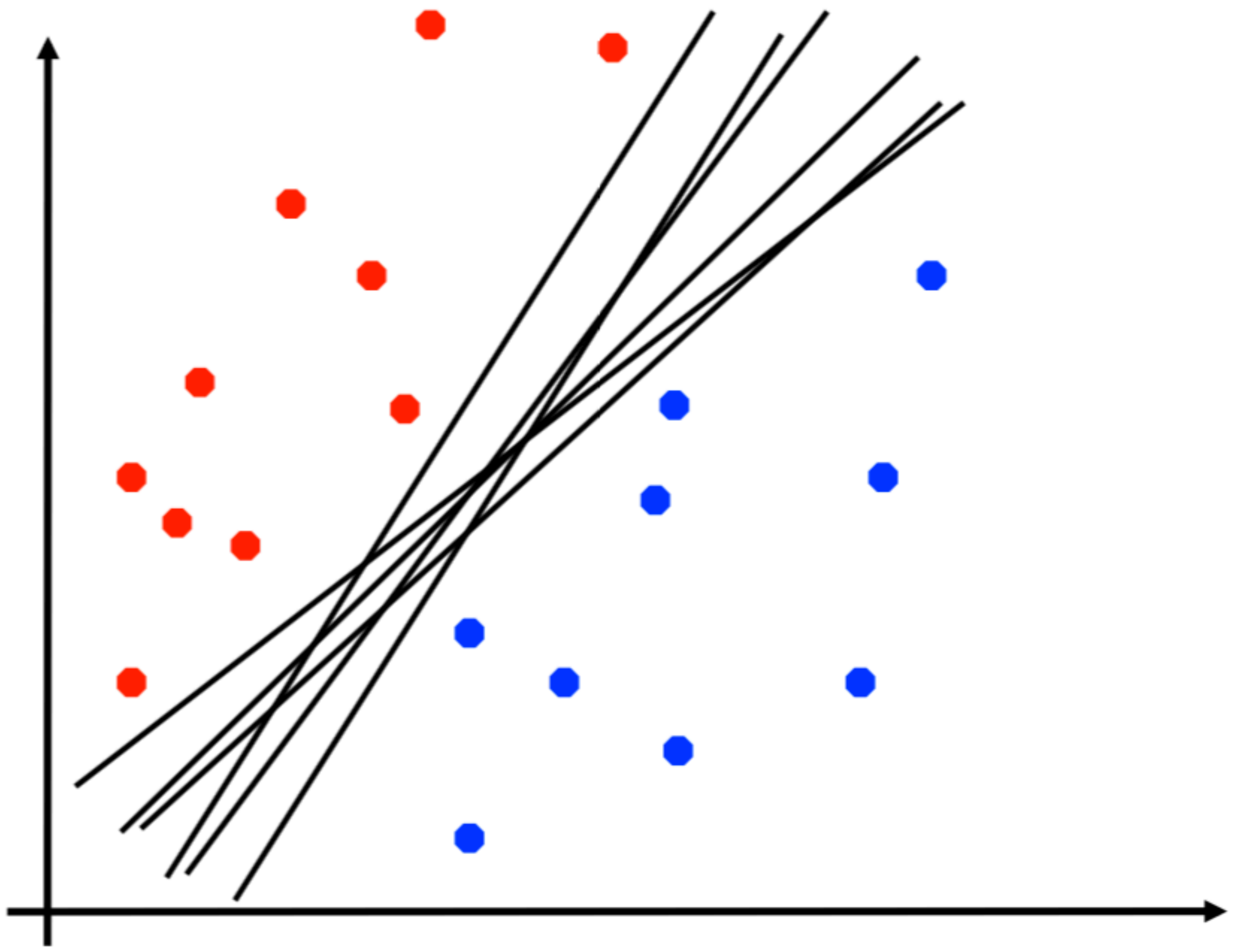
- Last time
  - Generalized linear models
  - Multi-class classification

- Today
  - Towards support vector machines
    - Perceptron algorithm (e.g., Hastie Section 4.5, Bishop Section 4.1.7)
  - Support vector machines

- Announcements
  - Homework 1 is due Wed, Sep 30 before class
  - Feedback https://forms.gle/VuCpbuRoPyZrohf57

Towards support vector machines - finding linear separators

Finding linear separators of two classes



[David Sontag]

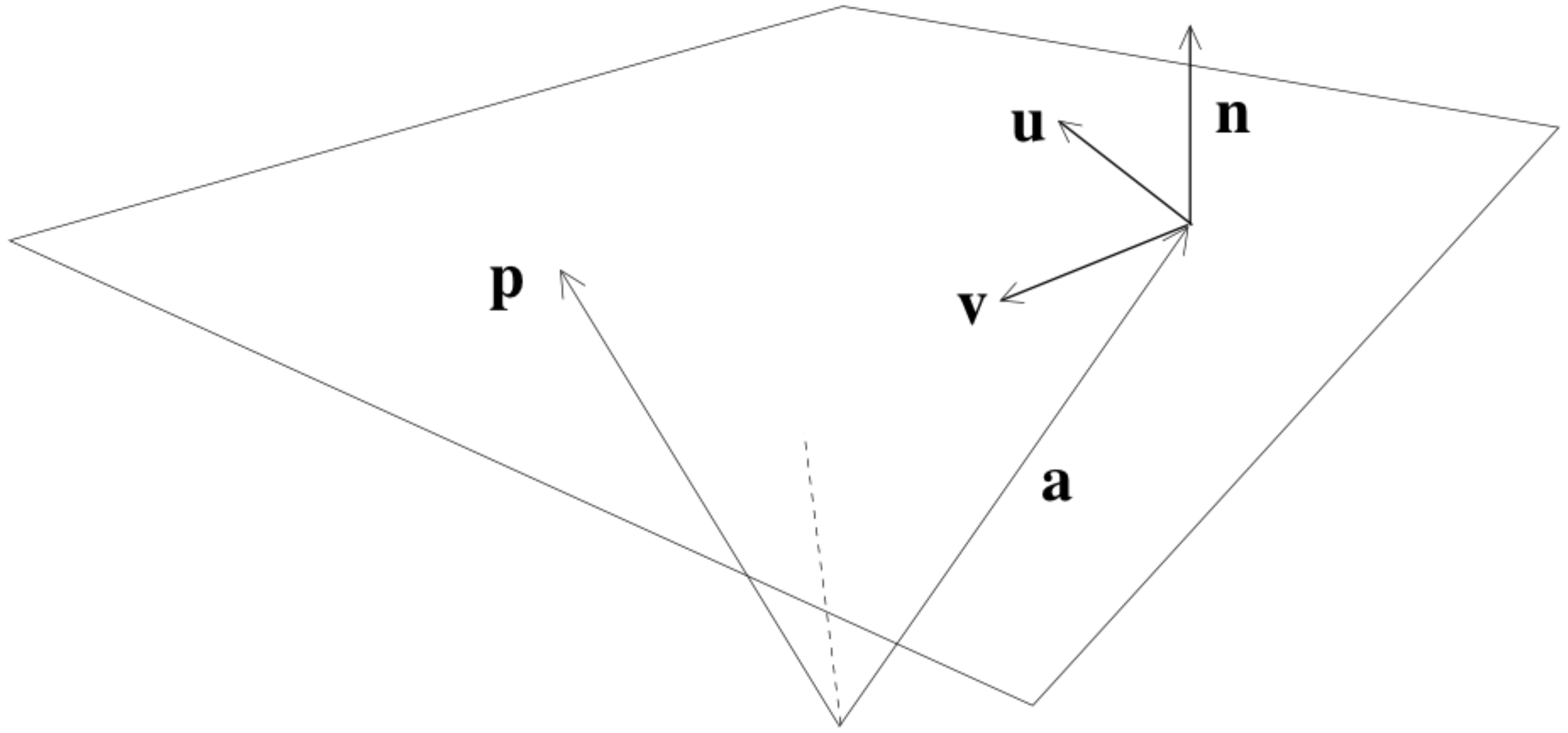
Geometry of linear separators (see blackboard)

A plane can be specified as the set of all points given by:

$$\mathbf{p} = \mathbf{a} + s\mathbf{u} + t\mathbf{v}, \qquad (s,t) \in \mathcal{R}.$$

Vector from origin to a point in the plane

Two non-parallel directions in the plane



Alternatively, it can be specified as:

$$(\mathbf{p} - \mathbf{a}) \cdot \mathbf{n} = 0 \Leftrightarrow \mathbf{p} \cdot \mathbf{n} = \mathbf{a} \cdot \mathbf{n}$$

Normal vector
(we will call this w)

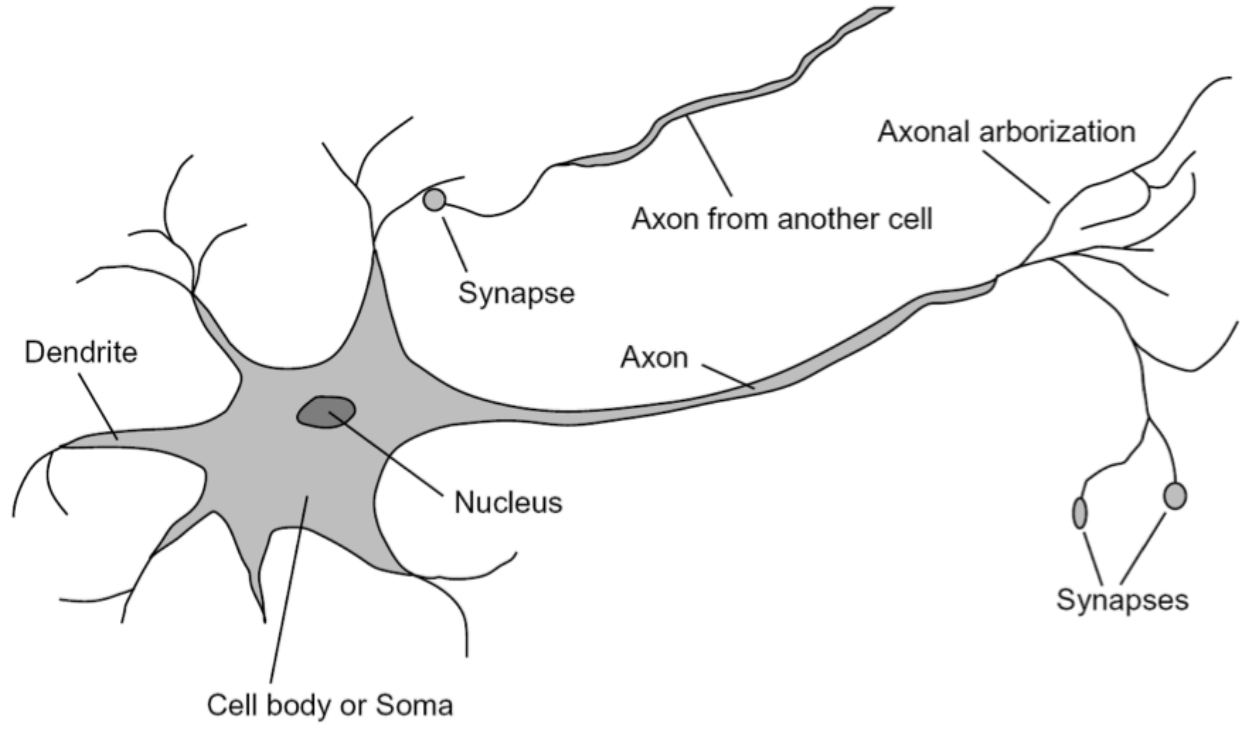
Only need to specify this dot product,
a scalar (we will call this the offset, b)

Figure: Barber, Section 29 (appendix)

[David Sontag]

Linear Classifiers

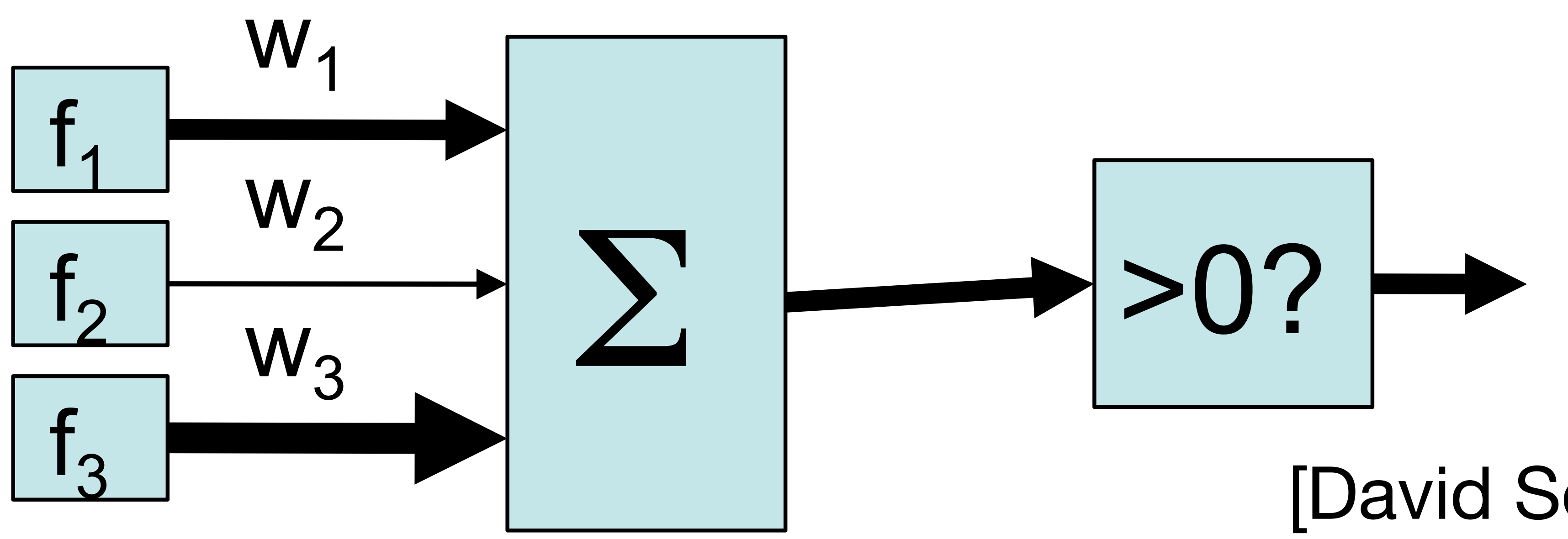
- Inputs are feature values
- Each feature has a weight
- Sum is the activation



Important note: changing notation!

$$\text{activation}_w(x) = \sum_i w_i \cdot f_i(x) = w \cdot f(x)$$

- If the activation is:
  - Positive, output *class 1*
  - Negative, output *class 2*



[David Sontag]

Example: Spam

- Imagine 3 features (spam is "positive" class):
  1. free (number of occurrences of "free")
  2. money (occurrences of "money")
  3. BIAS (intercept, always has value 1)

$$w \cdot f(x)$$

$$\sum_i w_i \cdot f_i(x)$$

x	$f(x)$	w
"free money"	BIAS : 1 free : 1 money : 1 ...	BIAS : -3 free : 4 money : 2 ...

$$\begin{aligned} &(1)(-3) \quad + \\ &(1)(4) \quad + \\ &(1)(2) \quad + \\ &\ldots \\ &= 3 \end{aligned}$$

w.f(x) > 0 ➔ SPAM!!!

Note: The BIAS term determines the threshold

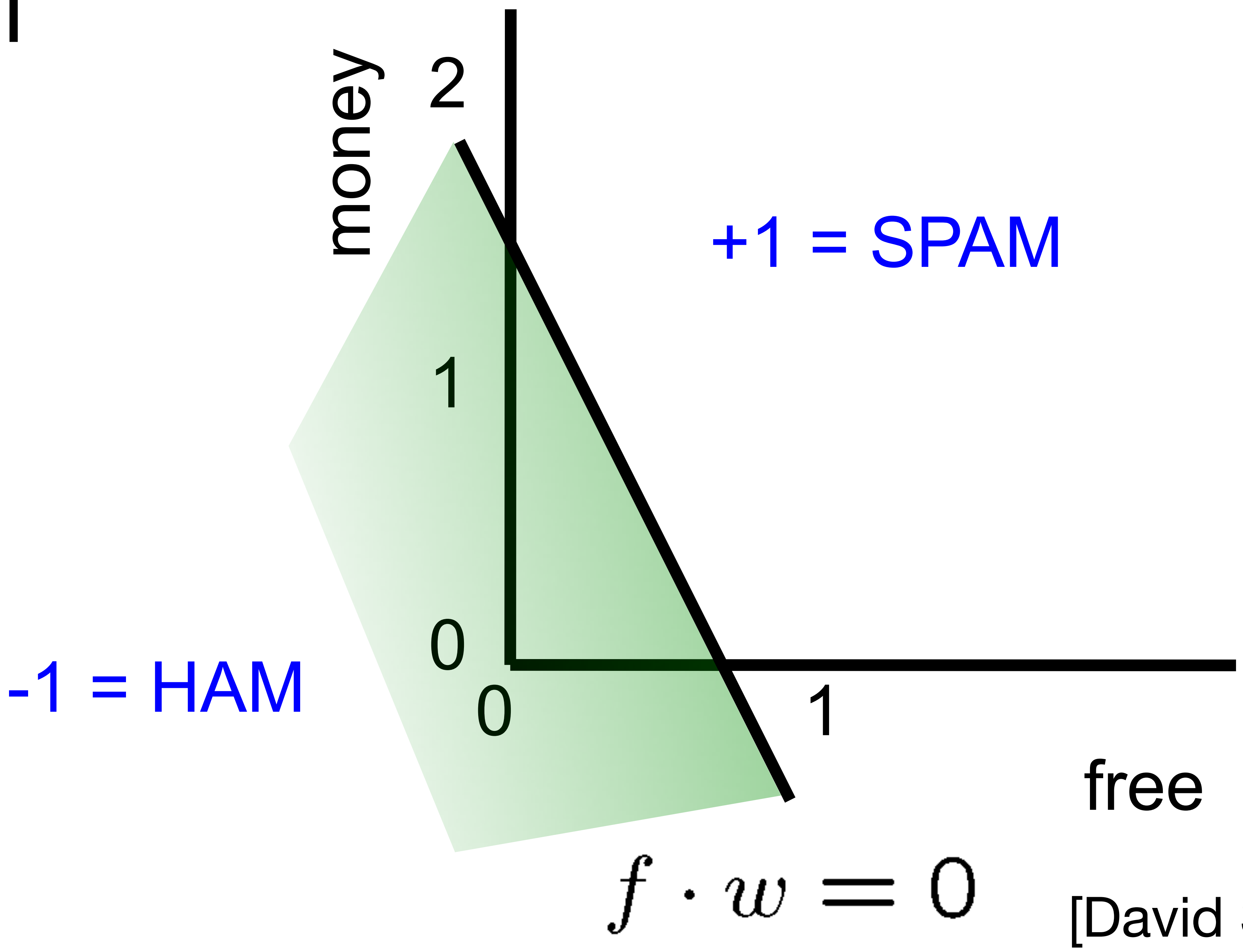
[David Sontag]

Binary Decision Rule

- In the space of feature vectors
  - Examples are points
  - Any weight vector is a hyperplane
  - One side corresponds to Y=+1
  - Other corresponds to Y=-1

w

```
BIAS  : -3
free  :  4
money :  2
...
```



[David Sontag]

The perceptron algorithm

- Start with weight vector = $\vec{0}$
- For each training instance (x_i, y_i):
  - Classify with current weights

$$y = \begin{cases} +1 & \text{if } \; w \cdot f(x_i) \geq 0 \\ -1 & \text{if } \; w \cdot f(x_i) < 0 \end{cases}$$

  - If correct (i.e., $y=y_i$), no change!
  - If wrong: update

$$w = w + y_i \, f(x_i)$$

[David Sontag]

Geometrical interpretation on board

Preceptron and logistic regression

- Logistic regression with $h_\theta(x) = g(\theta^T x)$ with $g(z) = 1/(1 + e^{-z})$
- Now set

$$g(z) = \begin{cases} 1\,, & z \geq 0 \\ 0\,, & z < 0 \end{cases}$$

and use SGD update rule as before

$$\theta^{(k+1)} = \theta^k + \alpha \left(y^{(i)} - h_{\theta^{(k)}}(x^{(i)}) \right) x^{(i)}$$

gives the perceptron algorithm (with $\alpha = 1/2$)

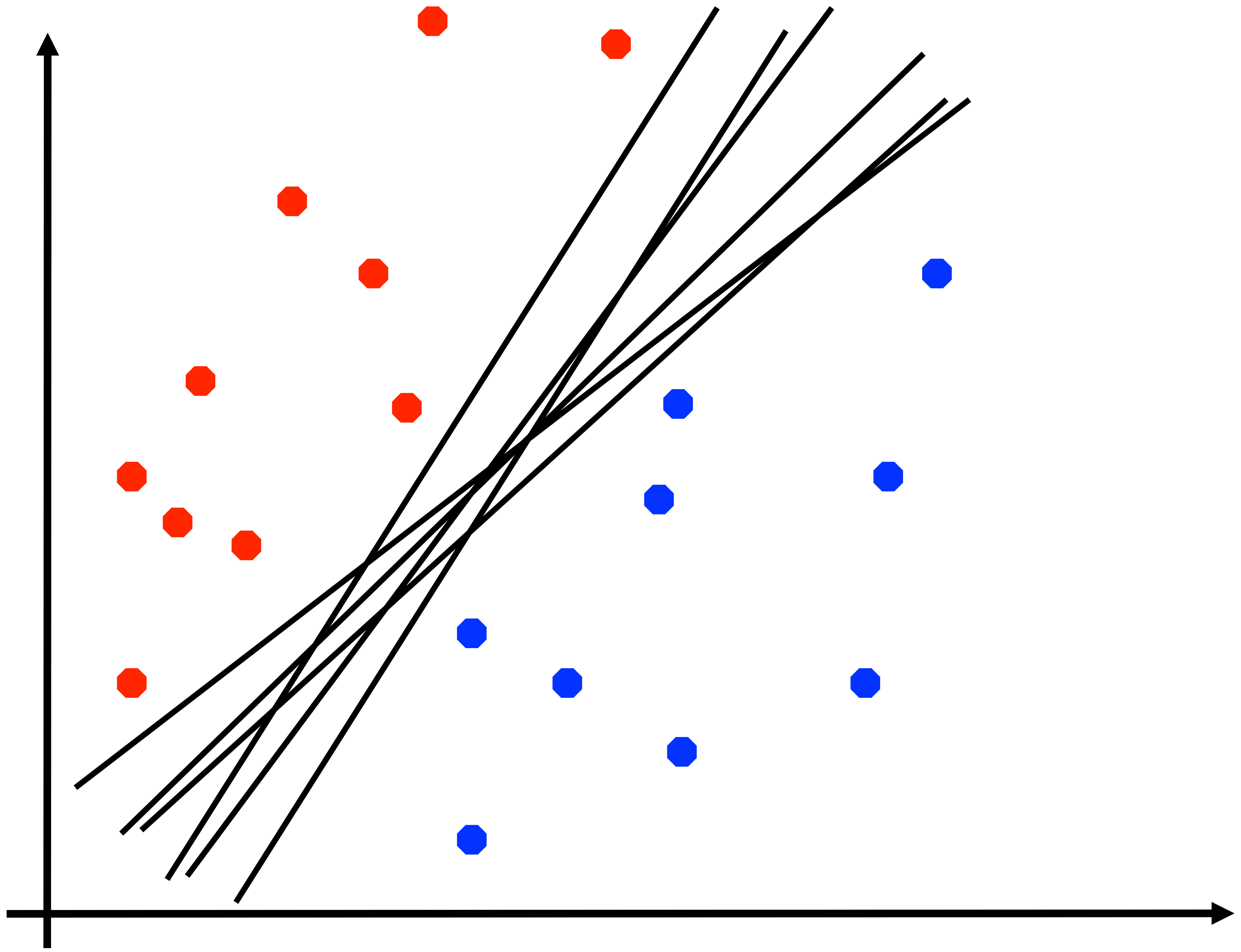
- If misclassified ($y^{(i)} - h_{\theta^{(k)}}(x^{(i)}) \neq 0$), then update weight $\theta^{(k)}$ with $x^{(i)}$
- Note that $y^{(i)} - h_{\theta^{(k)}}(x^{(i)}) = -2$ if $y^{(i)} = -1$ and $h_{\theta^{(k)}}(x^{(i)}) = 1$

Logistic regression

- Consider $g(z) = 1/(1 + e^{-z})$ and recall that the probability
$$p(y = 1 \,|\, x; \theta) = h_\theta(x) = g(\theta^T x)$$
- Predict label “1” if $h_\theta(x) \geq 0.5$ which is equivalent to $\theta^T x \geq 0$
- The larger $h_\theta(x)$, the more confident we are that we correctly predict the label “1”, i.e., $\theta^T x \gg 0$
- Similarly, confident in prediction “0” if $\theta^T x \ll 0$
- Note that θ is the normal vector of a hyperplane

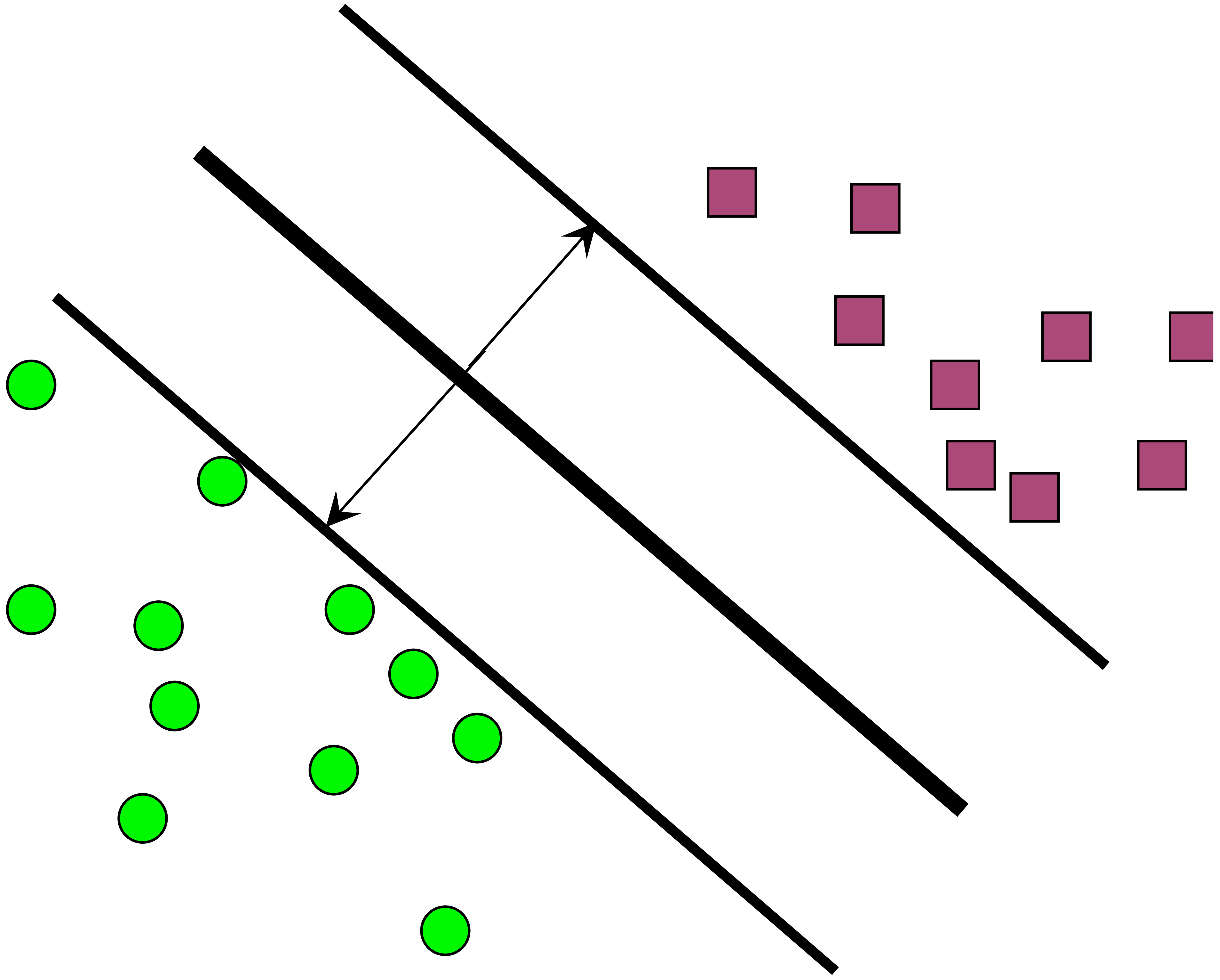
Linear Separators

- Which of these linear separators is optimal?



[David Sontag]

Find separator with largest margin

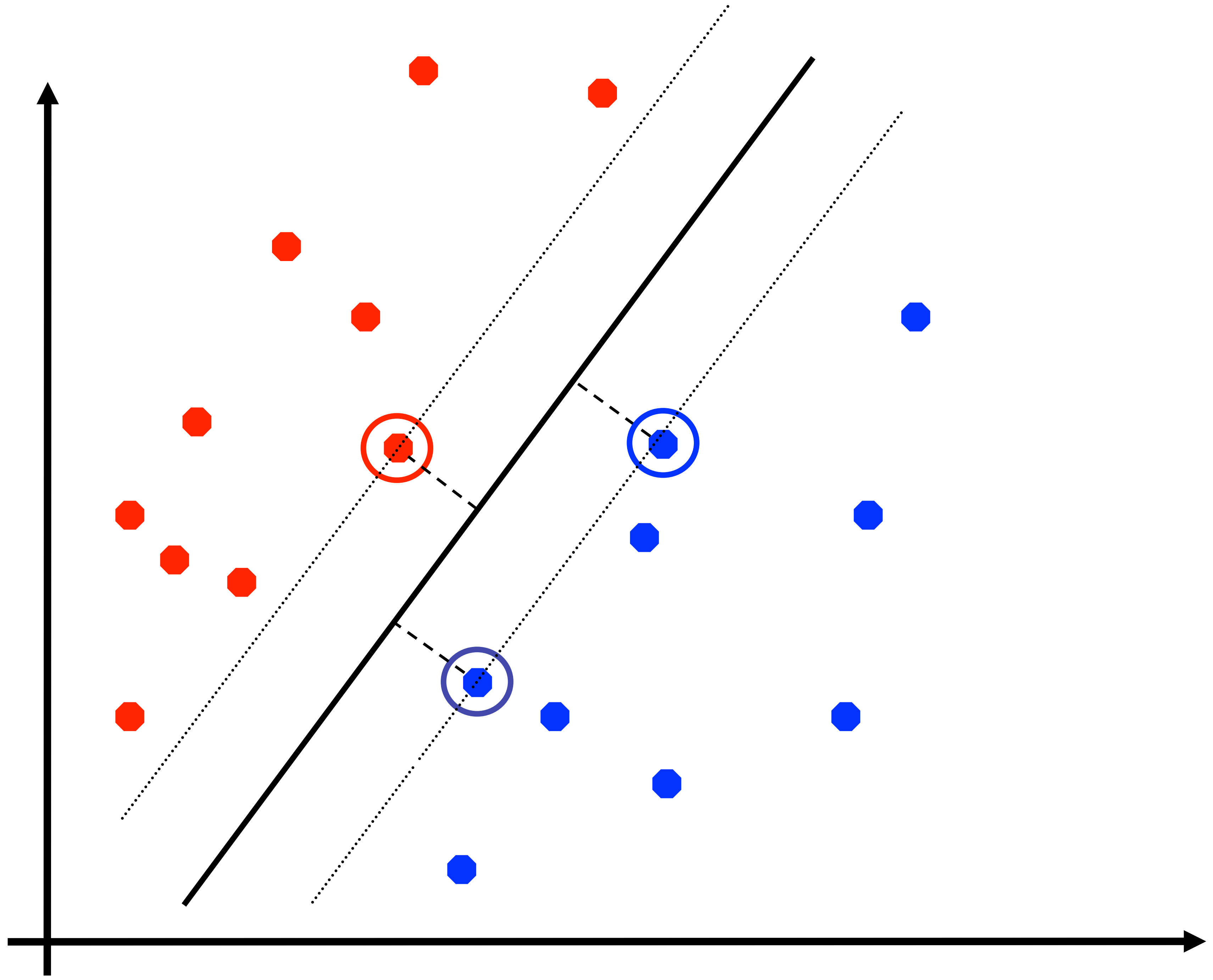


[David Sontag]

Support vector machines

Support vector machines

- SVMs (Vapnik, 1990's) choose the linear separator with the **largest margin**



- Good according to intuition, theory, practice

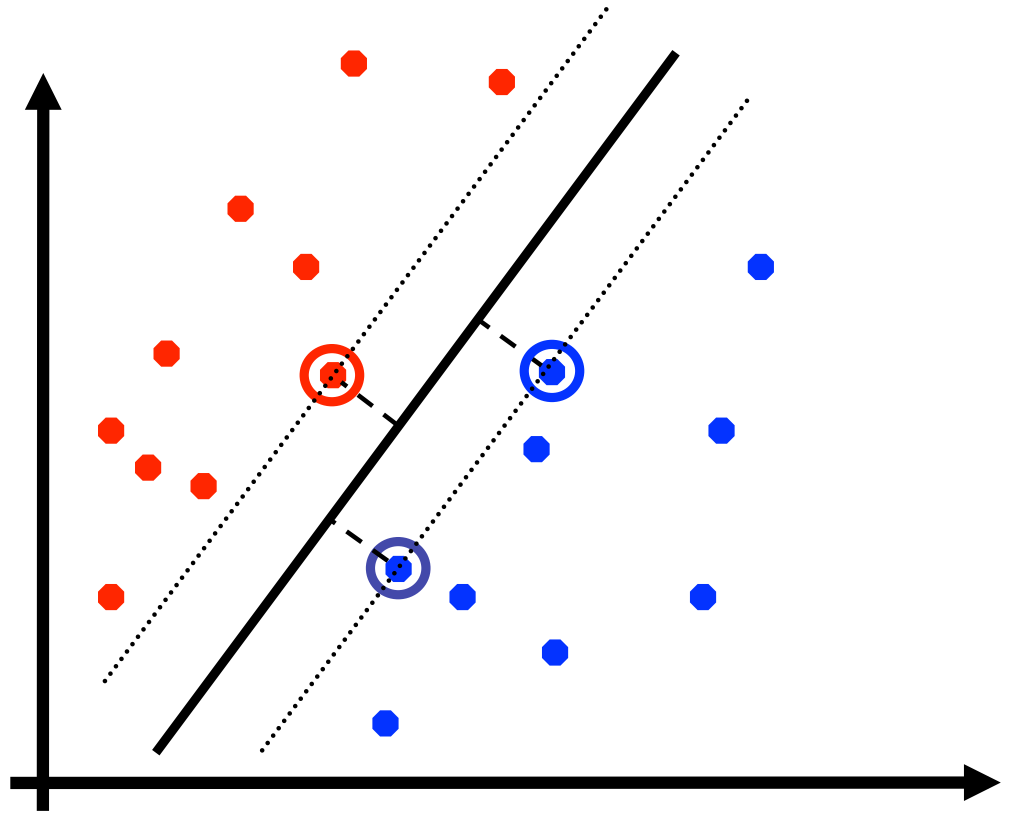
[David Sontag]

Support vector machines: 3 key ideas

1. Use **optimization** to find solution (i.e. a hyperplane) with few errors

2. Seek **large margin** separator to improve generalization

3. Use **kernel trick** to make large feature spaces computationally efficient



[David Sontag]

Notation

- Class labels $y \in \{-1,1\}$
- Parametrize with w, b rather than θ (intercept treated separately)

$$h_\theta(x) = g(w^T x + b)$$

with

$$g(z) = \begin{cases} 1\,, & z \geq 0 \\ 0\,, & z < 0 \end{cases}$$

Functional margin

- Define functional margin of training sample $(x^{(i)}, y^{(i)})$ w.r.t. (w, b) as
$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$$
    - If $y^{(i)} = 1$, then need $w^T x + b \gg 0$ for $\hat{\gamma}^{(i)}$ large
    - If $y^{(i)} = -1$, then need $w^T x + b \ll 0$ for $\hat{\gamma}^{(i)}$ large
    - Holds $y^{(i)}(w^T x + b) > 0$, then prediction correct
- Large functional margin = confident + correct prediction
- Scaling
    - For our choice $g(z) = 1?z \geq 0 : 0$ have
$$g(2w^T x + 2b) = g(w^T x + b)$$
which means that h_θ is invariant under scaling even though $\hat{\gamma}^{(i)}$ is not
—> normalize by enforcing $\|\boldsymbol{w}\| = 1$

Geometric margin

[board]

Geometric margin

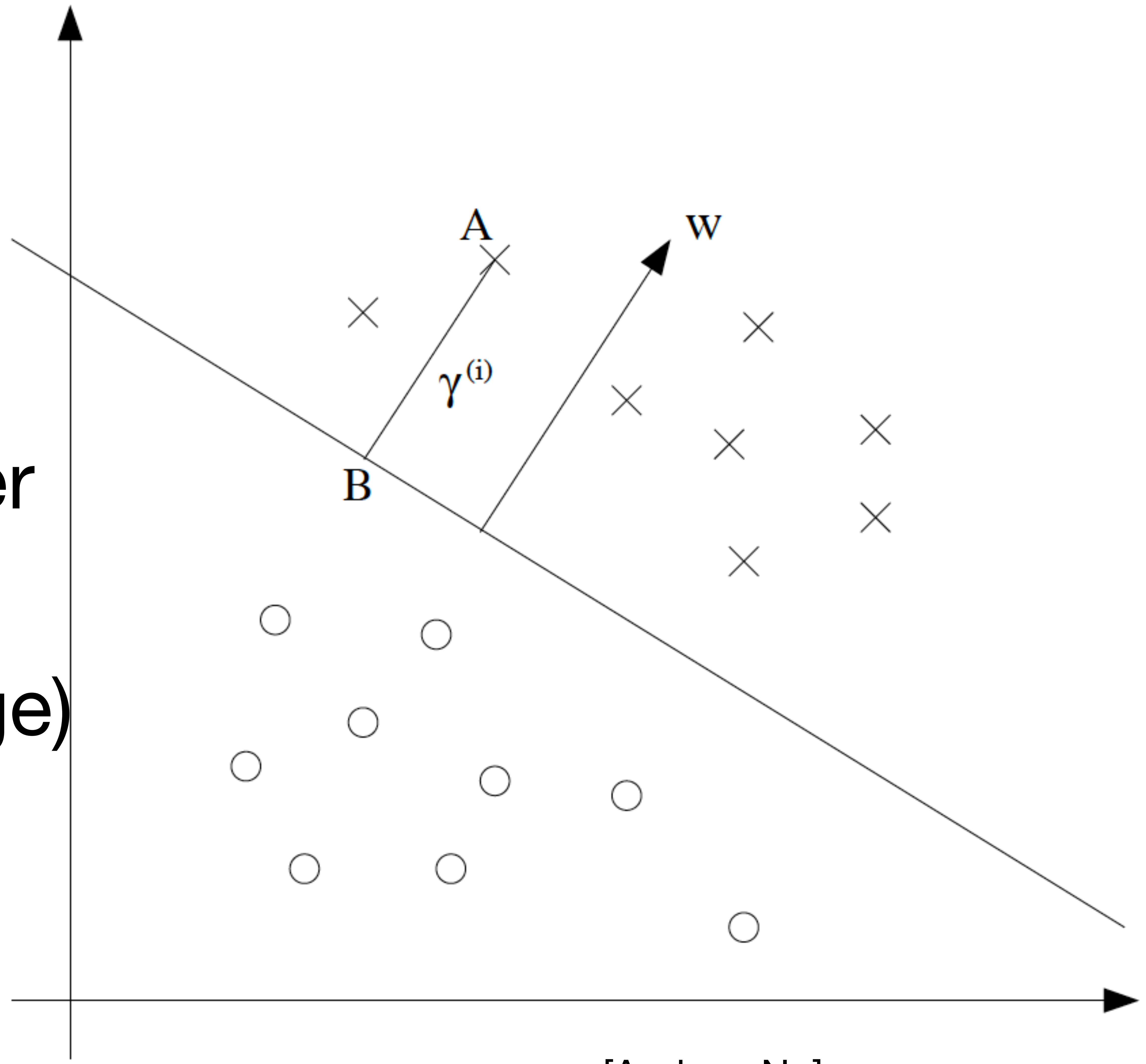
- Geometric margin of (w, b) w.r.t. $(x^{(i)}, y^{(i)})$

$$\gamma^{(i)} = y^{(i)} \left(\frac{w^T}{\|w\|} x^{(i)} + \frac{b}{\|w\|} \right)$$

- Geometric margin is invariant under scaling of w, b (e.g., replace w, b with $2w, 2b$ then $\gamma^{(i)}$ doesn't change)

- Geometric margin w.r.t. set $\mathscr{D}$

$$\gamma = \min_{i=1,\ldots,N} \gamma^{(i)}$$



[Andrew Ng]

Optimal margin classifier

- Find decision boundary that maximizes geometric margin
- **Assumption: Training set $\mathscr{D}$ is linearly separable**
- Pose optimization problem

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \textbf{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma\,, \quad i = 1,\ldots,N \\ & \|w\| = 1 \end{aligned}$$

- Constraint $\|w\| = 1$ ensures that functional margin $((\hat{\gamma}^{(i)} =)y^{(i)}(w^T x^{(i)} + b))$ is equal to geometric margin
- Constraint $\|w\| = 1$ leads to non-convex set (hard to optimize over) [in contrast to $\|w\| \leq 1$]

Optimal margin classifier (cont'd)

- Note that functional margin $\hat{\gamma}$ and geometric margin γ are related as
$$\gamma = \hat{\gamma}/||w||$$
- Optimize normalized functional margin
$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$
$$\textbf{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1,\ldots,N$$

- Got rid of constraint $||w|| = 1$ but introduced objective $\hat{\gamma}/||w||$

Optimal margin classifier (cont'd)

- Invoke that functional margin $\hat{\gamma}$ depends on scaling
  - Multiplying w, b by constant, multiplies $\hat{\gamma}$ by that constant
- Introducing constraint $\hat{\gamma} = 1$, which indeed is a scaling constraint on w, b and obtain

$$\begin{array}{ll} \min\limits_{w,b} & \frac{1}{2}\|w\|^2 \\ \textbf{s.t.} & y^{(i)}(w^T x^{(i)} + b) \geq 1\,, \quad i = 1,\ldots,N \end{array}$$

- Note: maximizing $\hat{\gamma}/\|w\|$ (with $\hat{\gamma} = 1$) is same as minimizing $\|w\|^2$
- Convex quadratic objective, linear constraints
- The solution is the optimal margin classifier