

CIS 5200: MACHINE LEARNING

SUPPORT VECTOR MACHINES (SVM)

Surbhi Goel

Content here draws from material by Jake/Shivani (UPenn), Vatsal Sharan (USC), Christopher De Sa (Cornell)



Spring 2023

LOGISTICS - UPCOMING

Homework:

- * HW2 is out and is due on **Friday, Feb 17, 2023** end of day
- * There is a ***survey***, don't miss!
- * HW1 solutions will be uploaded soon
- * HW1 grading will be done by Monday, Feb 12, 2023

OUTLINE - TODAY

- * Back to Binary Classification
- * Hard-margin SVMs
 - * Formulation
 - * Dual version
 - * Support vectors
- * Soft-margin SVMs
 - * Formulation
 - * Optimization viewpoint

SUPERVISED LEARNING - BINARY CLASSIFICATION

Input space: $\mathcal{X} \subseteq \mathbb{R}^d$

Output space: $\mathcal{Y} = \{-1, 1\}$

Predictor function: $f : \mathcal{X} \rightarrow \mathcal{Y}, f \in \mathcal{F}$

Loss function: $\ell(f(x), y) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{otherwise.} \end{cases}$

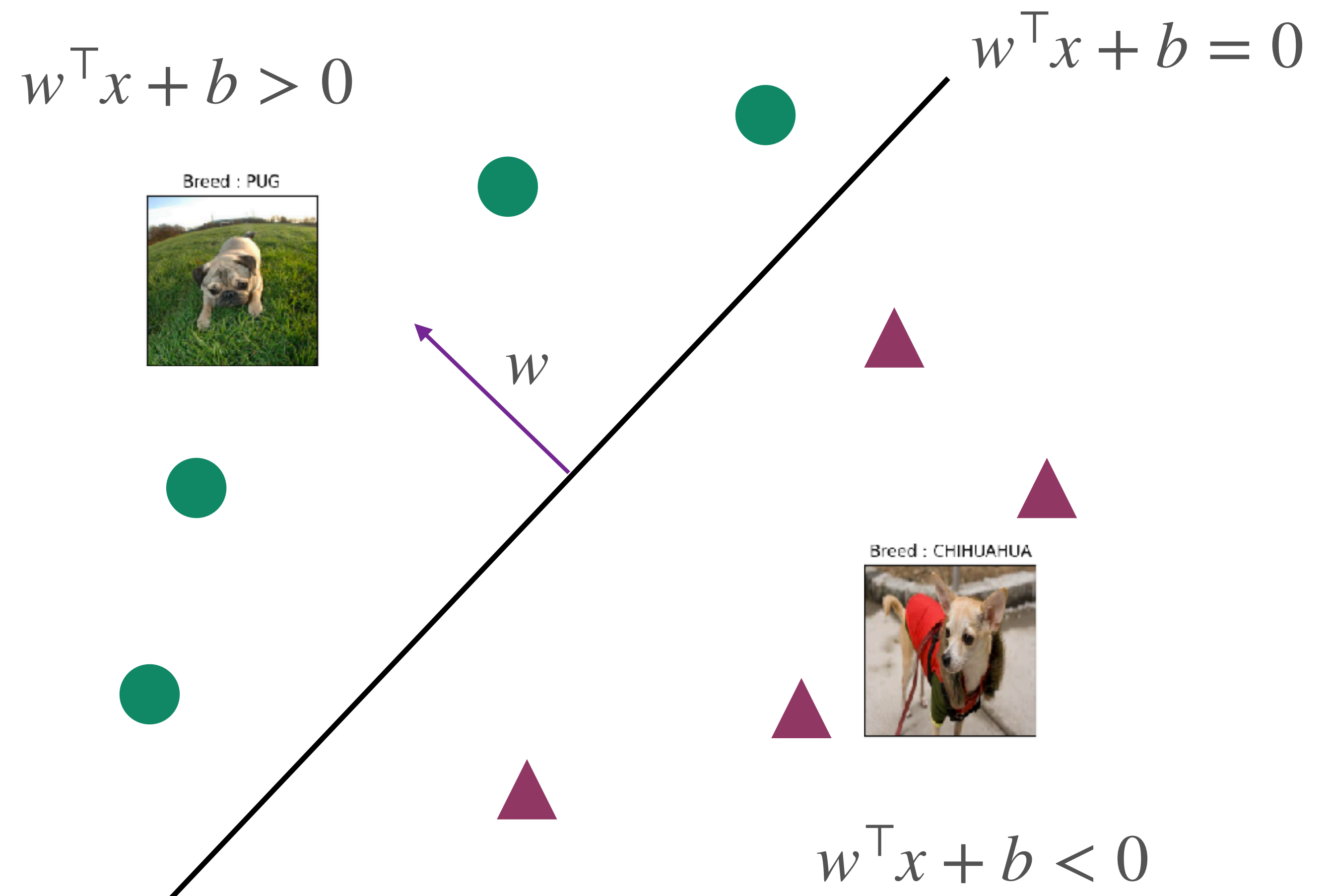
Data: $\{(x_1, y_1), \dots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from distribution \mathcal{D}

HYPOTHESIS CLASS - LINEAR CLASSIFIER

We will keep the bias

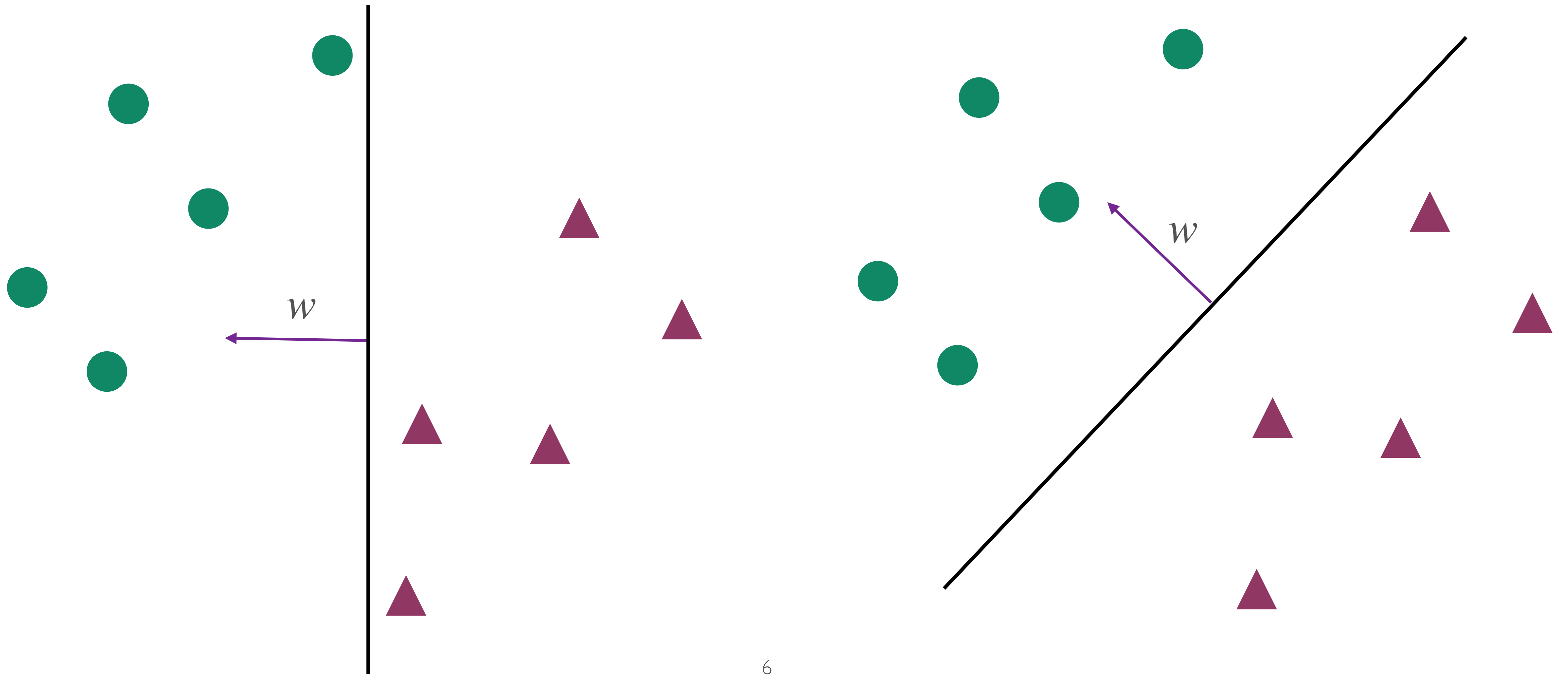
Linear Classifier: $\mathcal{F} := \{x \mapsto \text{sign}(w^\top x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$

$$\text{sign}(a) = \begin{cases} +1 & \text{if } a \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$



BEST SEPARATING HYPERPLANE - MAX-MARGIN

Which hyperplane is better?



BEST SEPARATING HYPERPLANE - MAX-MARGIN

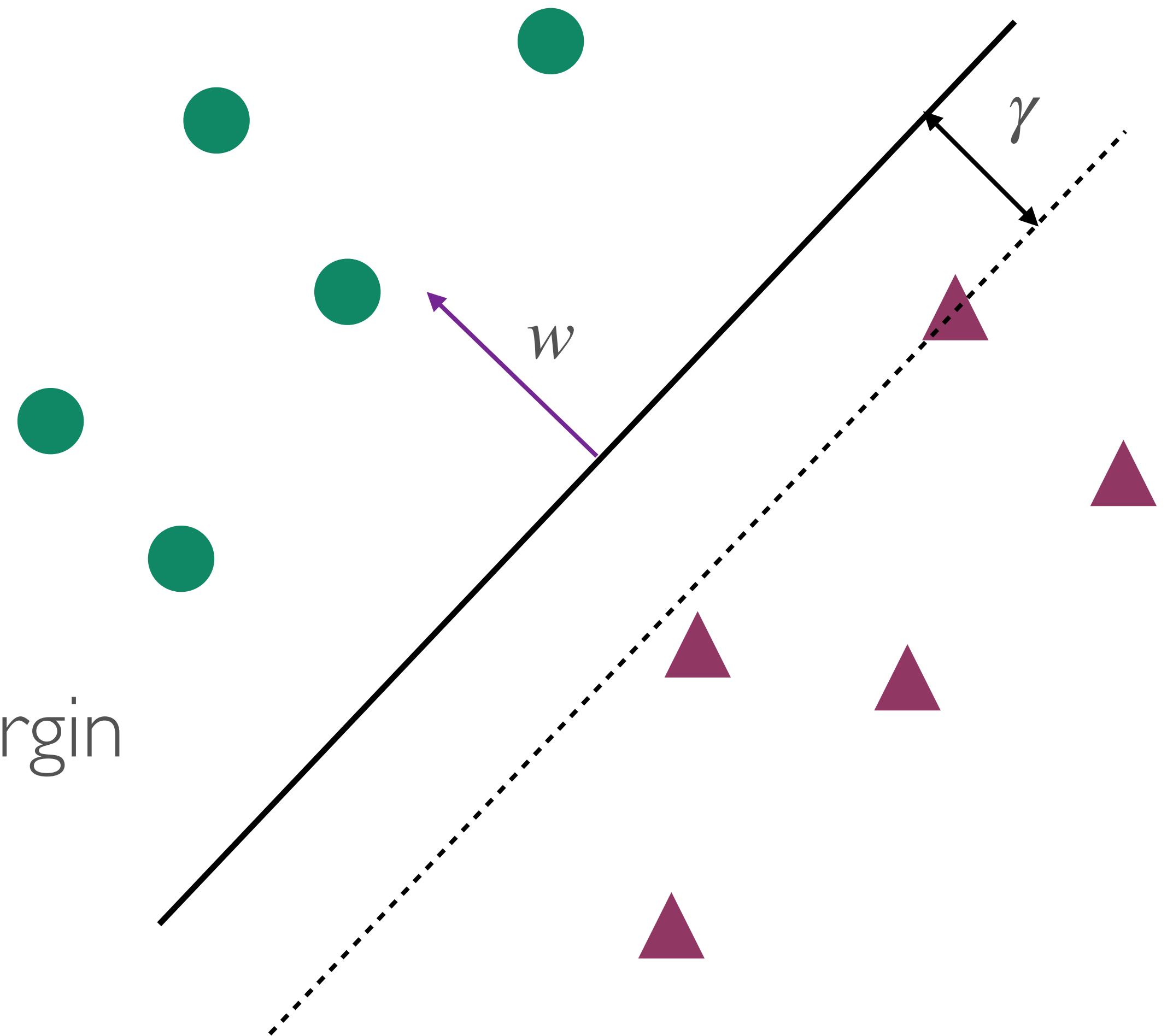
Margin of a hyperplane $w^T x + b = 0$

$$\gamma(w, b) = \min_{i \in [m]} \frac{|w^T x_i + b|}{\|w\|_2}$$

Distance of closest point from the hyperplane

SVM finds a hyperplane that maximizes margin

Margin Perceptron found a hyperplane with margin $\gamma/3$ not γ



OPTIMIZATION PROBLEM - MAX-MARGIN

$$\max_{w,b} \underbrace{\gamma(w,b)}_{\text{margin}}$$

such that

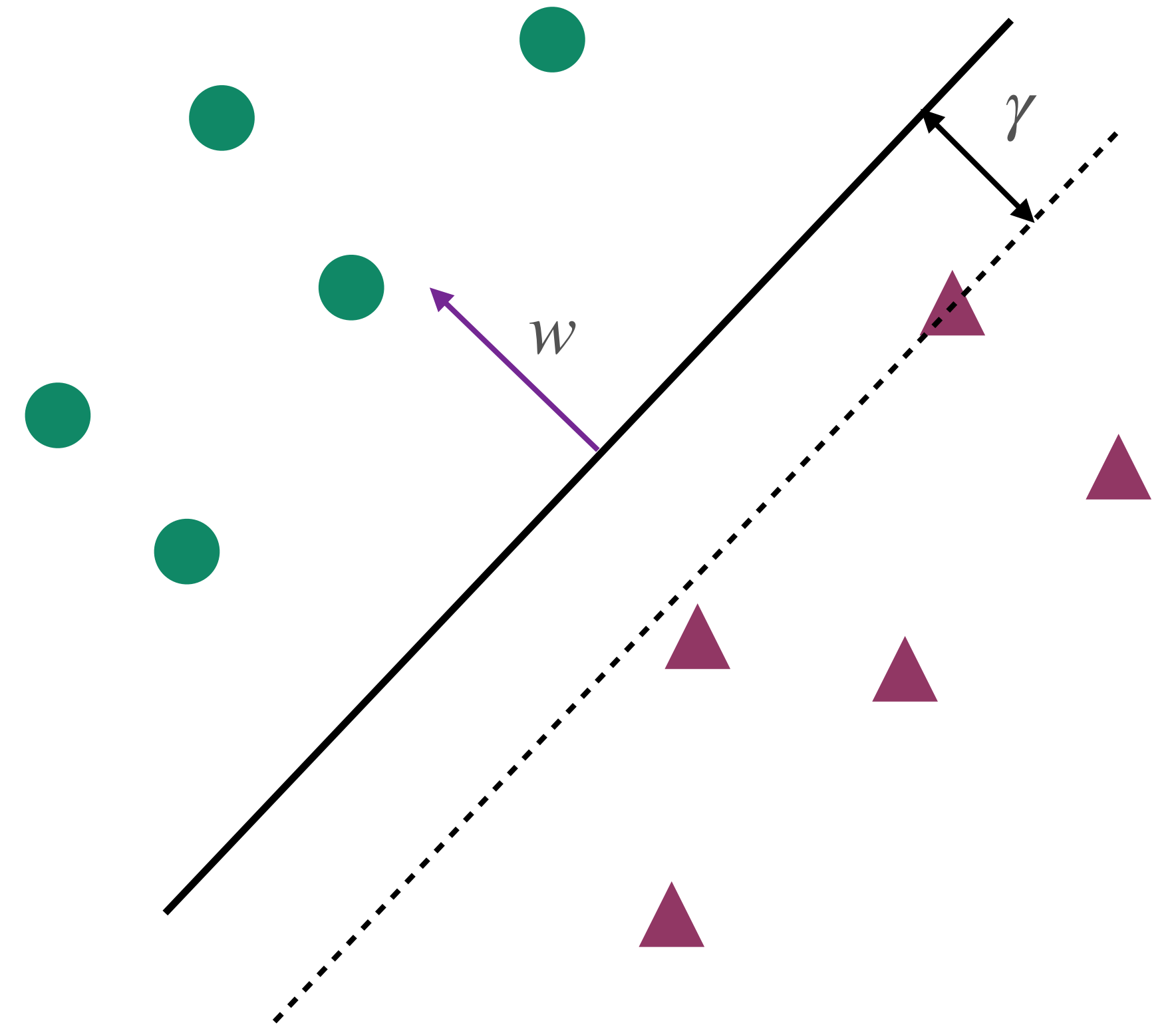
$$\underbrace{y_i(w^\top x_i + b) \geq 0, \forall i \in [m]}_{(w,b) \text{ linearly separates data}}$$

Substituting for margin:

$$\max_{w,b} \underbrace{\frac{1}{\|w\|_2} \min_{i \in [m]} |w^\top x_i + b|}_{\text{margin}}$$

such that

$$\underbrace{y_i(w^\top x_i + b) \geq 0, \forall i \in [m]}_{(w,b) \text{ linearly separates data}}$$



OPTIMIZATION PROBLEM - MAX-MARGIN

$$\max_{w,b} \underbrace{\frac{1}{\|w\|_2} \min_{i \in [m]} |w^\top x_i + b|}_{\text{margin}}$$

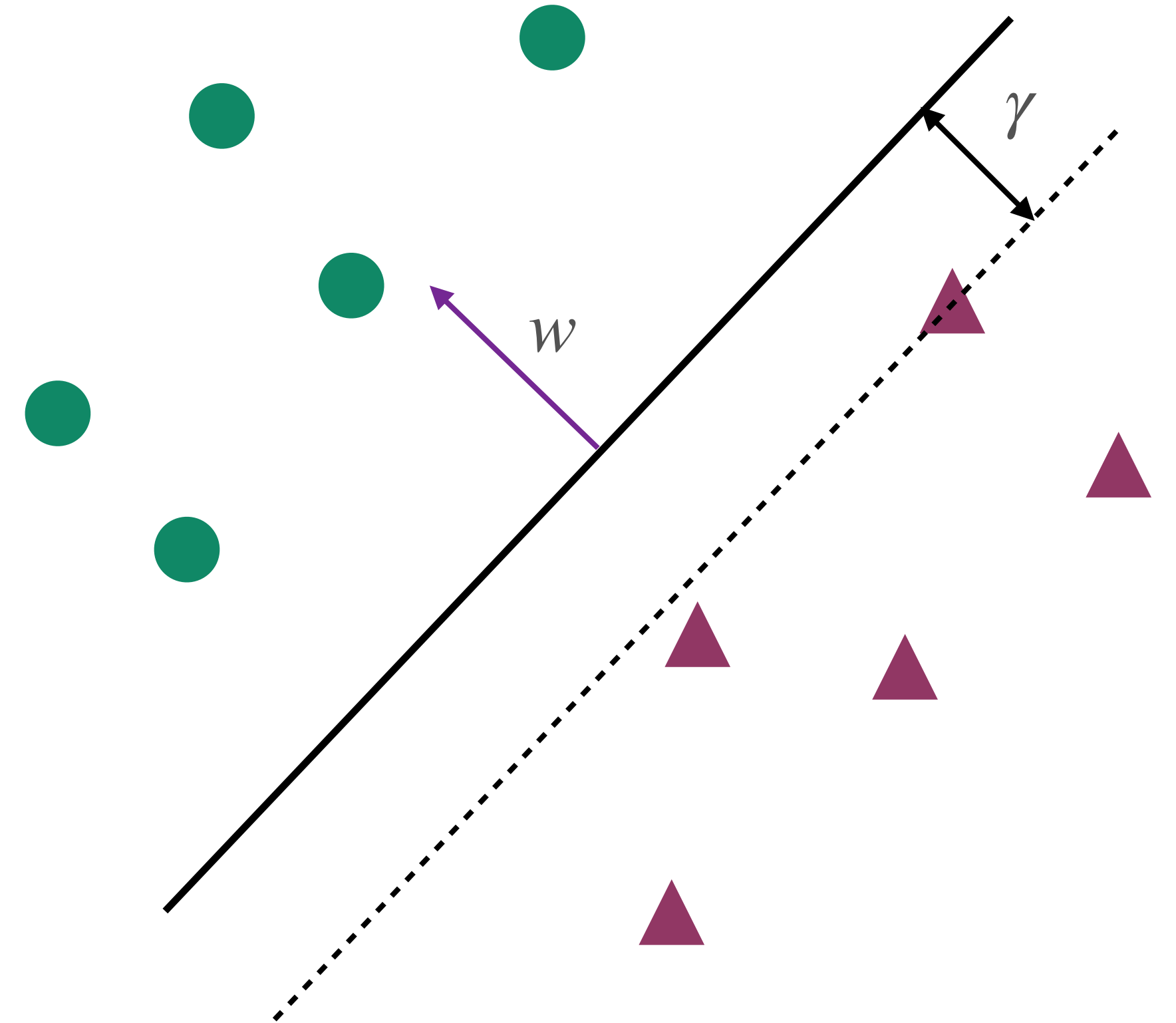
such that

$$\underbrace{y_i(w^\top x_i + b) \geq 0, \forall i \in [m]}_{(w,b) \text{ linearly separates data}}$$

Is there a unique solution?

We can fix the scale by setting $\min_{i \in [m]} |w^\top x_i + b| = 1$.

Puts a constraint on w, b



OPTIMIZATION PROBLEM - FIXED SCALE

Adding scale constraint:

$$\max_{w,b} \underbrace{\frac{1}{\|w\|_2}}_{\text{margin}}$$

$$\min_{w,b} \frac{1}{2} \|w\|_2^2$$

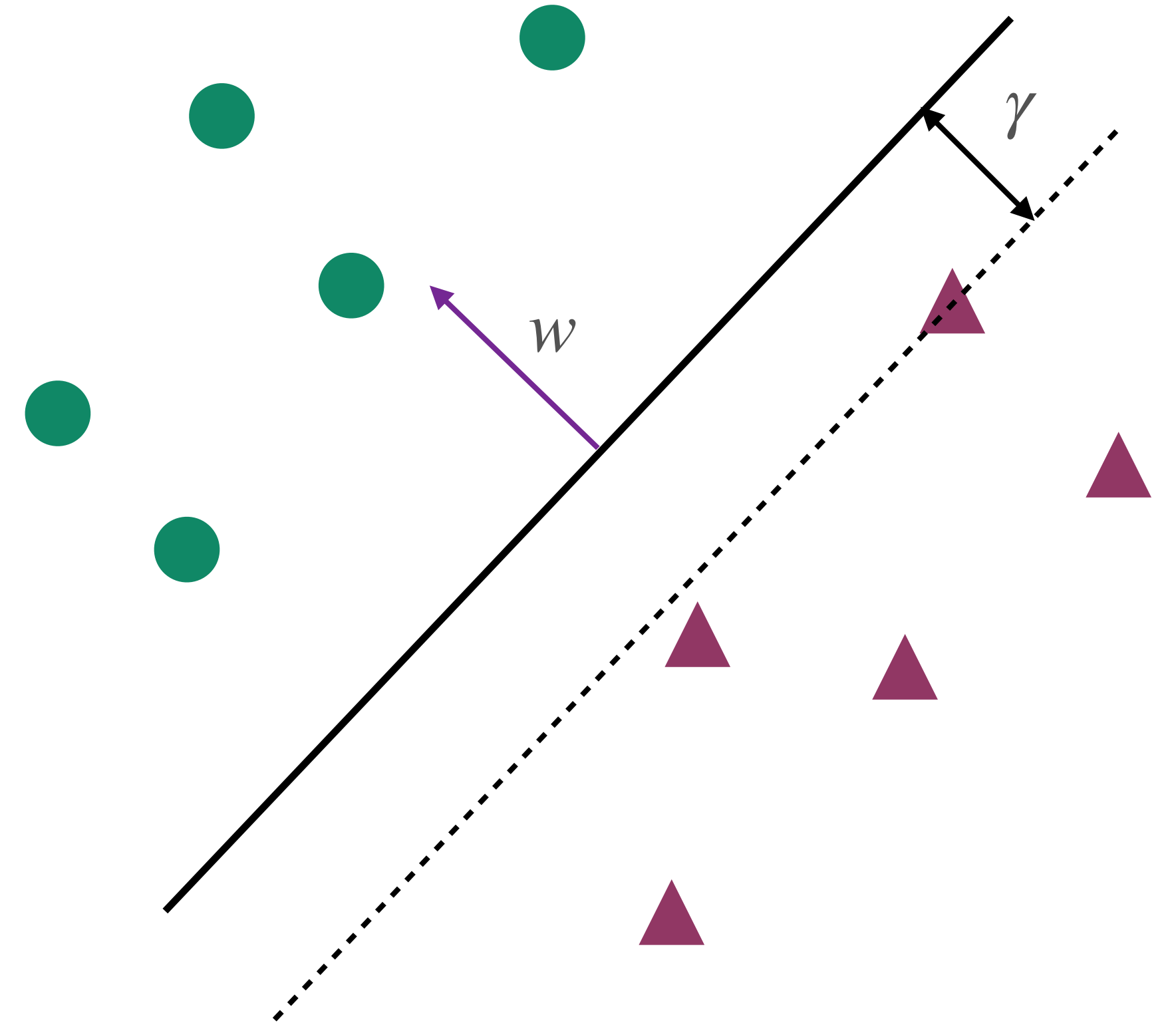
such that

$$y_i(w^\top x_i + b) \geq 0, \forall i \in [m]$$

(w,b) linearly separates data

$$\min_{i \in [m]} |w^\top x_i + b| = 1$$

fixed scale

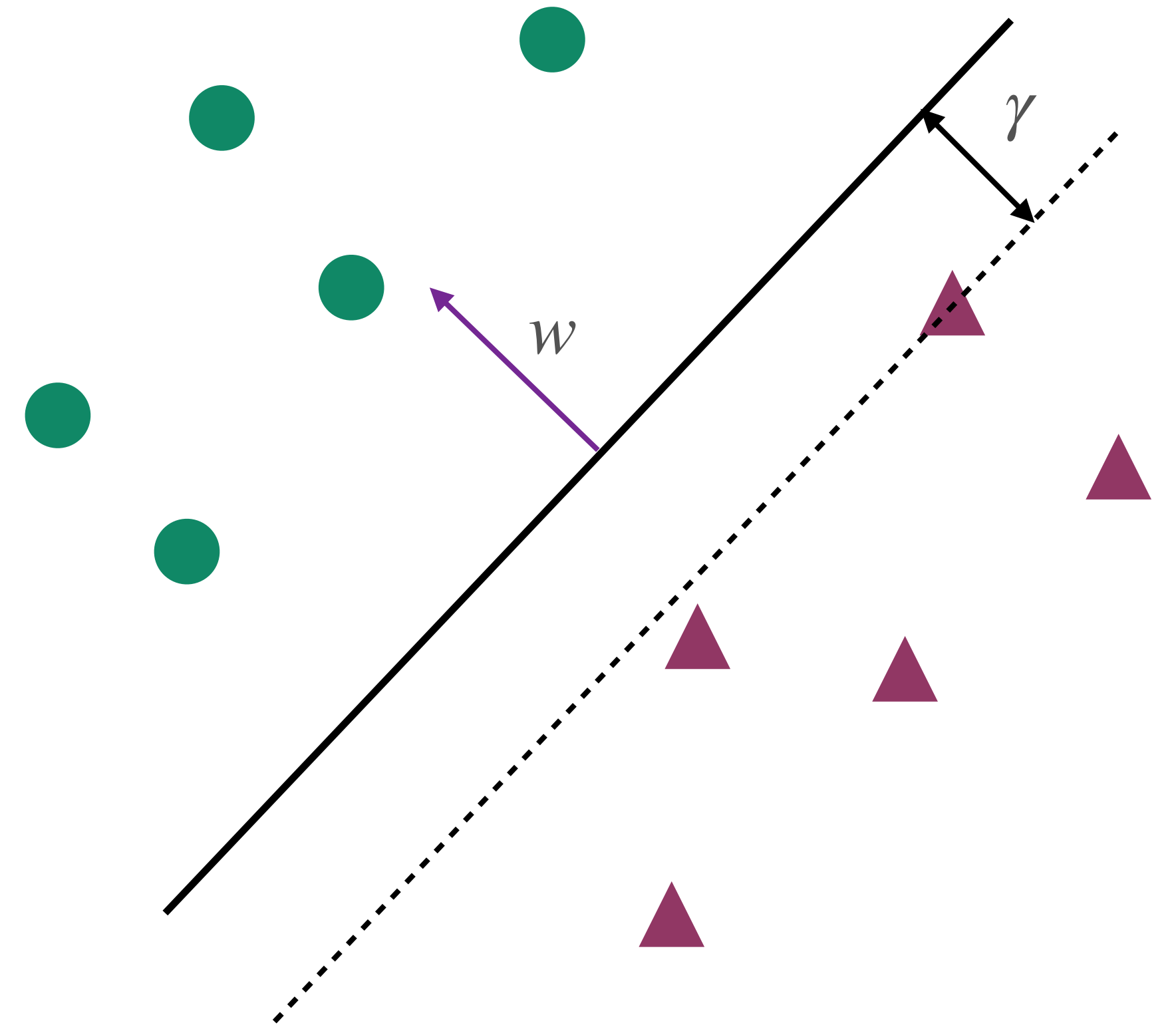


What kind of optimization problem is this?

OPTIMIZATION PROBLEM - QUADRATIC PROGRAM

$$\begin{aligned} & \min_{w,b} \quad \frac{1}{2} \|w\|_2^2 \\ & \text{such that} \quad y_i(w^\top x_i + b) \geq 0, \forall i \in [m] \\ & \quad \min_{i \in [m]} |w^\top x_i + b| = 1 \end{aligned}$$

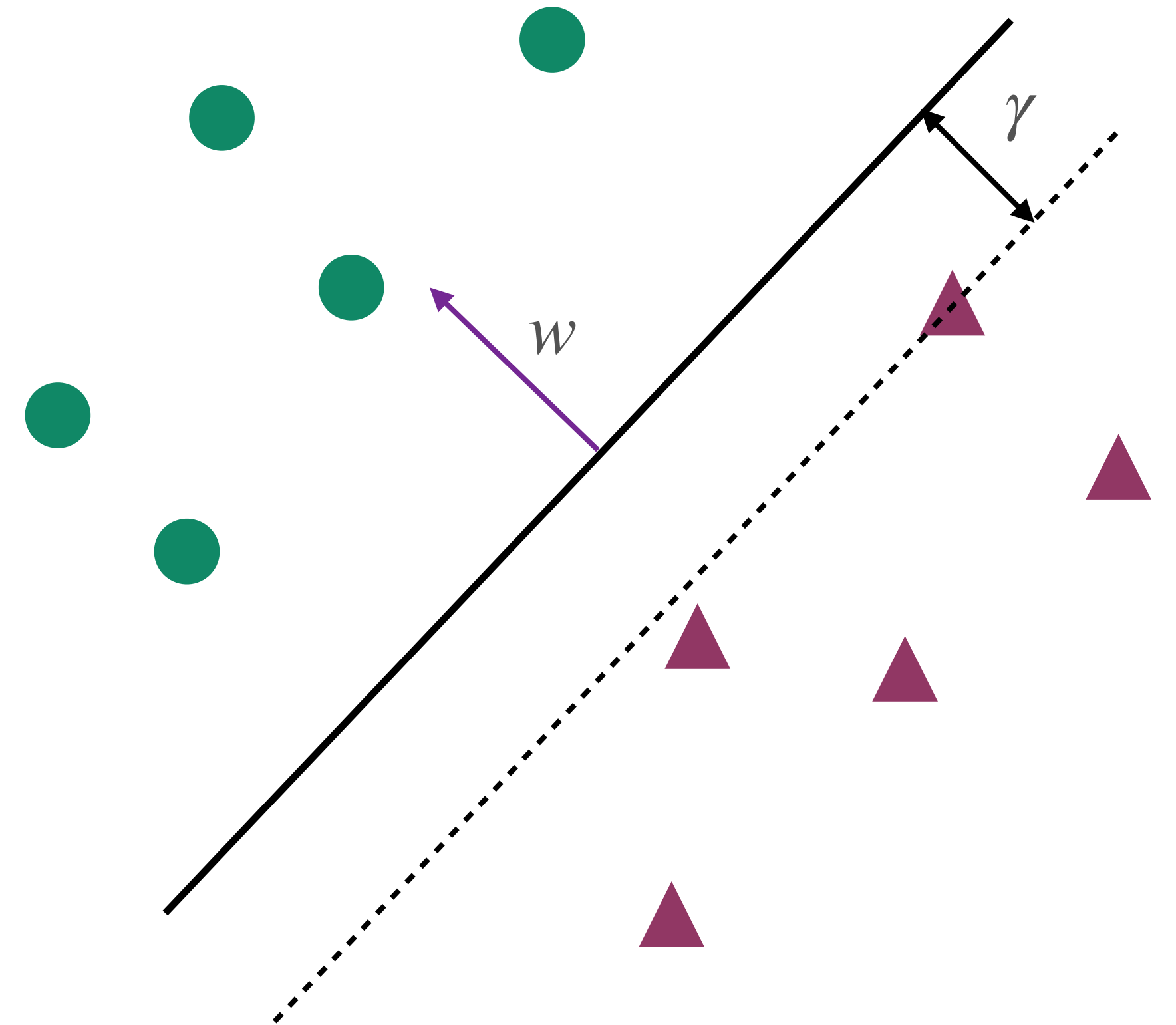
Is this a convex problem?



OPTIMIZATION PROBLEM - CONVEXTRICK

$$\begin{aligned} &\min_{w,b} \quad \frac{1}{2} \|w\|_2^2 \\ &\text{such that} \quad y_i(w^\top x_i + b) \geq 1, \forall i \in [m] \end{aligned}$$

$$\begin{aligned} &\min_{w,b} \quad \frac{1}{2} \|w\|_2^2 \\ &\text{such that} \quad y_i(w^\top x_i + b) \geq 0, \forall i \in [m] \\ &\quad \min_{i \in [m]} |w^\top x_i + b| = 1 \end{aligned}$$



This is a convex QP! Can use existing solvers!

Homework: Work out why these two are equivalent!

RECAP - DUALITY

Primal:

$$\begin{array}{ll} \min_{w} & J(w) \\ \text{such that} & c_i(w) \leq 0, \forall i \in [m] \end{array}$$

Lagrangian:

$$\mathcal{L}(w, \alpha) = J(w) + \sum_{i=1}^m \alpha_i c_i(w)$$

Strong duality:

$$J^* = \min_w \max_{\alpha \geq 0} \mathcal{L}(w, \alpha) = \max_{\alpha \geq 0} \min_w \mathcal{L}(w, \alpha) = D^*$$

Dual

$$\begin{array}{ll} \max_{\alpha} & D(\alpha) \\ \text{such that} & \alpha_i \geq 0, \forall i \in [m] \end{array}$$

Lagrange (dual):

$$D(\alpha) = \min_w \mathcal{L}(w, \alpha)$$

KKT conditions for optimal w, α :

1. $\nabla_w \mathcal{L}(w, \alpha) = 0$
2. $\alpha_i c_i(w) = 0$ for all $i \in [m]$

OPTIMIZATION - DUAL

Primal:

$$\begin{array}{ll} \min_{w,b} & \frac{1}{2} \|w\|_2^2 \\ \text{such that} & 1 - y_i(w^\top x_i + b) \leq 0, \forall i \in [m] \end{array}$$

Lagrangian:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^m \alpha_i (1 - y_i(w^\top x_i + b))$$

Dual

$$\begin{array}{ll} \max_{\alpha} & D(\alpha) \\ \text{such that} & \alpha_i \leq 0, \forall i \in [m] \end{array}$$

Lagrange (dual):

$$D(\alpha) = \min_{w,b} \mathcal{L}(w, b, \alpha)$$

Convex QP satisfies strong duality/ KKT conditions

OPTIMIZATION - DUAL

Lagrangian:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^m \alpha_i (1 - y_i (w^\top x_i + b))$$

Lagrange (dual):

$$D(\alpha) = \min_{w, b} \mathcal{L}(w, b, \alpha) = \min_{w, b} \left(\frac{1}{2} \|w\|_2^2 + \sum_{i=1}^m \alpha_i (1 - y_i (w^\top x_i + b)) \right)$$

Unconstrained convex optimization problem so we can minimize by setting gradient to 0

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^m \alpha_i y_i x_i = 0 \implies w = \sum_{i=1}^m \alpha_i y_i x_i$$

$$\nabla_b \mathcal{L}(w, b, \alpha) = - \sum_{i=1}^m \alpha_i y_i = 0 \implies \sum_{i=1}^m \alpha_i y_i = 0.$$

OPTIMIZATION - DUAL

$$\begin{array}{ll} \max_{\alpha} & D(\alpha) \\ \text{such that} & \alpha_i \geq 0, \forall i \in [m] \end{array} \quad \longrightarrow \quad \begin{array}{ll} \max_{\alpha} & -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (x_i^\top x_j) + \sum_{i=1}^m \alpha_i \\ \text{such that} & \sum_{i=1}^m \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \forall i \in [m] \end{array}$$

$$\text{Solve for } \alpha \implies w = \sum_{i=1}^m \alpha_i y_i x_i$$

OPTIMIZATION - SUPPORT VECTORS

Complementary slackness conditions for optimal w, b, α :

$$\alpha_i(1 - y_i(w^\top x_i + b)) = 0 \text{ for all } i \in [m]$$

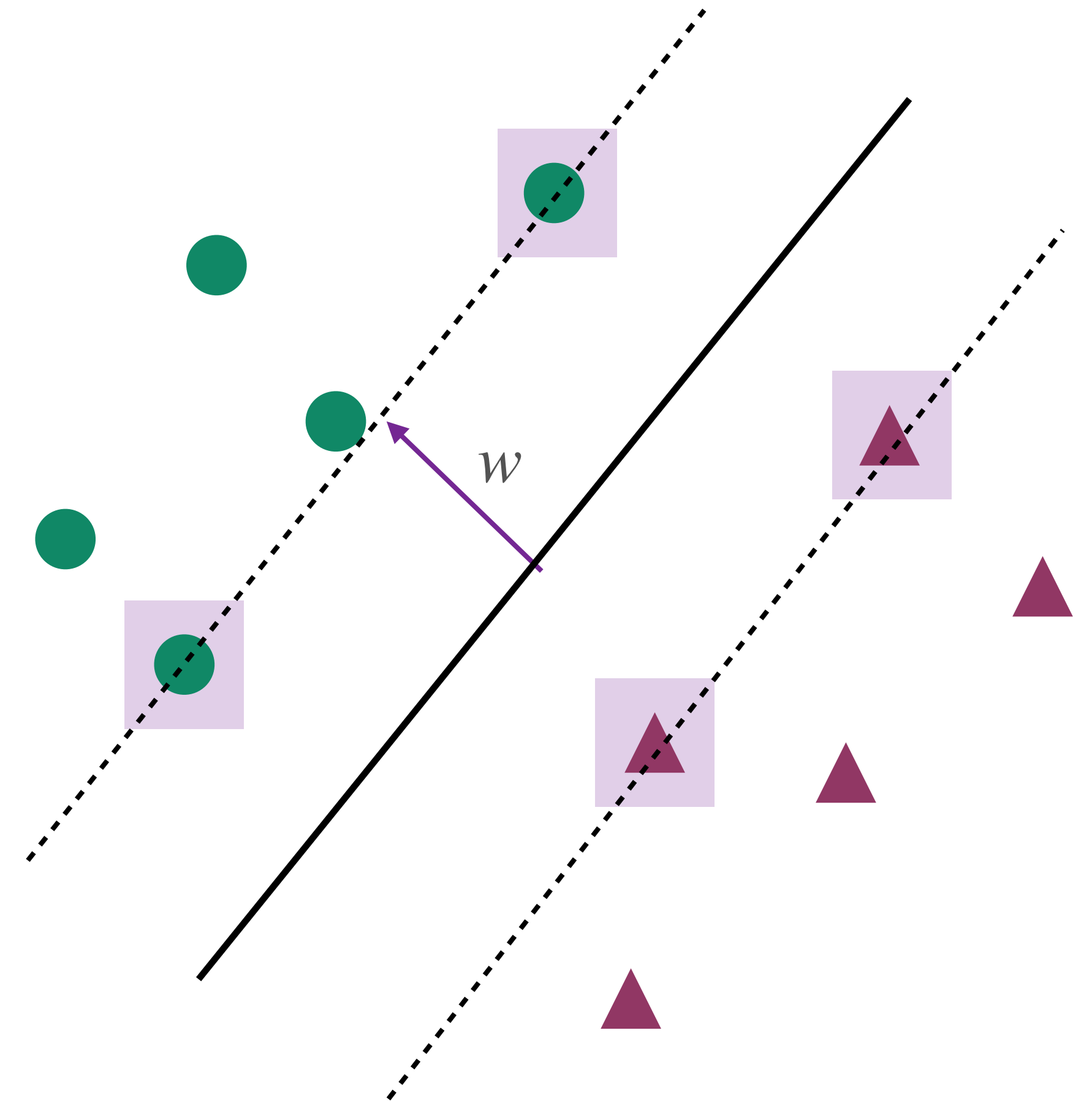
either $\alpha_i = 0$ or $y_i(w^\top x_i + b) = 1$

Support vectors:

$$SV = \{i \in [m] : \alpha_i > 0\}$$

$$w = \sum_{i=1}^m \alpha_i y_i x_i \implies w = \sum_{i \in SV} \alpha_i y_i x_i$$

$$b = y_i - w^\top x_i \text{ for any } i \in SV$$



SVM - PRIMAL & DUAL

Primal

$$\min_{w,b} \quad \frac{1}{2} \|w\|_2^2$$

such that

$$y_i(w^\top x_i + b) \geq 1, \forall i \in [m]$$

d + 1 variables

Dual

$$\max_{\alpha} \quad -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (x_i^\top x_j) + \sum_{i=1}^m \alpha_i$$

such that

$$\sum_{i=1}^m \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \forall i \in [m]$$

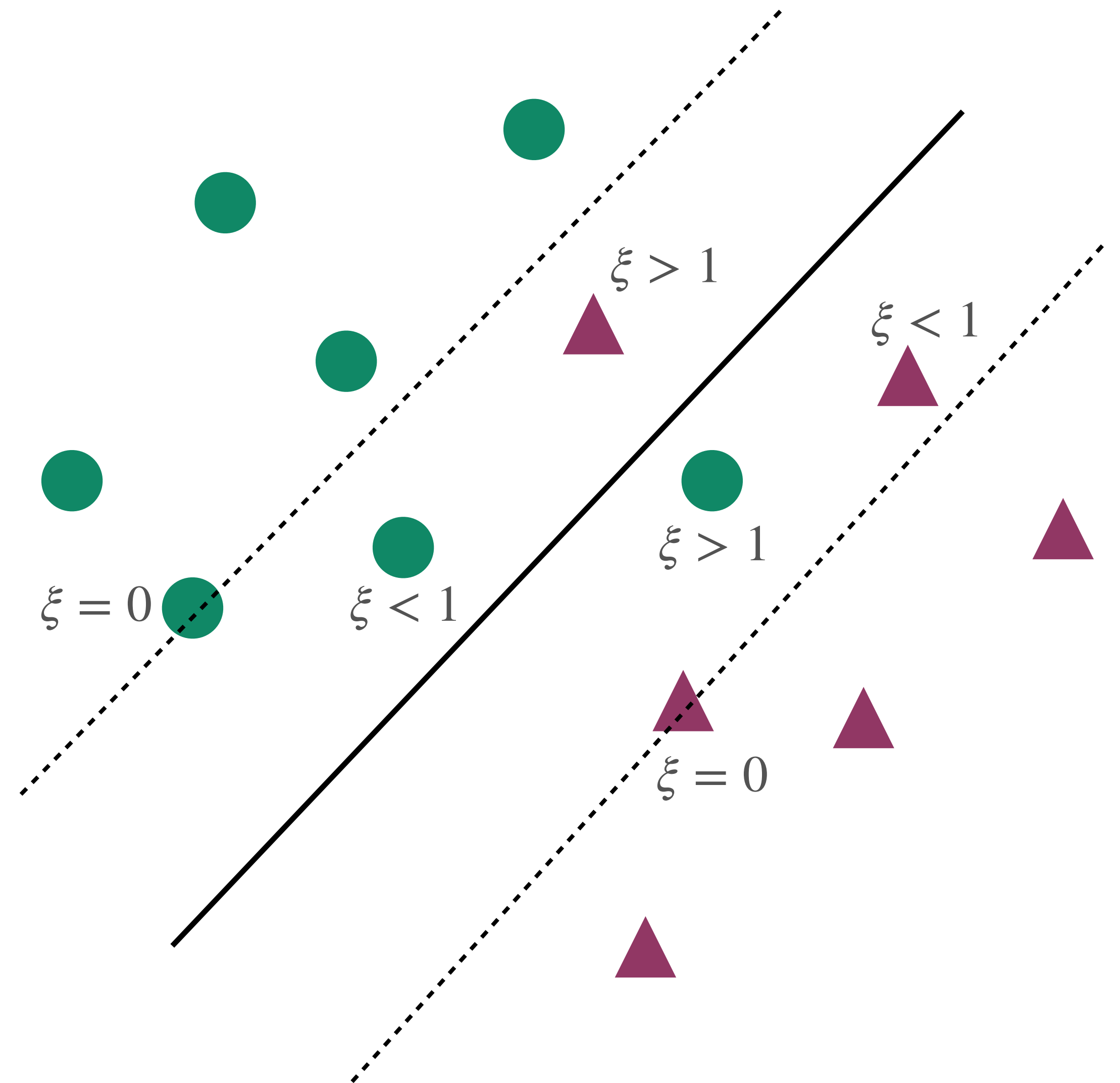
m variables

DATA - NON-SEPARABLE

$$\begin{array}{ll} \min_{w,b} & \frac{1}{2} \|w\|_2^2 \\ \text{such that} & y_i(w^\top x_i + b) \geq 1, \forall i \in [m] \end{array}$$

$$\begin{array}{ll} \min_{w,b} & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^m \xi_i \\ \text{such that} & y_i(w^\top x_i + b) \geq 1 - \xi_i, \forall i \in [m] \\ & \xi_i \geq 0, \forall i \in [m] \end{array}$$

Slack



SOFT-SVM - PRIMAL & DUAL

Primal

$$\min_{w,b,\xi_i} \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^m \xi_i$$

such that

$$y_i(w^\top x_i + b) \geq 1 - \xi_i, \forall i \in [m]$$
$$\xi_i \geq 0, \forall i \in [m]$$

$d + m + 1$ variables

Dual

$$\max_{\alpha} \quad -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (x_i^\top x_j) + \sum_{i=1}^m \alpha_i$$

such that

$$\sum_{i=1}^m \alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C, \forall i \in [m]$$

m variables

SOFT-SVM - LOSS MINIMIZATION VIEW

$$\begin{aligned} \min_{w,b,\xi_i} \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^m \xi_i \\ \text{such that} \quad & y_i(w^\top x_i + b) \geq 1 - \xi_i, \forall i \in [m] \\ & \xi_i \geq 0, \forall i \in [m] \end{aligned}$$

Is equivalent to the following loss minimization problem for $C = \frac{1}{2\lambda m}$:

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

ℓ_2 -regularized hinge loss minimization

