7 February 2023

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: $f : \mathcal{X} \to \mathcal{Y}, f \in \mathcal{F}$

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

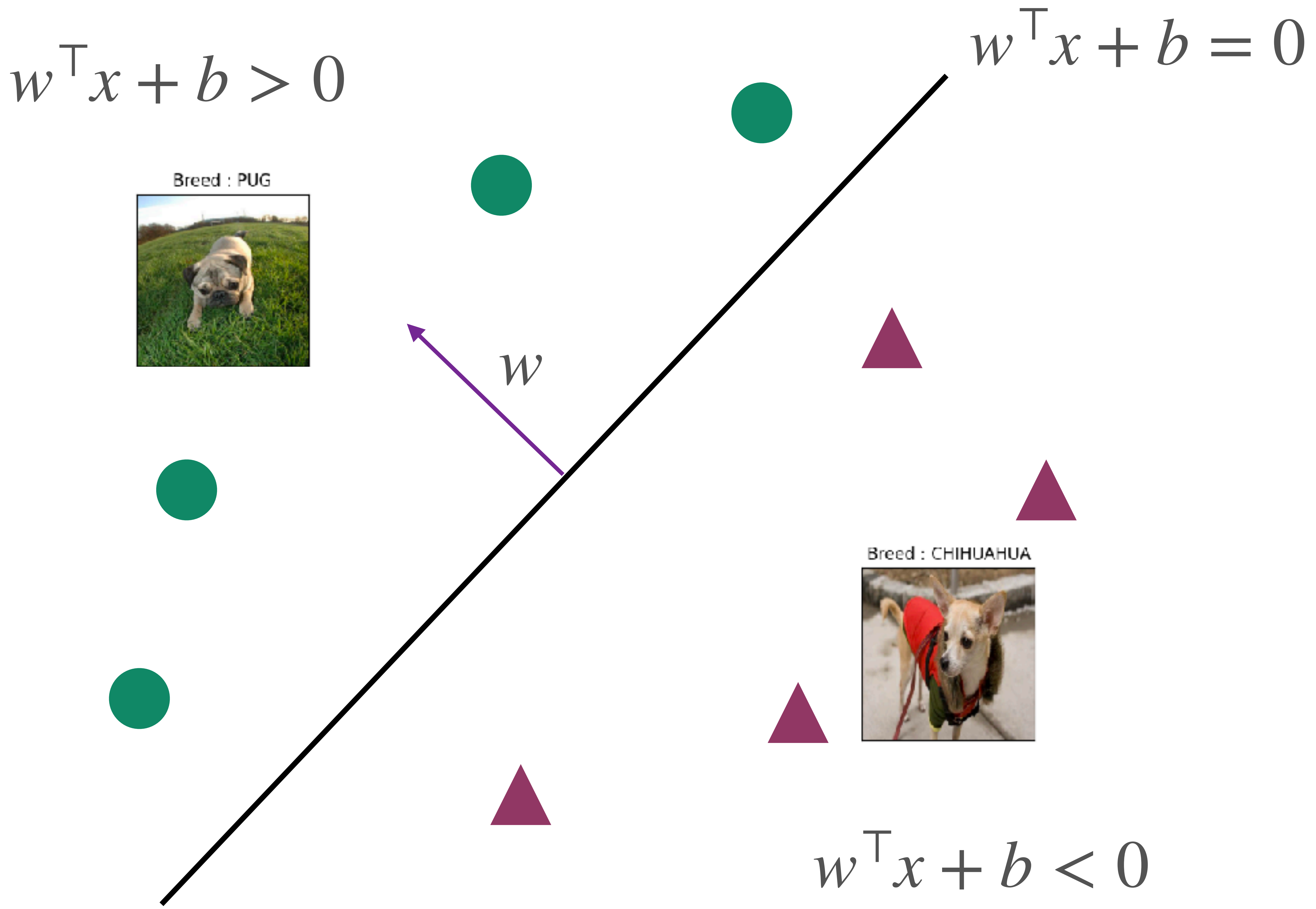
Data: $\{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from distribution $\mathcal{D}$

HYPOTHESIS CLASS - LINEAR CLASSIFIER

We will keep the bias

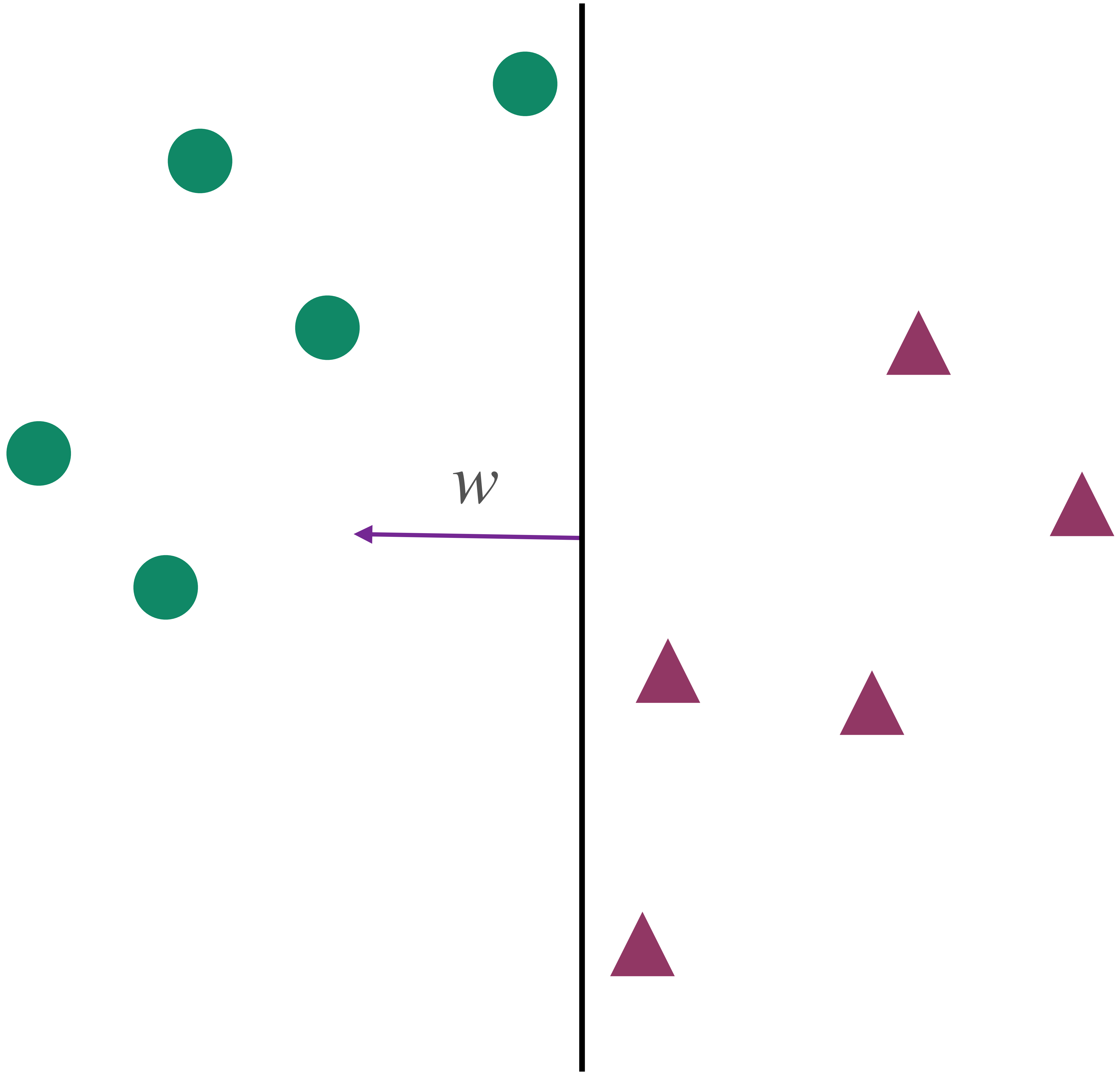
Linear Classifier: $\mathcal{F} := \{x \mapsto \mathrm{sign}(w^\top x + b) \,|\, w \in \mathbb{R}^d, b \in \mathbb{R}\}$

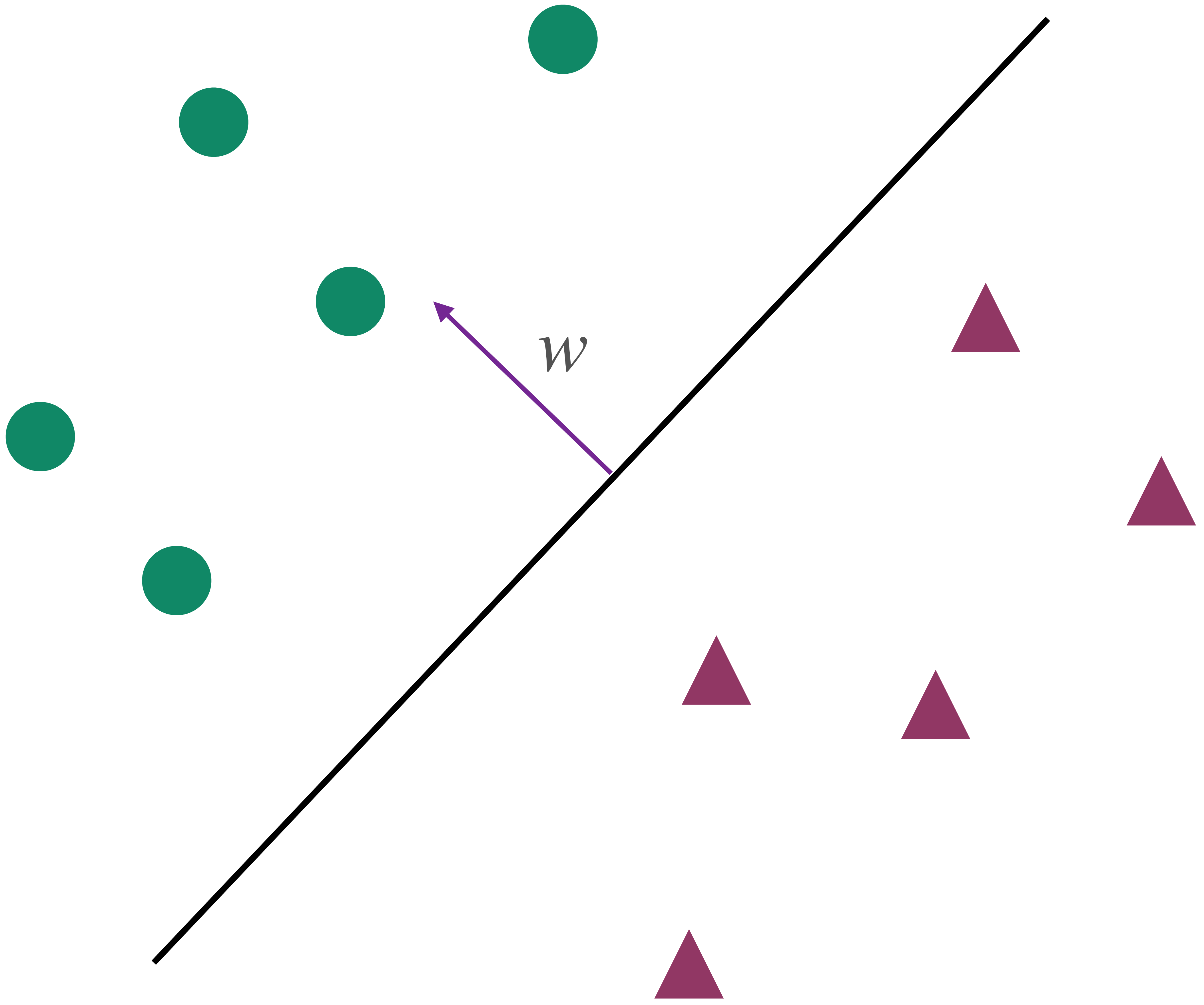
$$\mathrm{sign}(a) = \begin{cases} +1 & \text{if } a \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$



BEST SEPARATING HYPERPLANE - MAX-MARGIN

Which hyperplane is better?



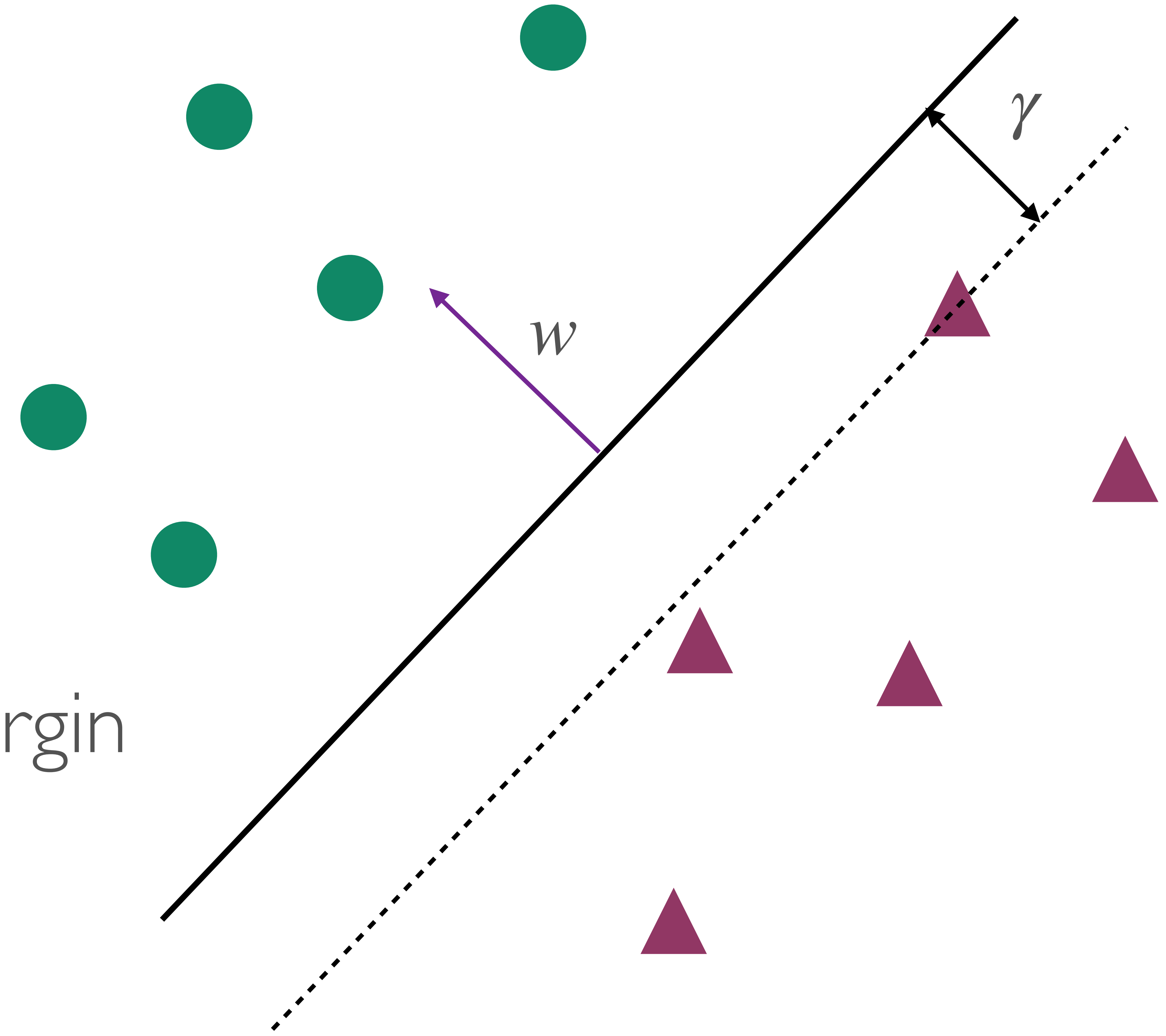


BEST SEPARATING HYPERPLANE - MAX-MARGIN

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

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

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

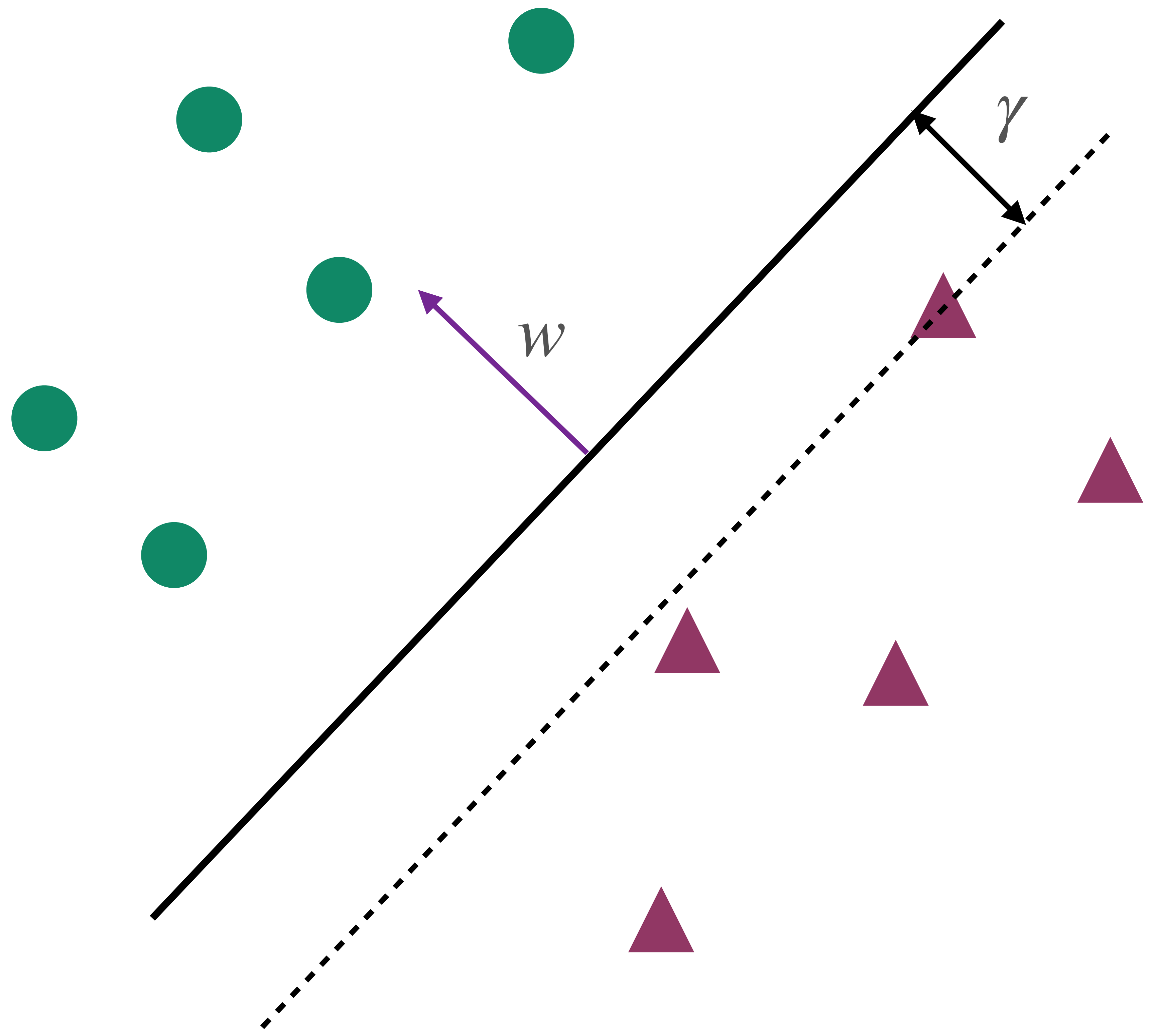
$$\max_{w,b} \quad \underbrace{\gamma(w,b)}_{\text{margin}}$$

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

Substituting for margin:

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

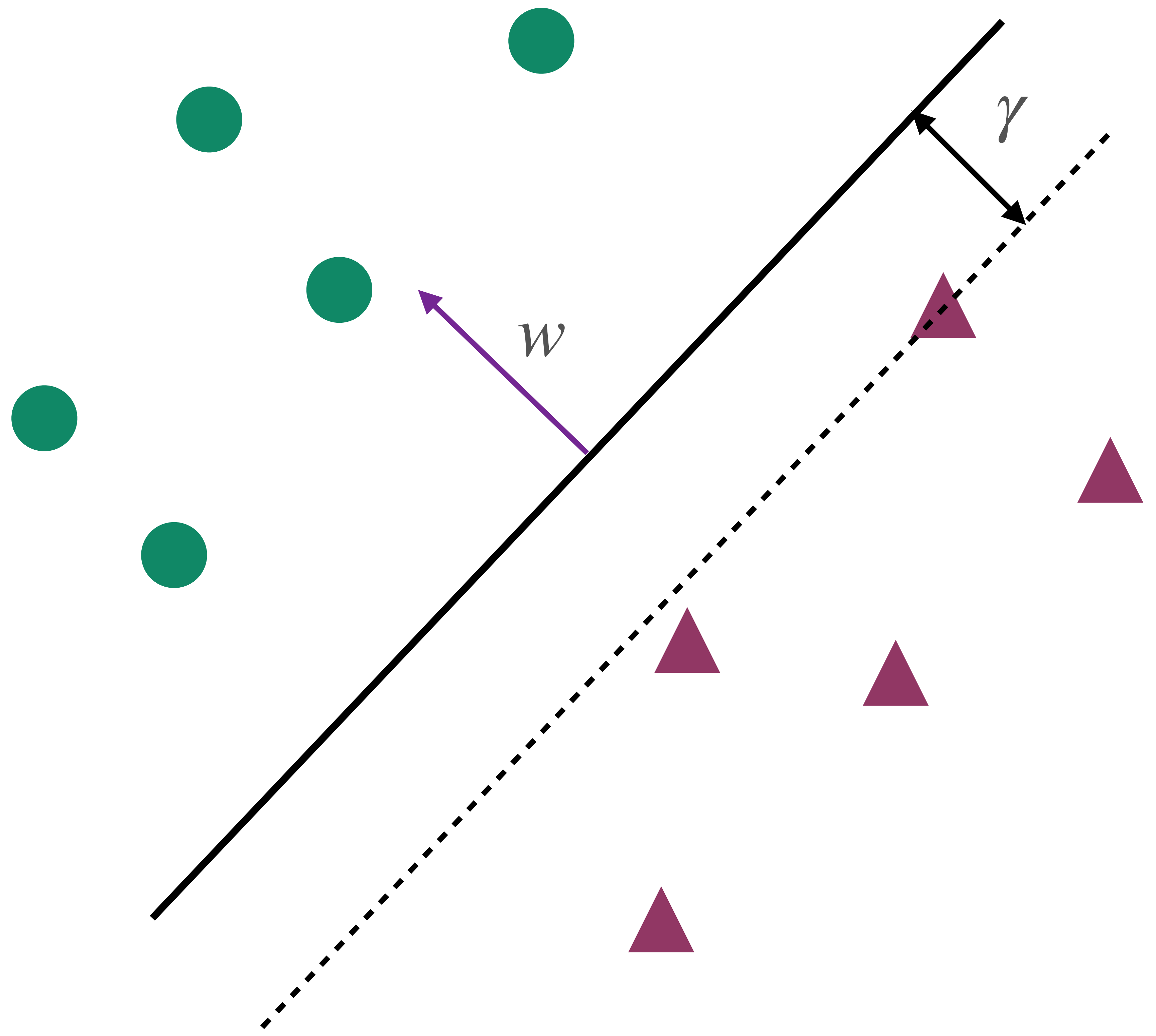
$$\text{such that} \quad \underbrace{y_i(w^\top x_i + b) \geq 0, \forall i \in [m]}_{(w,b)\text{ linearly separates data}}$$



OPTIMIZATION PROBLEM - MAX-MARGIN

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

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

Puts a constraint on w, b

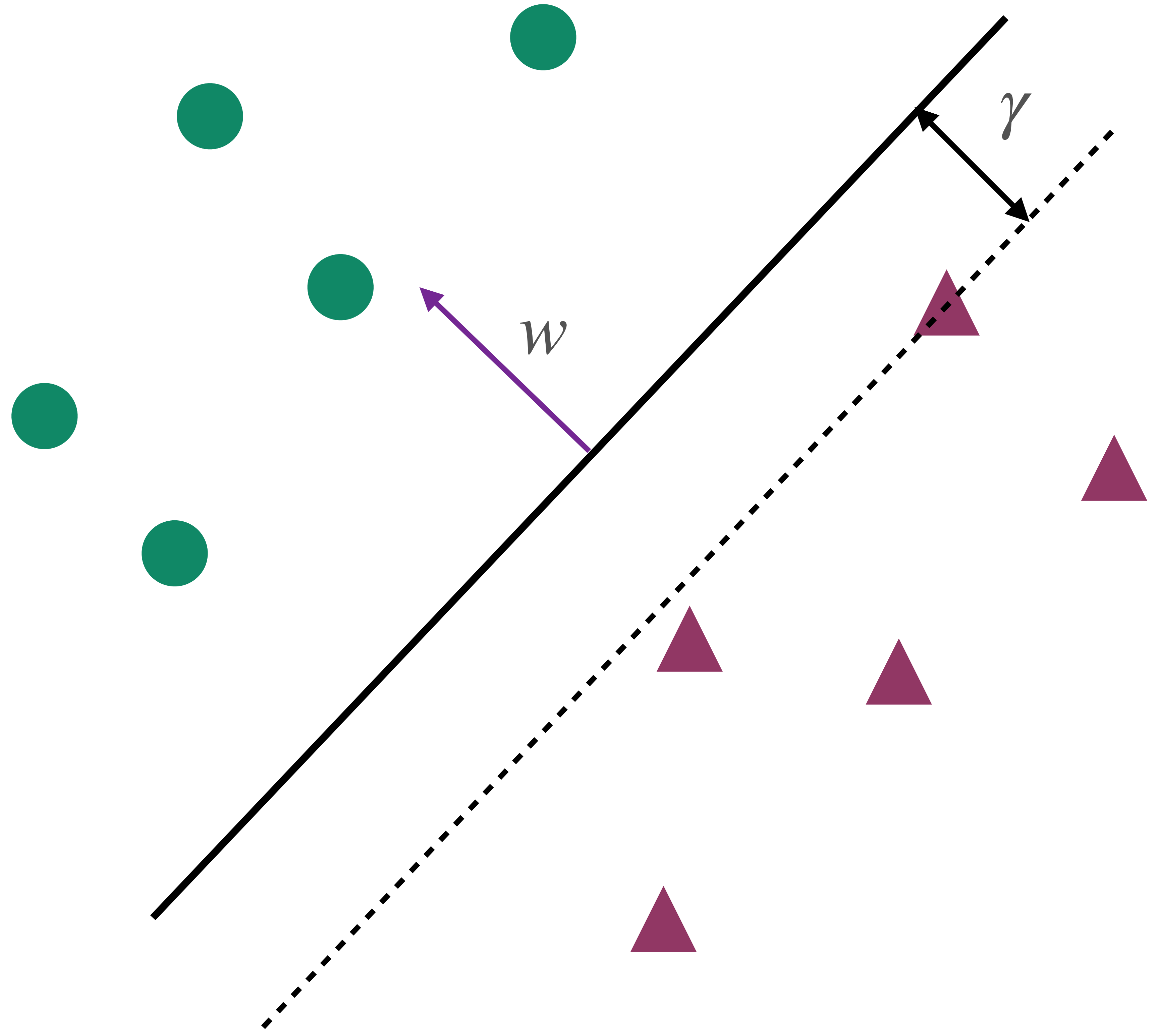
OPTIMIZATION PROBLEM - FIXED SCALE

Adding scale constraint:

$$\cancel{\max_{w,b} \underbrace{\frac{1}{\|w\|_2}}_{\text{margin}}} \qquad \min_{w,b} \quad \frac{1}{2}\|w\|_2^2$$

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

$$\underbrace{\min_{i \in [m]} |w^\top x_i + b| = 1}_{\text{fixed scale}}$$



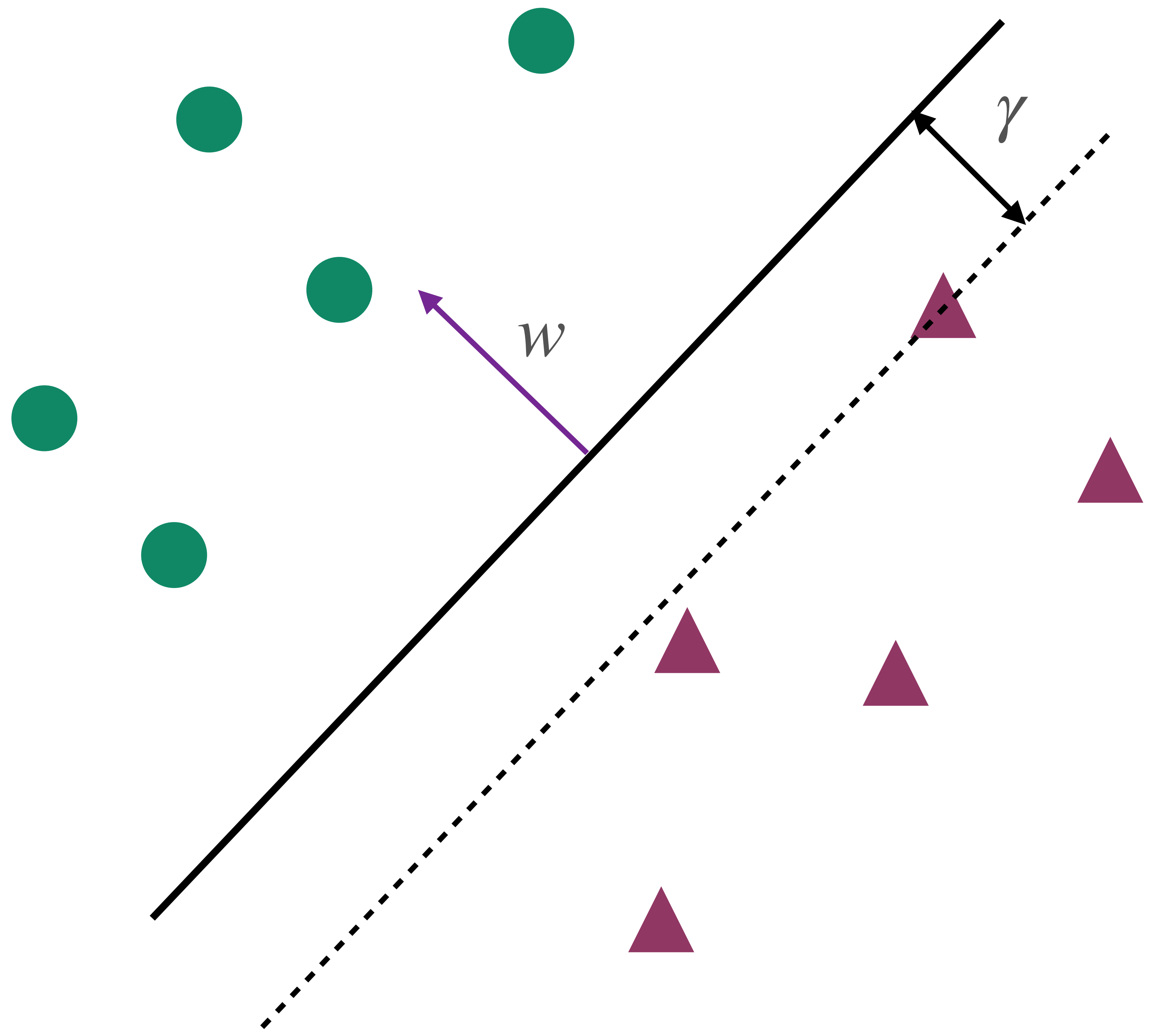
What kind of optimization problem is this?

OPTIMIZATION PROBLEM - QUADRATIC PROGRAM

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

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

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



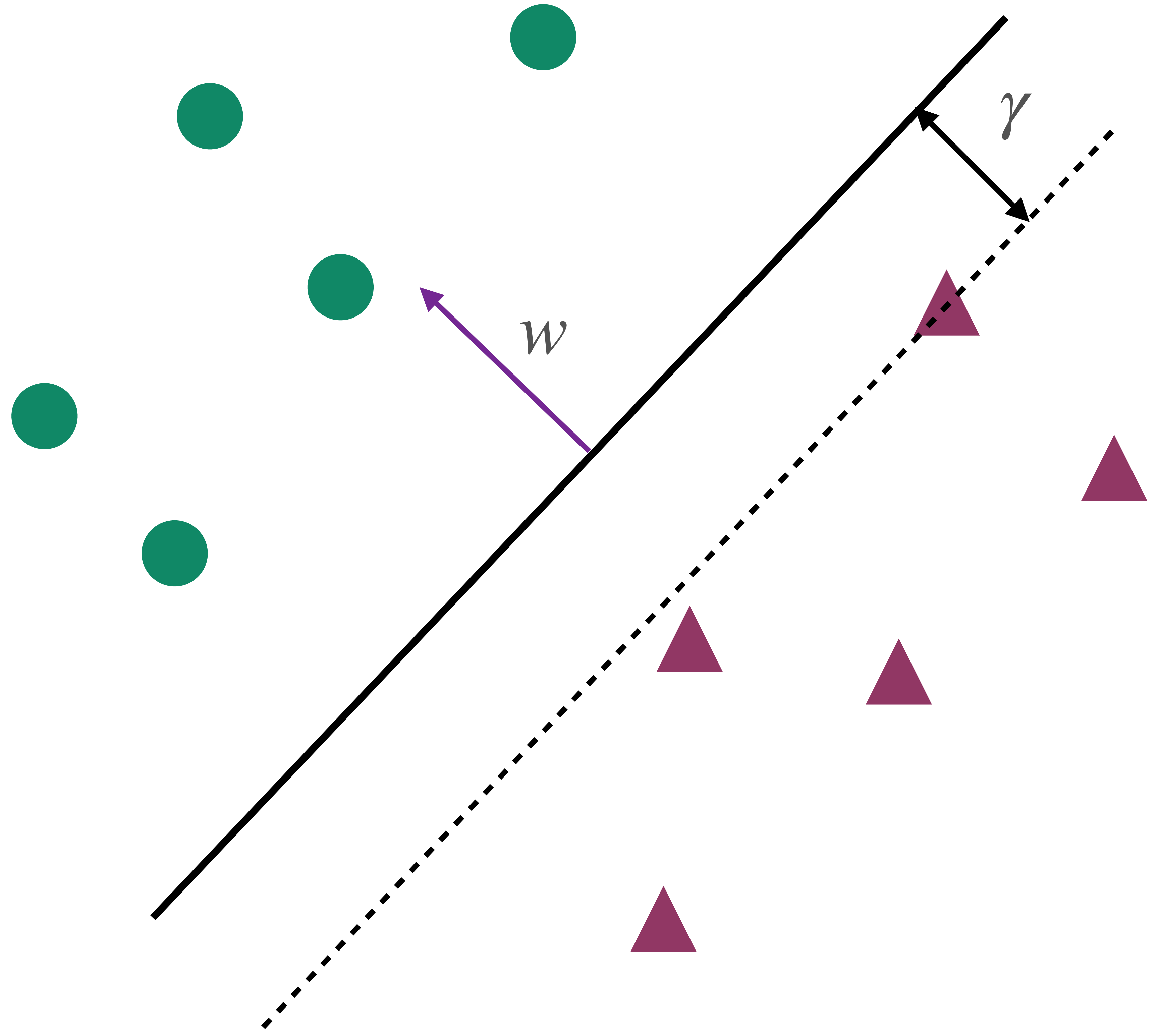
Is this a convex problem?

OPTIMIZATION PROBLEM - CONVEX TRICK

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

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

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



This is a convex QP! Can use existing solvers!

Homework: Work out why these two are equivalent!

RECAP - DUALITY

Primal:

$$\min_{w} \quad J(w)$$

$$\text{such that} \quad c_i(z) \leq 0, \forall i \in [m]$$

Dual

$$\max_{\alpha} \quad D(\alpha)$$

$$\text{such that} \quad \alpha_i \leq 0, \forall i \in [m]$$

Lagrangian:

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

Lagrange (dual):

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

Strong duality:

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

KKT conditions for optimal w, α:

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

OPTIMIZATION - DUAL

Primal:

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

$$\text{such that} \quad 1 - y_i(w^\top x_i + b) \leq 0, \forall i \in [m]$$

Dual

$$\max_{\alpha} \quad D(\alpha)$$

$$\text{such that} \quad \alpha_i \leq 0, \forall i \in [m]$$

Lagrangian:

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

Lagrange (dual):

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

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

$$\max_{\alpha} \quad D(\alpha)$$

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

$\Rightarrow$

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

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

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

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:

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

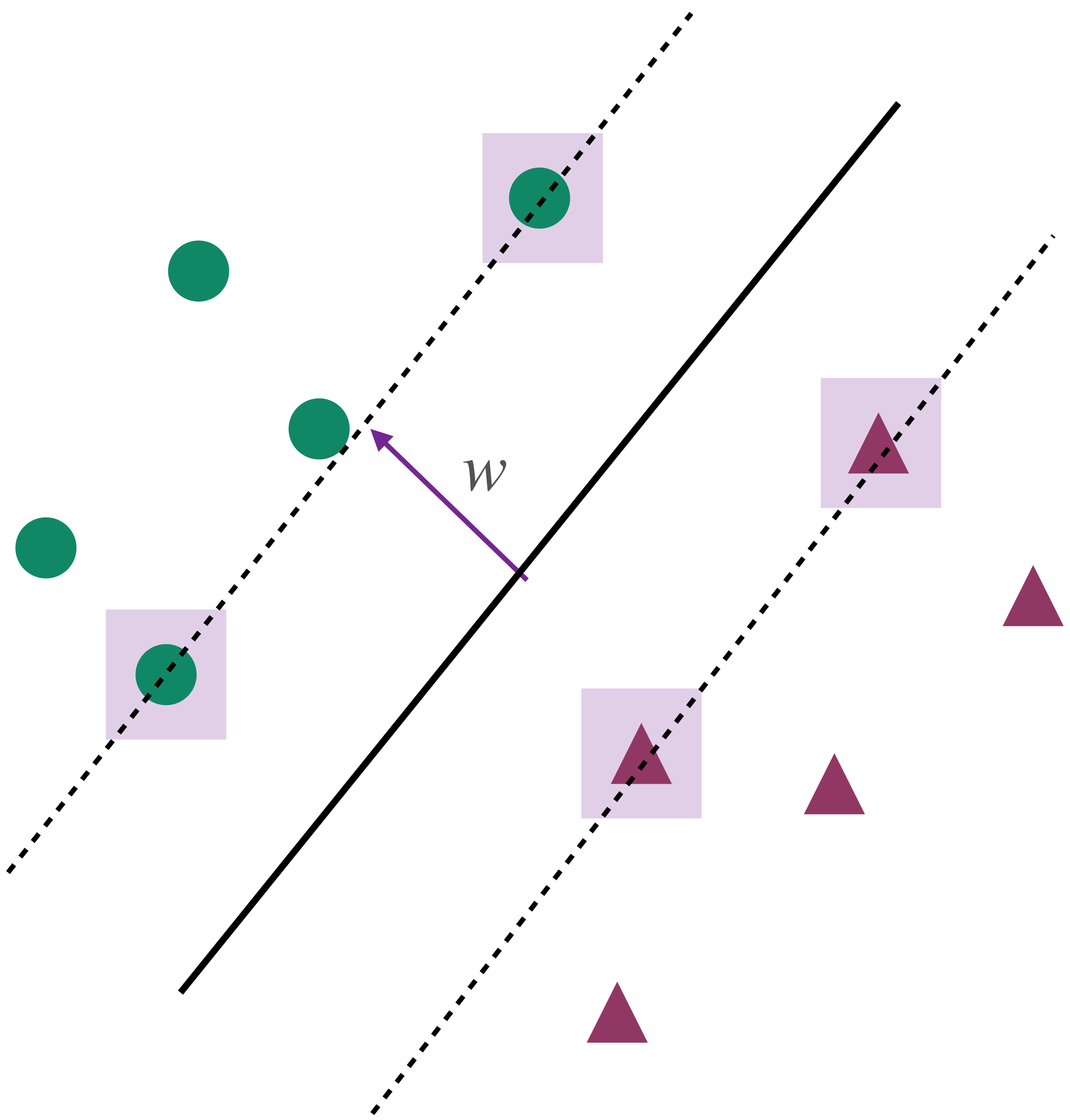
either $\alpha_i = 0$ or $y_i(w^\top x_i + b) = 0$

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

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

$$\text{such that} \quad y_i(w^\top x_i + b) \geq 1, \forall i \in [m]$$

Dual

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

$$\text{such that} \quad \sum_{i=1}^{m}\alpha_i y_i = 0$$

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

$d+1$ *variables*

m *variables*

DATA - NON-SEPARABLE

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

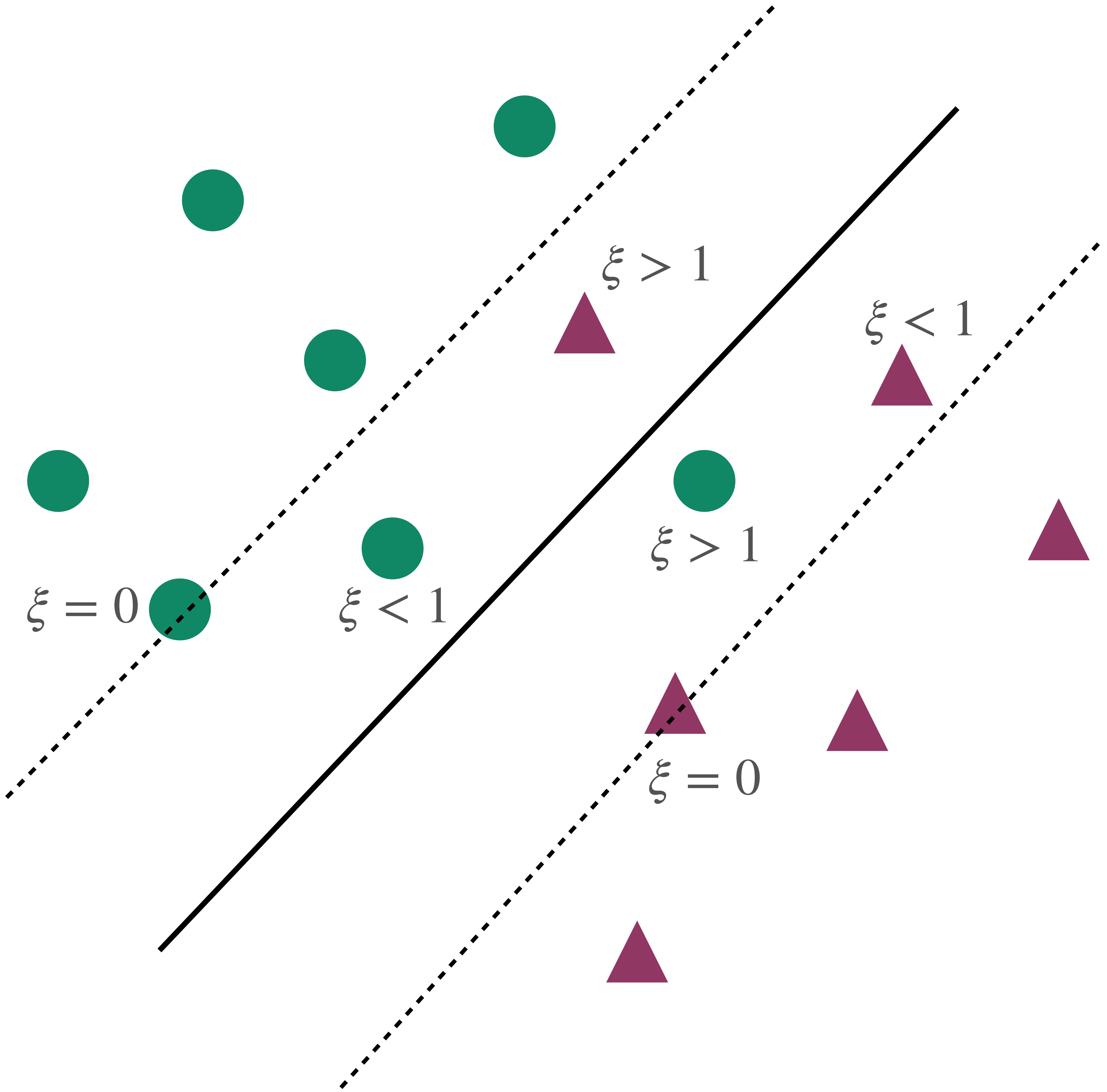
such that $\quad y_i(w^\top x_i + b) \geq 1, \forall i \in [m]$

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

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

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

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

$\quad \xi_i \geq 0, \forall i \in [m]$

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

$\quad 0 \leq \alpha_i \leq C, \forall i \in [m]$

$d + m + 1$ *variables*

m *variables*

SOFT-SVM - LOSS MINIMIZATION VIEW

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

$$\text{such that} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i, \forall i \in [m]$$

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

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

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

ℓ_2-regularized hinge loss minimization