$$\min_x f(x)$$

**First-order necessary condition**

$$-\frac{\partial f(x)}{\partial x} = \mathbf{0}$$

$$\min_x f(x)$$
$$\text{subj to } g(x) = b$$

$$-\frac{\partial f(x)}{\partial x} = \lambda \frac{\partial g(x)}{\partial x}$$

$$\min_x f(x)$$
$$\text{subj to } g(x) \geq b$$

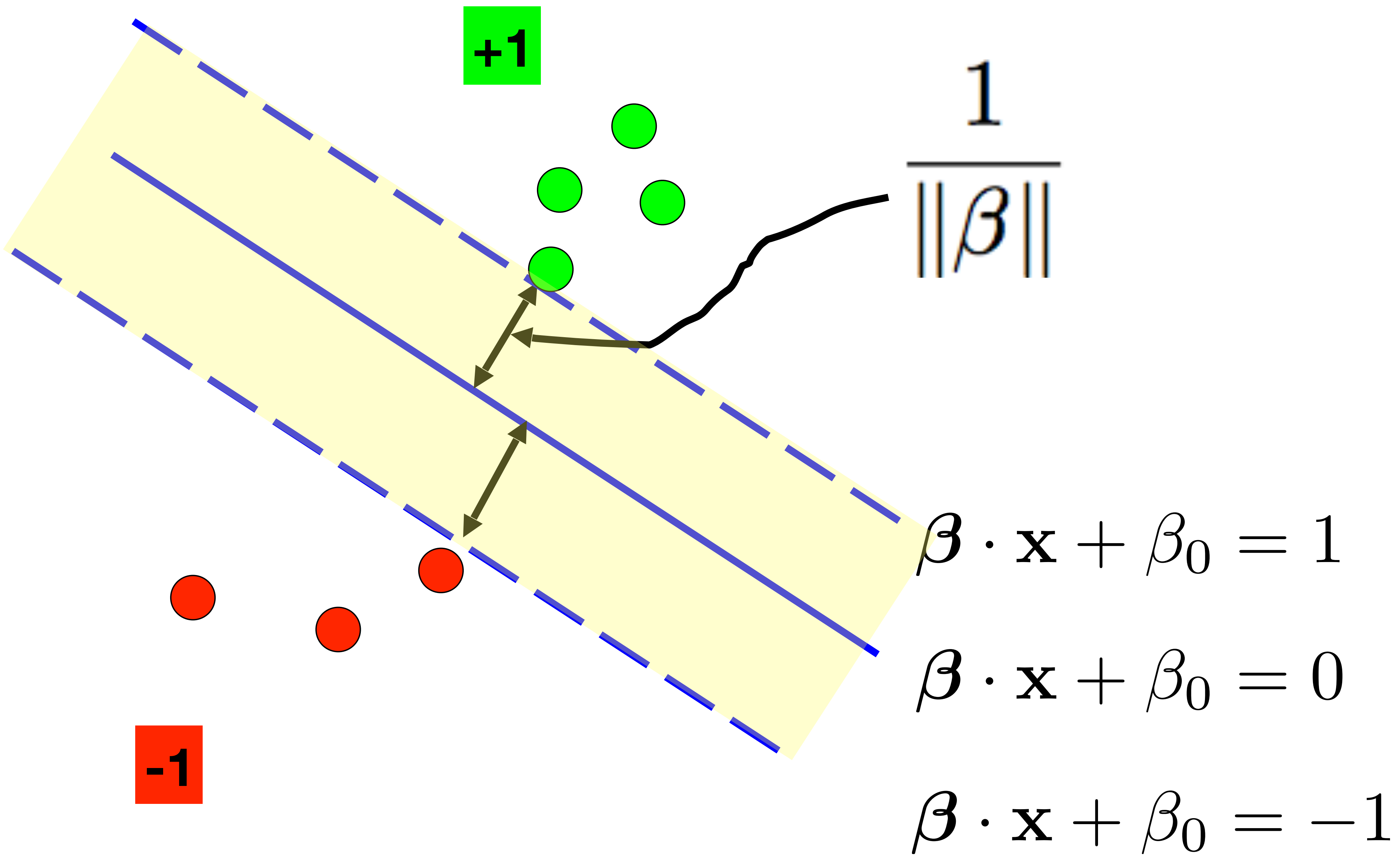
$$-\frac{\partial f(x)}{\partial x} = -\lambda \frac{\partial g(x)}{\partial x}$$
$$\lambda \geq 0$$
$$g(x) - b \geq 0$$
$$\lambda(g(x) - b) = 0$$

$$\text{Define } L(x, \lambda) = f(x) - \lambda\big(g(x) - b\big)$$

$$\boxed{\frac{\partial}{\partial x} L = 0}$$

If x is a local optimum for the constrained optimization, then it must satisfy the **KKT conditions**.

- x is active (lambda >= 0)
- x is inactive (lambda = 0)




- Convex quadratic optimization problem with affine constraints.
- Any local optimum is a global optimum.
- **KKT conditions** are sufficient and necessary
- Equivalence between **the Primal** and **the Dual**.

## Max-Margin Problem

$$\min_{\boldsymbol{\beta},\beta_0} \quad \frac{1}{2}||\boldsymbol{\beta}||^2 \tag{1}$$
$$\text{subject to} \quad y_i(\boldsymbol{\beta} \cdot \mathbf{x}_i + \beta_0) - 1 \geq 0,$$

where $\boldsymbol{\beta} \cdot \mathbf{x}_i = \boldsymbol{\beta}^t \mathbf{x}_i$ denotes the (Euclidian) inner product between two vectors. The constraints are imposed to make sure that the points are on the correct side of the dashed lines, i.e.,

$$\boldsymbol{\beta} \cdot \mathbf{x}_i + \beta_0 \geq +1 \quad \text{for } y_i = +1,$$
$$\boldsymbol{\beta} \cdot \mathbf{x}_i + \beta_0 \leq -1 \quad \text{for } y_i = -1.$$

### Primal

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|^2$$

$$\text{subj to } y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) - 1 \geq 0,$$
$$i = 1, \ldots, n$$

### Support Vectors

Data points with non-zero Lagrange multiplier lambda_i, i.e, data points on the dashed lines.

### Dual

$$\max_{\lambda_{1:n}} \sum \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subj to } \sum \lambda_i y_i = 0,$$
$$\lambda_i \geq 0$$

### KKT conditions

$$\sum_i \lambda_i y_i \mathbf{x}_i = \boldsymbol{\beta}$$

$$\sum_i \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

$$y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) - 1 \geq 0$$

$$\lambda_i \left[ y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) - 1 \right] = 0$$

### Why work with Dual?

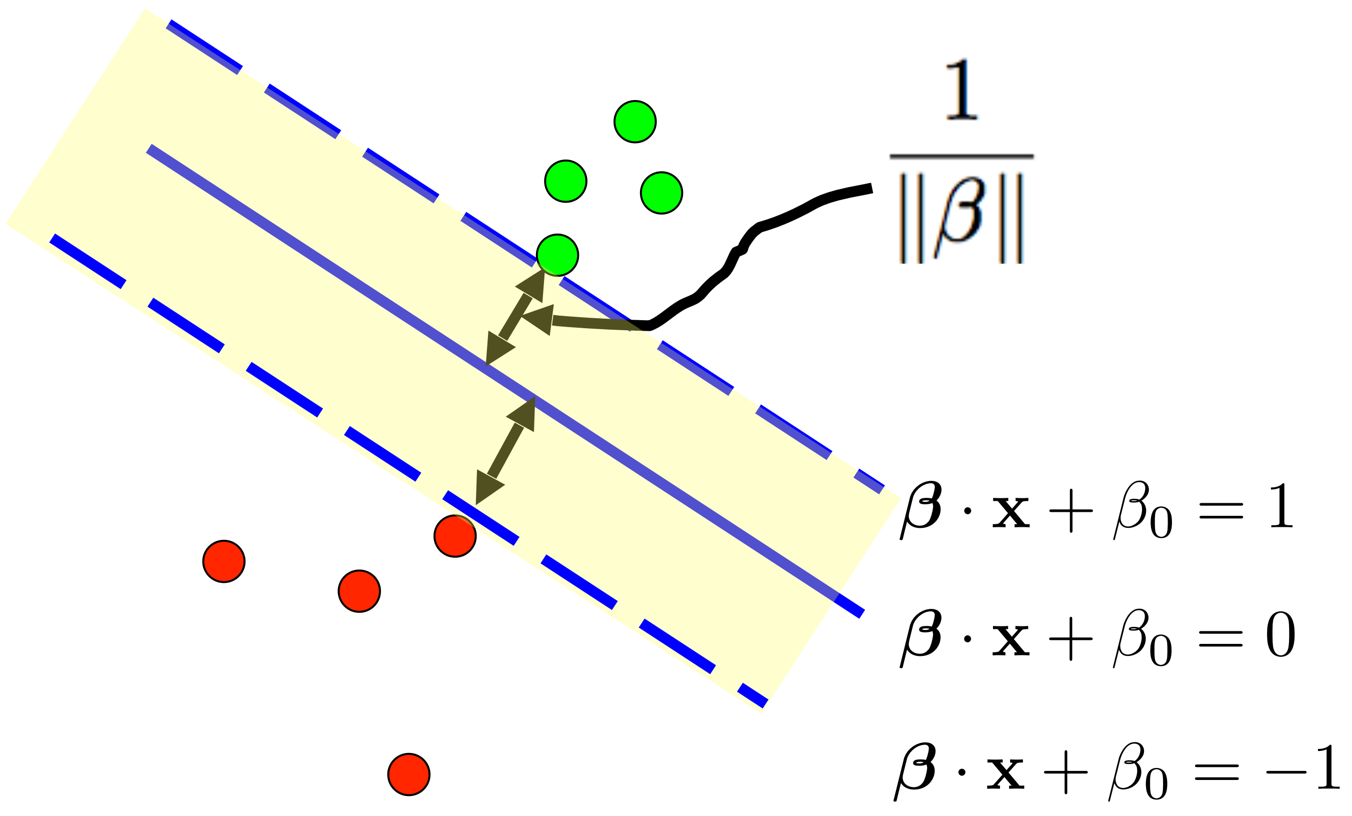
1. Easier to solve
2. Many lambda_i's are zero
3. Leads to kernel trick

### Lagrange function

$$L(\boldsymbol{\beta}, \beta_0, \lambda_{1:n})$$
$$= \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_i \lambda_i \left[ y_i(\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) - 1 \right]$$
$$= \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_i \lambda_i y_i(\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) + \sum_i \lambda_i$$

**Hard Margin**

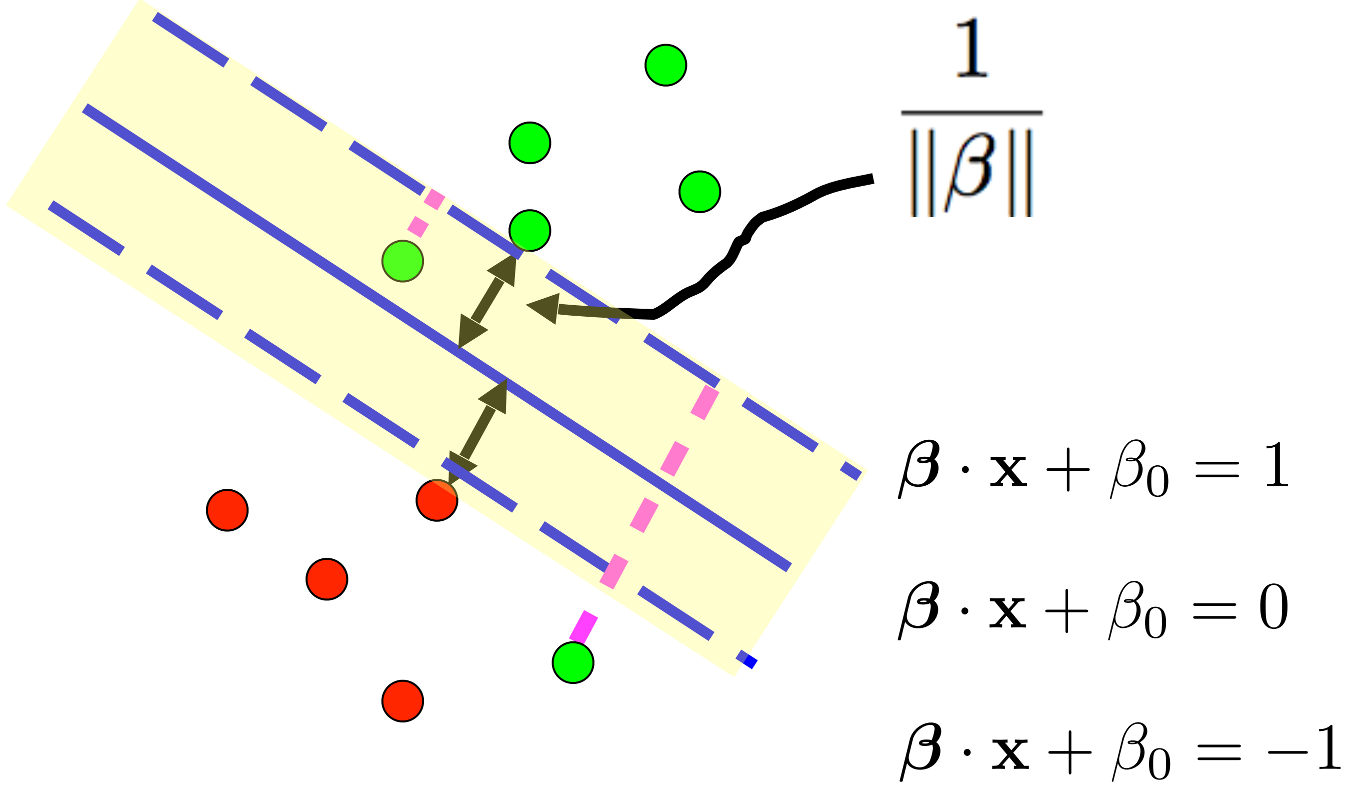
Linear SVM for Separable Data




1. Formulate the **Primal** Problem (dim = p+1)
2. Solve the **Dual** Problem (dim = n)
3. **KKT Conditions** link the two sets of solutions
4. **SV**: data points on the dashed line

**Soft Margin**

Linear SVM for Non-separable/Separable Data

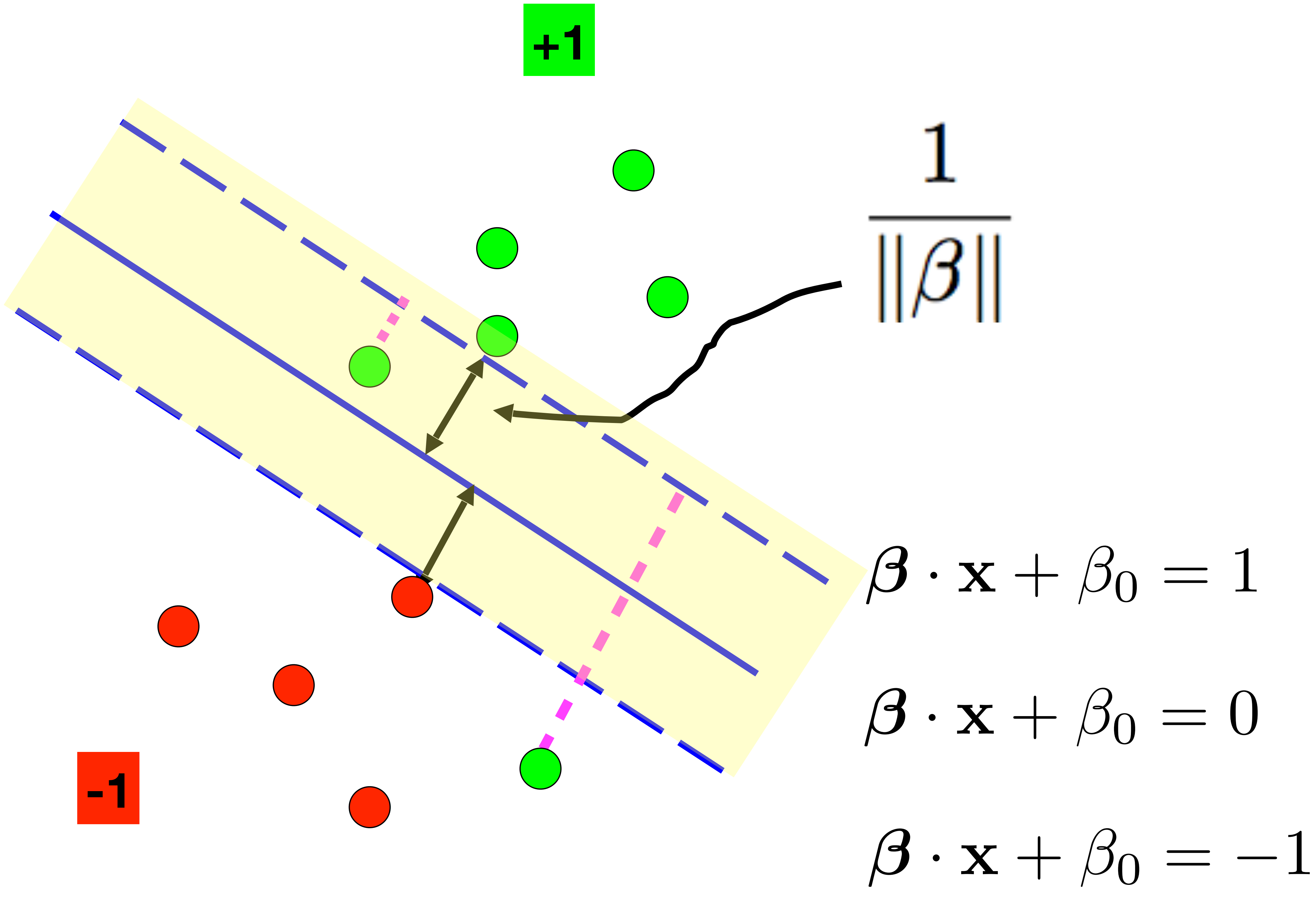



**Kernel Machine**

Nonlinear SVM for Separable/Non-separable Data

**Some Practical Issues**

1. Binary decision to probability
2. Multiclass SVM

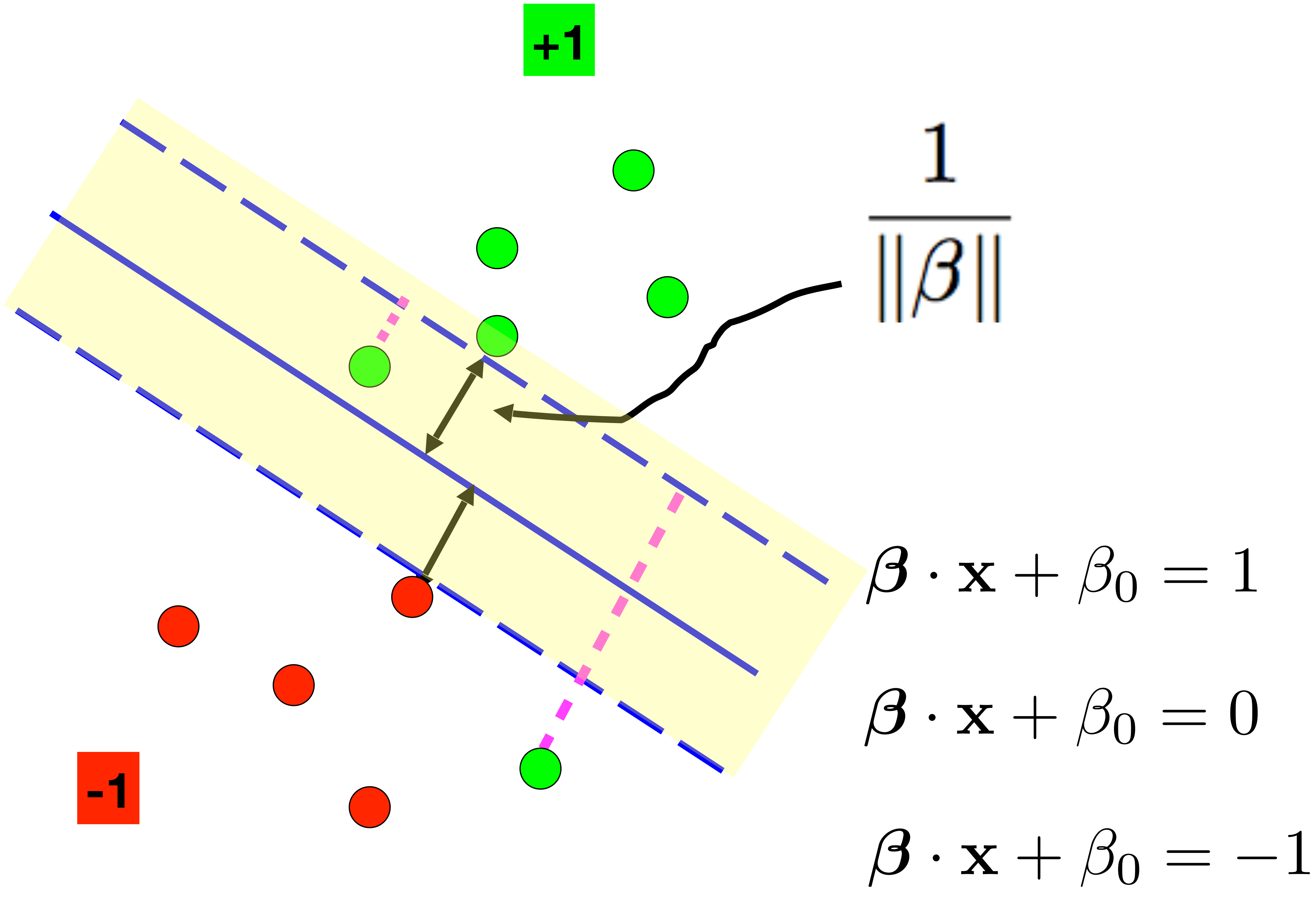



## Max-Margin with Slack Variables

Then we introduce a slack variable $\xi_i$ for each sample, and formulate the max-margin problem as follows

$$\begin{aligned} \min_{\boldsymbol{\beta}, \beta_0} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum \xi_i \qquad (3) \\ \text{subject to} \quad & y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) - 1 + \xi_i \geq 0, \\ & \xi_i \geq 0. \end{aligned}$$

Note that $\xi_i > 0$ only for samples that are on the wrong side of the dashed line, and $\xi_i$ is automatically (by the optimization) set to be 0 for samples that are on the correct side of the dashed line.




- Convex quadratic optimization problem with affine constraints (2n constraints).
- Any local optimum is a global optimum.
- **KKT conditions** are sufficient and necessary
- Equivalence between **the Primal** and **the Dual**.

## Max-Margin with Slack Variables

Then we introduce a slack variable $\xi_i$ for each sample, and formulate the max-margin problem as follows

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum \xi_i \qquad (3) \\ \text{subject to} \quad & y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) - 1 + \xi_i \geq 0, \\ & \xi_i \geq 0. \end{aligned}$$

Note that $\xi_i > 0$ only for samples that are on the wrong side of the dashed line, and $\xi_i$ is automatically (by the optimization) set to be 0 for samples that are on the correct side of the dashed line.

## Primal

$$\min_{\boldsymbol{\beta},\beta_0,\xi_{1:n}} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum \xi_i$$

$$\text{subj to } y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i,$$
$$\xi_i \geq 0$$

## Dual

$$\max_{\lambda_{1:n}} \quad \sum \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subj to } \sum \lambda_i y_i = 0, \;\; \gamma \geq \lambda_i \geq 0$$

Support Vectors

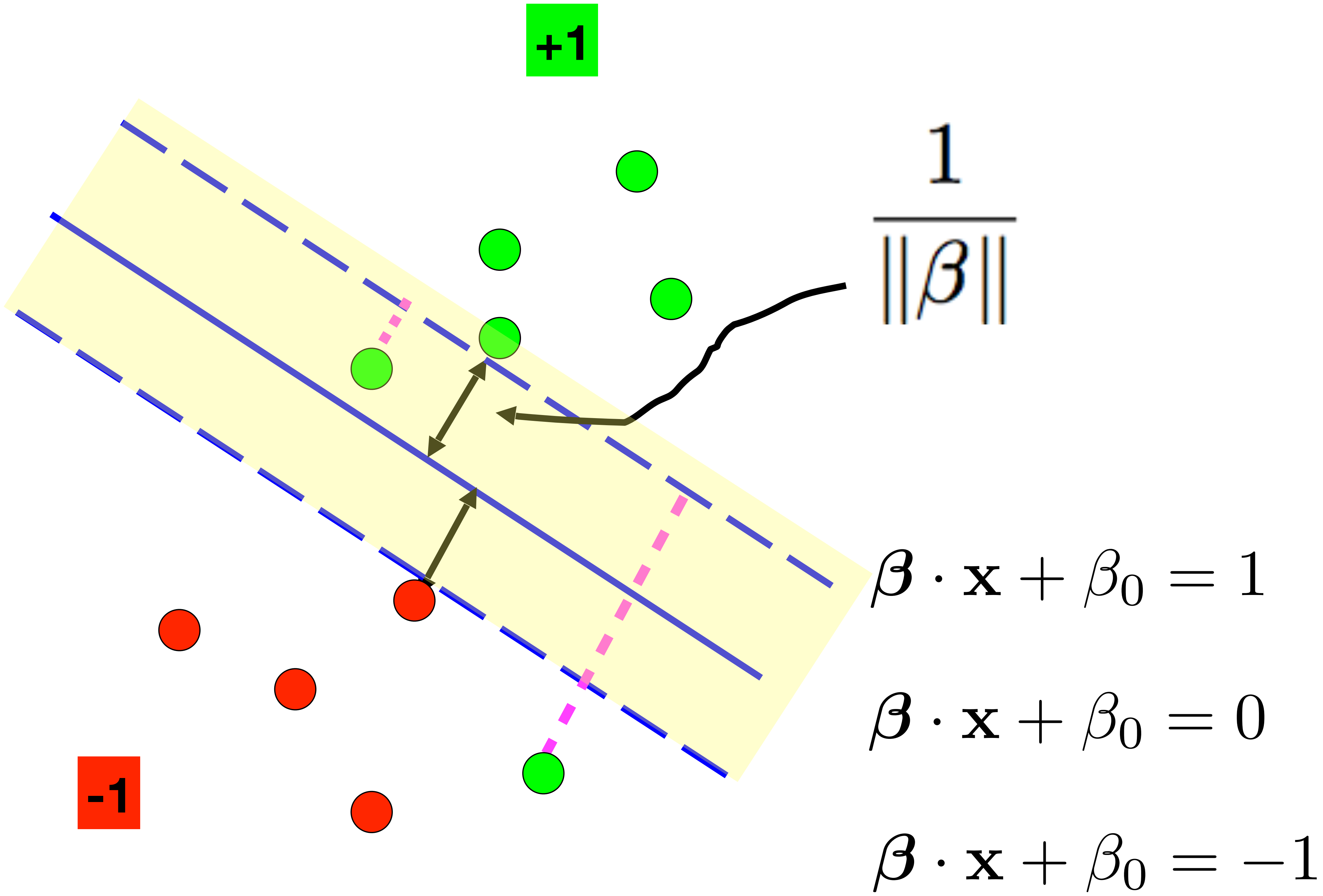
$$\{i : \lambda_i > 0\}$$

## KKT conditions

$$\sum \lambda_i y_i \mathbf{x}_i = \boldsymbol{\beta}$$
$$\sum \lambda_i y_i = 0$$
$$(\gamma - \lambda_i - \eta_i) = 0$$
$$\lambda_i \geq 0, \;\; \eta_i \geq 0$$
$$y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) - 1 + \xi_i \geq 0$$
$$\xi_i \geq 0$$
$$\lambda_i \left[ y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) - 1 + \xi_i \right] = 0$$
$$\eta_i \xi_i = 0$$

## Lagrange function

$$L(\boldsymbol{\beta}, \beta_0, \xi_{1:n}, \lambda_{1:n}, \eta_{1:n})$$
$$= \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_i \xi_i - \sum_i \lambda_i \left[ y_i(\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) - 1 + \xi_i \right] - \sum_i \eta_i \xi$$
$$= \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_i \lambda_i y_i (\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) + \sum_i \lambda_i + \underline{\sum_i (\gamma - \lambda_i - \eta_i)\xi_i}$$




- Convex quadratic optimization problem with affine constraints (2n constraints).
- Any local optimum is a global optimum.
- **KKT conditions** are sufficient and necessary
- Equivalence between **the Primal** and **the Dual**.

gamma = C (SVMinR_JSS2006.pdf)

**Increaase** gamma →

- less number of support vectors,
- complex model,
- overfitting

## Max-Margin with Slack Variables

Then we introduce a slack variable $\xi_i$ for each sample, and formulate the max-margin problem as follows

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum \xi_i \\ \text{subject to} \quad & y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) - 1 + \xi_i \geq 0, \\ & \xi_i \geq 0. \end{aligned} \tag{3}$$

Note that $\xi_i > 0$ only for samples that are on the wrong side of the dashed line, and $\xi_i$ is automatically (by the optimization) set to be 0 for samples that are on the correct side of the dashed line.

**Primal**

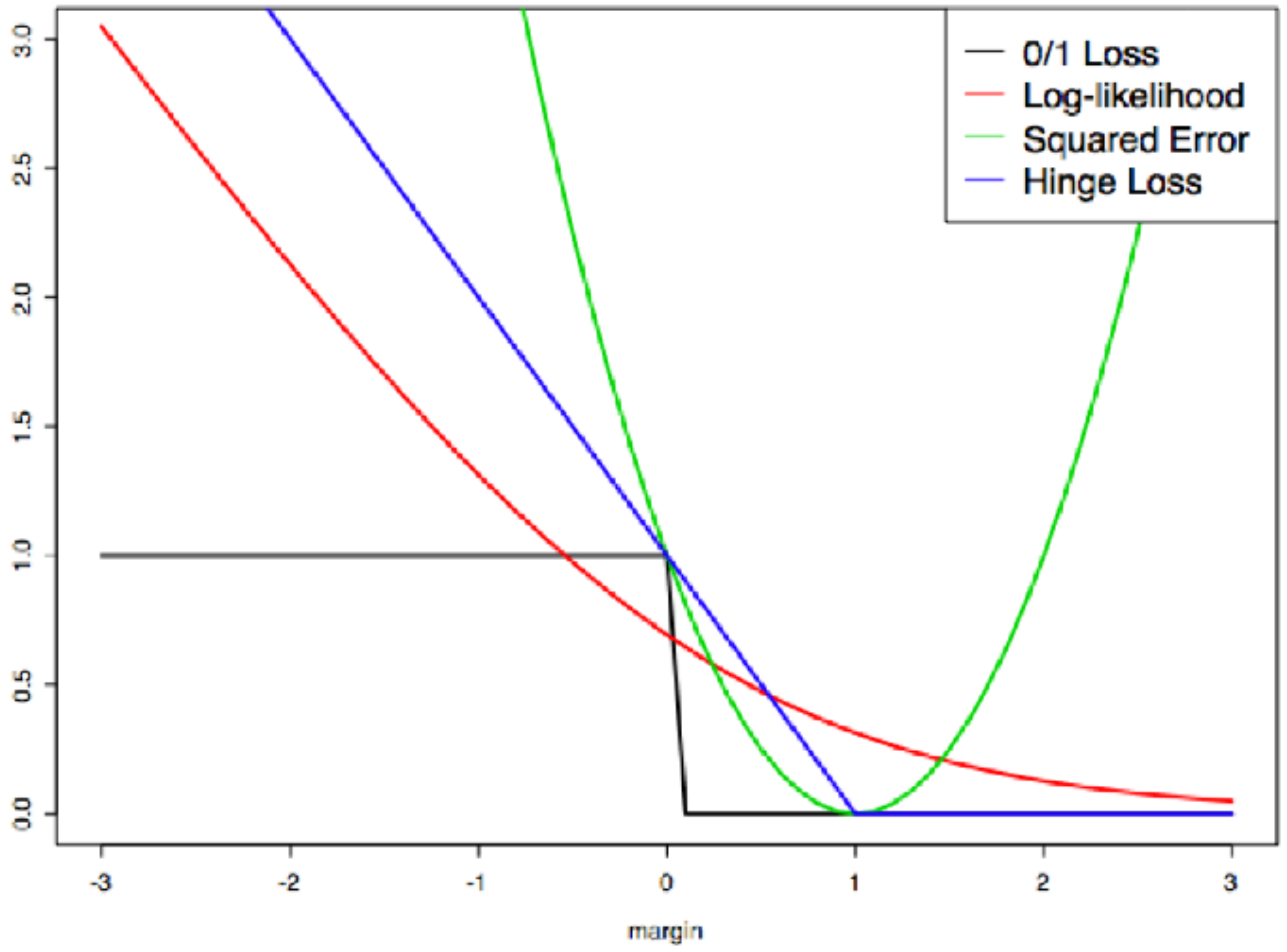
$$\min_{\boldsymbol{\beta},\beta_0,\xi_{1:n}} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum \xi_i$$

$$\text{subj to } y_i(\mathbf{x}_i \cdot \boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i,$$

$$\xi_i \geq 0$$

$$1 < y_i f(\mathbf{x}_i) \Longrightarrow 1 - y_i f(\mathbf{x}_i) < 0, \quad \xi_i = 0$$

$$1 \geq y_i f(\mathbf{x}_i), \Longrightarrow 1 - y_i f(\mathbf{x}_i) = \xi_i$$




## SVM as a penalization method

**Hinge Loss**

Let $f(x) = \mathbf{x} \cdot \boldsymbol{\beta} + \beta_0$ and $y_i \in \{-1, 1\}$. Then

$$\min_{\boldsymbol{\beta},\beta_0} \sum_{i=1}^{n} [1 - y_i f(\mathbf{x}_i)]_+ + \nu \|\boldsymbol{\beta}\|^2 \tag{6}$$

has the same solution as the linear SVM (3), when the tuning parameter $\nu$ is properly chose (which will depend on $\gamma$ in (3)). So SVM is a special case of the following Loss + Penalty framework

**Reciprocally related**