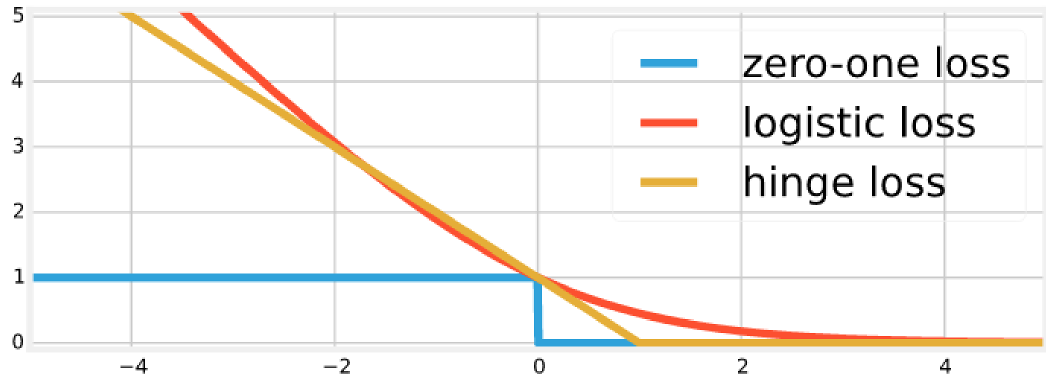
SUPPORT VECTOR MACHINES

Alexey Chervonenkis, Vladimir Vapnik

HINGE LOSS

LOSS FUNCTIONS



Training Loss

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{\text{data }(x,y)} \text{Loss}\big(y(\theta^\top x)\big)$$

Zero-One Loss

$$\text{Loss}_{01}(z) = [\![\, z \leq 0 \,]\!]$$

Hinge Loss

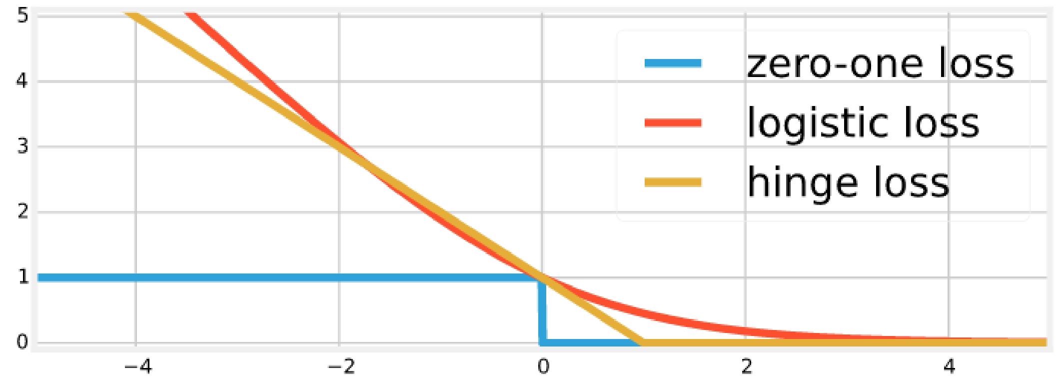
$$\text{Loss}_{\text{H}}(z) = \max\{1 - z, 0\}$$

CONVEX!

Penalize large mistakes more.

Penalize near-mistakes, i.e. $0 \leq z \leq 1$.

HINGE LOSS



Find θ that minimizes

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} \mathrm{Loss}_{\mathrm{H}}(z)$$

$$= \frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} \max\{1 - y(\theta^\top x), 0\}$$

Gradient

$$\nabla_{\mathrm{z}}\mathrm{Loss}_{\mathrm{H}}(z) = \begin{cases} 0 & \text{if } z > 1, \\ -1 & \text{otherwise.} \end{cases}$$

$$\nabla_{\theta}\mathrm{Loss}_{\mathrm{H}}\big(y(\theta^\top x)\big) = \begin{cases} 0 & \text{if } y(\theta^\top x) > 1, \\ -yx & \text{otherwise.} \end{cases}$$

LAGRANGE MULTIPLIERS

CONSTRAINED OPTIMIZATION

Want to minimize some function $f(x)$, but
there are some *constraints* on the values of x.

Method 1 (Dual Problem)

Solve a *dual optimization problem* where the constraints are nicer,
and where it is easier to implement gradient descent.

Method 2 (Exact Solution)

Solve the *Lagrangian* system of equations.

EQUALITY CONSTRAINTS

Problem.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \dots, h_l(x) = 0 \end{array}$$

Lagrangian.

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Example.

$$\begin{array}{ll} \text{minimize} & f(x) = n_1 \log x_1 + \cdots + n_d \log x_d \\ \text{subject to} & h(x) = x_1 + \cdots + x_d - 1 = 0 \end{array}$$

$$L(x, \lambda) = n_1 \log x_1 + \cdots + n_d \log x_d + \lambda(x_1 + \cdots + x_d - 1)$$

TWO-PLAYER GAME

$$L(x,\lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Rules.

- You get to choose the value of x.
  Your goal is to minimize $L(x,\lambda)$.
- Your adversary gets to choose the value of λ.
  His goal is to maximize $L(x,\lambda)$.

PRIMAL GAME

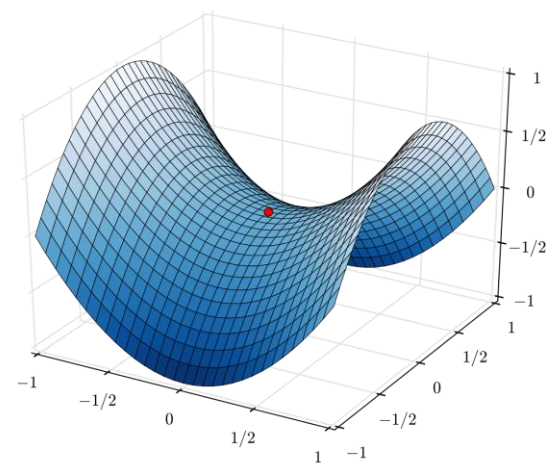
$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Primal Game. You go first.

Your Strategy.

- Ensure that $h_1(x) = 0, \ldots, h_l(x) = 0$.
- Find x that minimizes $f(x)$.

Final Score. $p^* = \min_x \max_\lambda L(x, \lambda)$



The optimal x^*, λ^* are *saddle points* of $L(x, \lambda)$.

DUAL GAME

$$L(x,\lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Dual Game. You go second.

Adversary's Strategy.

- For each λ, compute $\ell(\lambda) = \min_x L(x,\lambda)$
- Find λ that maximizes $\ell(\lambda)$.

Final Score. $d^* = \max_\lambda \min_x L(x,\lambda)$

MAX-MIN INEQUALITY

Primal. $p^* = \min_x \max_\lambda L(x, \lambda)$

Dual. $d^* = \max_\lambda \min_x L(x, \lambda)$

"you do better if you have the last say"

$$\boxed{\begin{aligned} p^* &= \min_x \max_\lambda L(x, \lambda) \\ &\geq \max_\lambda \min_x L(x, \lambda) = d^* \end{aligned}}$$

If $p^* = d^*$, we can solve the primal by solving the dual.

Challenge. Can you prove $p^* \geq d^*$? (*not in syllabus)

MAX-MIN INEQUALITY

Example.

	$x = 1$	$x = 2$
$\lambda = 1$	1	4
$\lambda = 2$	3	2

Primal. $p^* = \min_x \max_\lambda L(x, \lambda) = 3$

Dual. $d^* = \max_\lambda \min_x L(x, \lambda) = 2$

EXACT SOLUTION

Problem.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \dots, h_l(x) = 0 \end{array}$$

Lagrange multipliers.

1. Write down the Lagrangian.

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

2. Solve for critical points x, λ.

$$\nabla_x L(x, \lambda) = 0, \quad h_1(x) = 0, \dots, h_l(x) = 0$$

3. Pick critical point which gives global minimum.

EXAMPLE

$$\begin{aligned} &\text{minimize} \quad f(x) = n_1 \log x_1 + \cdots + n_d \log x_d \\ &\text{subject to} \quad h(x) = x_1 + \cdots + x_d - 1 = 0 \end{aligned}$$

Lagrangian

$$L(x, \lambda) = n_1 \log x_1 + \cdots + n_d \log x_d + \lambda(x_1 + \cdots + x_d - 1)$$

Critical points

$$\begin{aligned} 0 &= x_1 + \cdots + x_d - 1 \\ 0 &= n_i / x_i + \lambda \end{aligned} \quad \Rightarrow \quad \begin{aligned} (-\lambda) &= n_1 + \cdots + n_d \\ x_i &= n_i / (-\lambda) \end{aligned}$$

Inequality Constraints (Primal-Dual)

Primal Problem.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_1(x) \leq 0, \dots, g_m(x) \leq 0 \end{array}$$

Lagrangian.

$$L(x, \alpha) = f(x) + \alpha_1 g_1(x) + \cdots + \alpha_m g_m(x)$$

Dual Problem.

$$\begin{array}{ll} \text{maximize} & \ell(\alpha) \\ \text{subject to} & \alpha_1 \geq 0, \dots, \alpha_m \geq 0 \end{array} \qquad \text{where } \ell(\alpha) = \min_{x \in \mathbb{R}^d} L(x, \alpha)$$

Box constraints are easier to work with!

INEQUALITY CONSTRAINTS (EXACT SOLN)

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_1(x) \le 0, \dots, g_m(x) \le 0 \end{array}$$

Lagrangian.

$$L(x, \alpha) = f(x) + \alpha_1 g_1(x) + \cdots + \alpha_m g_m(x)$$

Solve for x, α satisfying

Karush-Kuhn-Tucker (KKT) Conditions

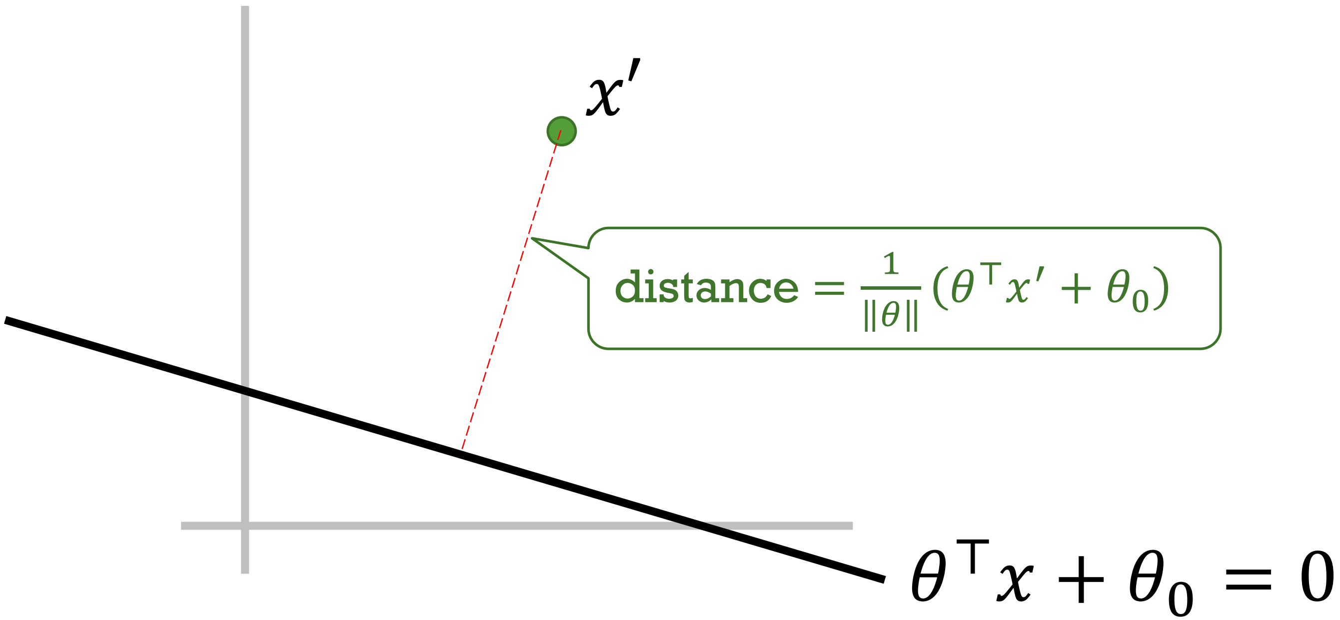
1. $\nabla_x L(x, \alpha) = 0$
2. $g_1(x) \le 0, \dots, g_m(x) \le 0$
3. $\alpha_1 \ge 0, \dots, \alpha_m \ge 0$
4. $\alpha_1 g_1(x) = 0, \dots, \alpha_m g_m(x) = 0$

Complementary Slackness

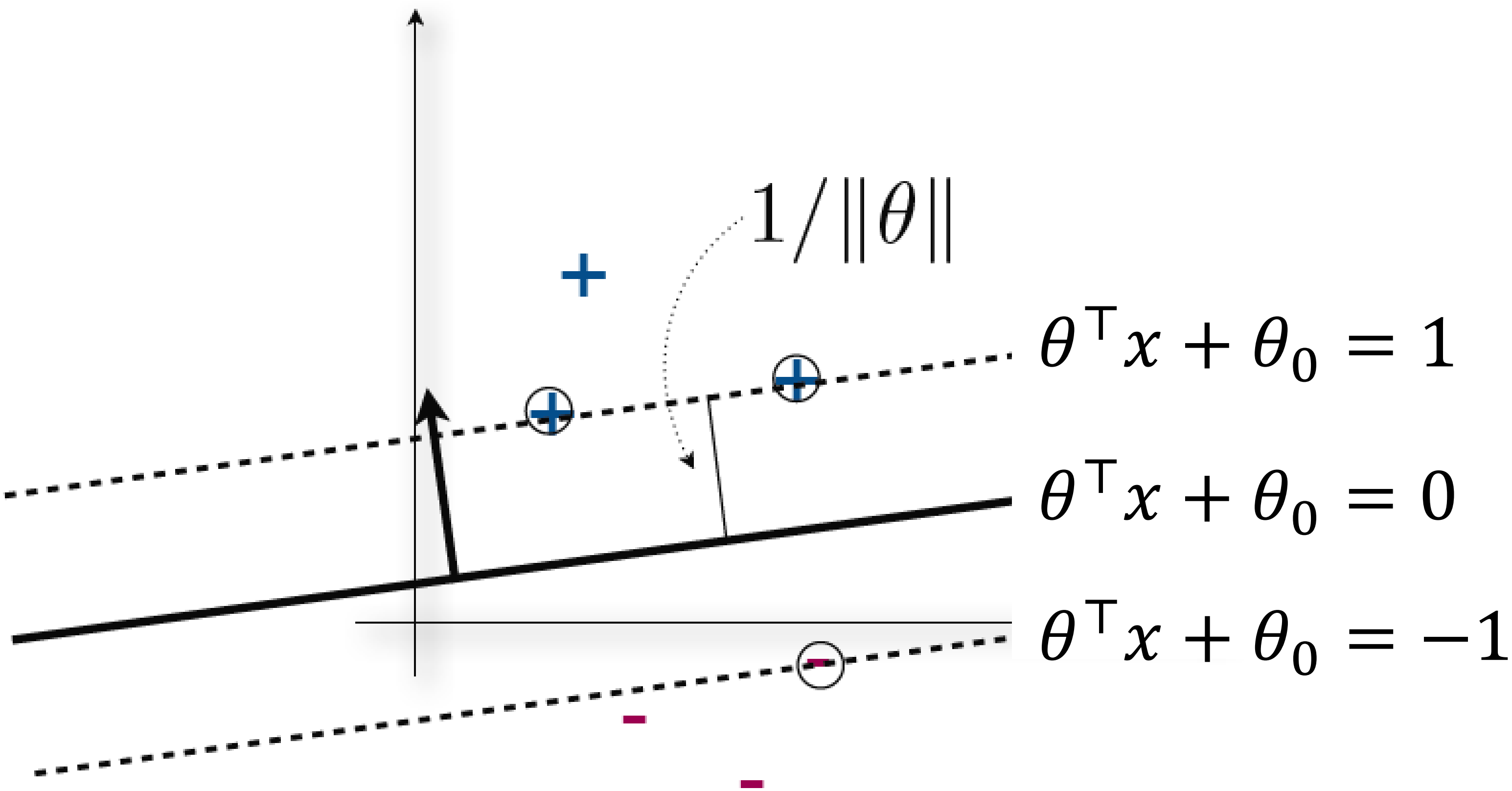
5 min break

MAXIMUM MARGINS

COMPUTING THE MARGIN



COMPUTING THE MARGIN



MAXIMUM MARGIN

Unfortunately, this only applies to data that is linearly separable.

Our goal is to

maximize $1/\|\theta\|$

subject to $y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

Or equivalently,

minimize $\frac{1}{2}\|\theta\|^2$

subject to $y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

LAGRANGIAN

Drop θ_0 for now

minimize $\frac{1}{2}\|\theta\|^2$

subject to $y(\theta^\top x) \geq 1$ for all data (x, y)

Lagrangian. $L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{(x,y)} \alpha_{x,y} \left(1 - y(\theta^\top x)\right)$

To find $\ell(\alpha) = \min\limits_{\theta} L(\theta, \alpha)$, we solve

$$0 = \nabla_\theta L(\theta, \alpha) = \theta - \sum_{(x,y)} \alpha_{x,y}\, yx$$

to get $\theta = \sum_{(x,y)} \alpha_{x,y}\, yx$. Substituting into $L(\theta, \alpha)$ gives

$$\ell(\alpha) = \sum_{(x,y)} \alpha_{x,y} \; - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'} yy'(x^\top x').$$

PRIMAL-DUAL

It can be shown that the primal and dual problems are equivalent (*strong duality*).

Primal.

minimize $\quad \frac{1}{2}\|\theta\|^2$

subject to $\quad y(\theta^\top x) \geq 1$ for all data (x, y)

Dual.

maximize $\quad \sum_{(x,y)} \alpha_{x,y} - \frac{1}{2}\sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'} yy'(x^\top x')$

subject to $\quad \alpha_{x,y} \geq 0$ for all (x, y)

After solving the dual to get the optimal $\alpha_{x,y}$'s,
we obtain the optimal θ using $\theta = \sum_{(x,y)} \alpha_{x,y}\, yx$.

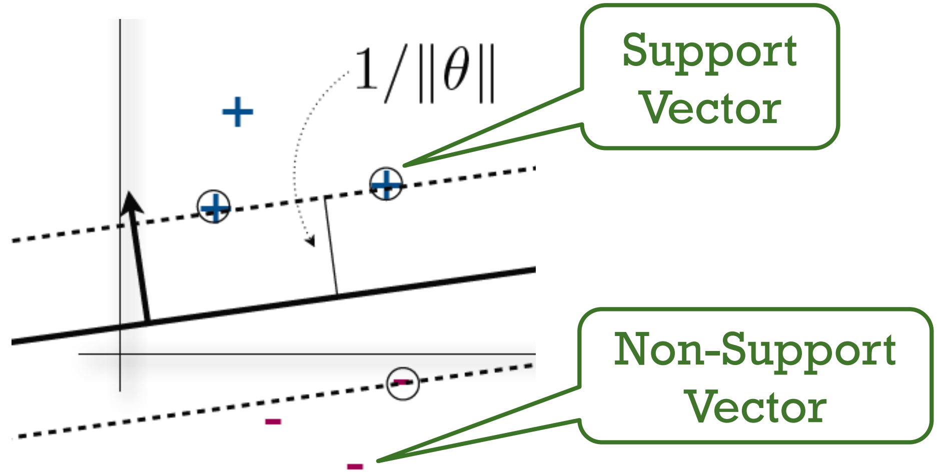
SUPPORT VECTORS

Complementary Slackness.

$\hat{\alpha}_{x,y} > 0$:	$y(\hat{\theta}^\top x) = 1$	Support Vectors
$\hat{\alpha}_{x,y} = 0$:	$y(\hat{\theta}^\top x) > 1$	Non-Support Vectors

Sparsity

Since very few data points are support vectors, most of the $\hat{\alpha}_{x,y}$ will be zero.



KERNEL TRICK

Learning.

$$\ell(\alpha) = \sum_{(x,y)} \alpha_{x,y} - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, y y' (x^\top x')$$

Prediction.

$$h(x;\theta) = \mathrm{sign}(\theta^\top x) = \mathrm{sign}\left(\sum_{(x',y')} \alpha_{x',y'}\, y' (x^\top x') \right)$$

For the dual, we don't need the feature vectors x, x'.
Knowing just the dot products $(x^\top x')$ is enough.

Recall that $(x^\top x')$ is a measure of similarity between x and x'.
This similarity function is also called a *kernel*.

EXTENSIONS

SVM WITH OFFSET

Primal.

minimize $\quad \frac{1}{2}\|\theta\|^2$

subject to $\quad y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

Dual.

maximize $\quad \sum_{(x,y)} \alpha_{x,y} - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, y y' (x^\top x')$

subject to $\quad \alpha_{x,y} \geq 0$ for all (x, y)

$\sum_{(x,y)} \alpha_{x,y}\, y = 0$

Parameters. $\quad \hat{\theta} \;= \sum_{(x,y)} \alpha_{x,y}\, y x$

$\hat{\theta}_0 = y - \hat{\theta}^\top x \quad$ where (x, y) is a support vector

SVM WITH ERRORS

Primal.

$$\begin{aligned}
&\text{minimize} \quad && \frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n}\sum_{(x,y)} \xi_{x,y} && \\
&\text{subject to} \quad && y(\theta^\top x + \theta_0) \geq 1 - \xi_{x,y} && \text{for all data } (x,y) \\
& && \xi_{x,y} \geq 0 && \text{for all data } (x,y)
\end{aligned}$$

Slack variables allow constraints to be violated for a cost.

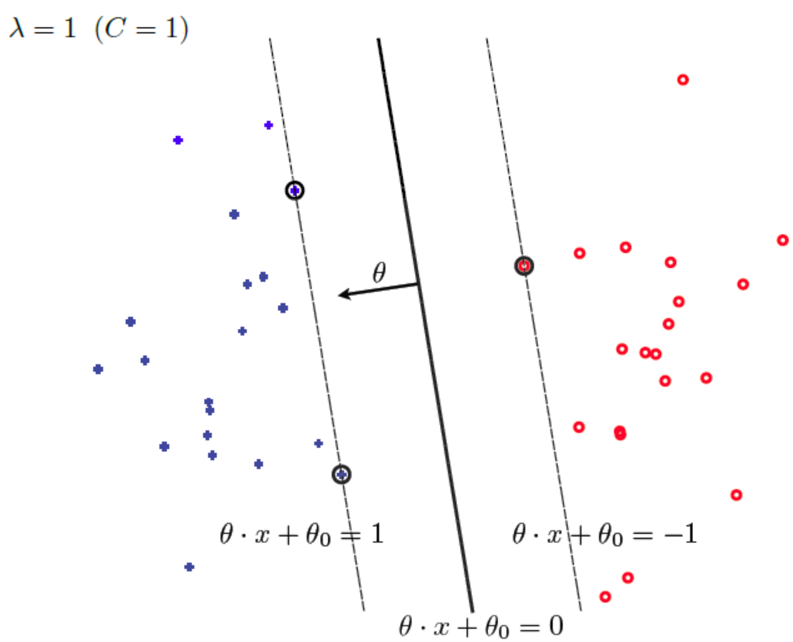
Equivalent Primal.

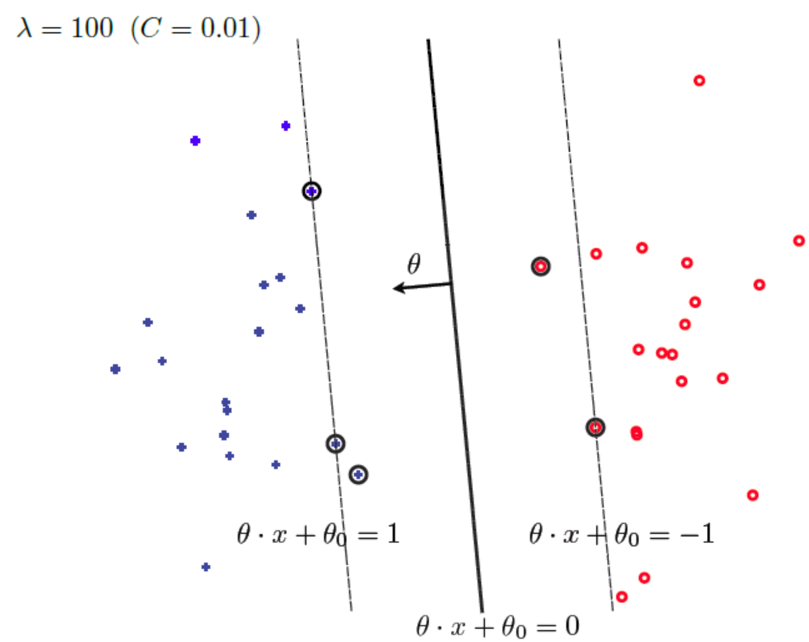
$$\text{minimize} \quad \frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n}\sum_{(x,y)} \mathrm{Loss}_{\mathrm{H}}\big(y(\theta^\top x + \theta_0)\big)$$

Hinge-loss classifier with regularization!

SVM WITH ERRORS

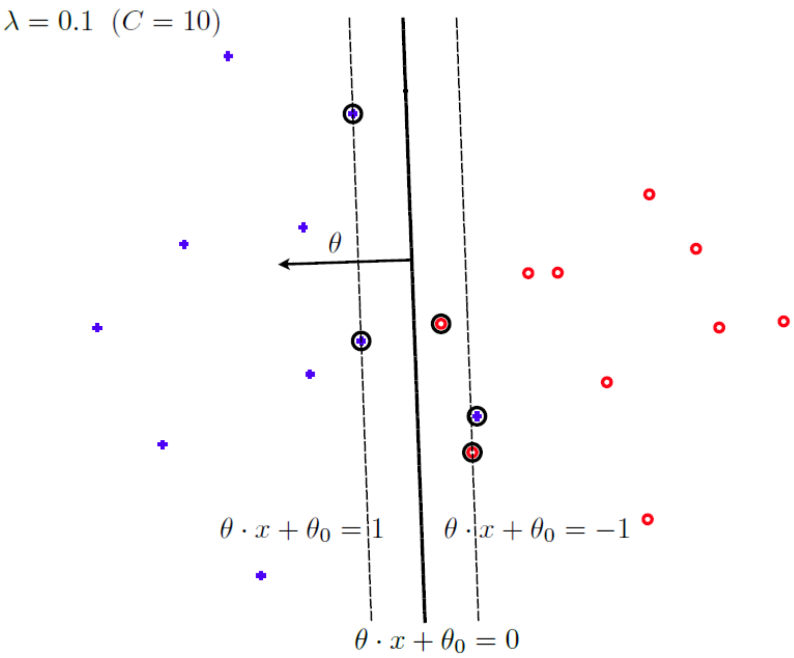
Linearly Separable.

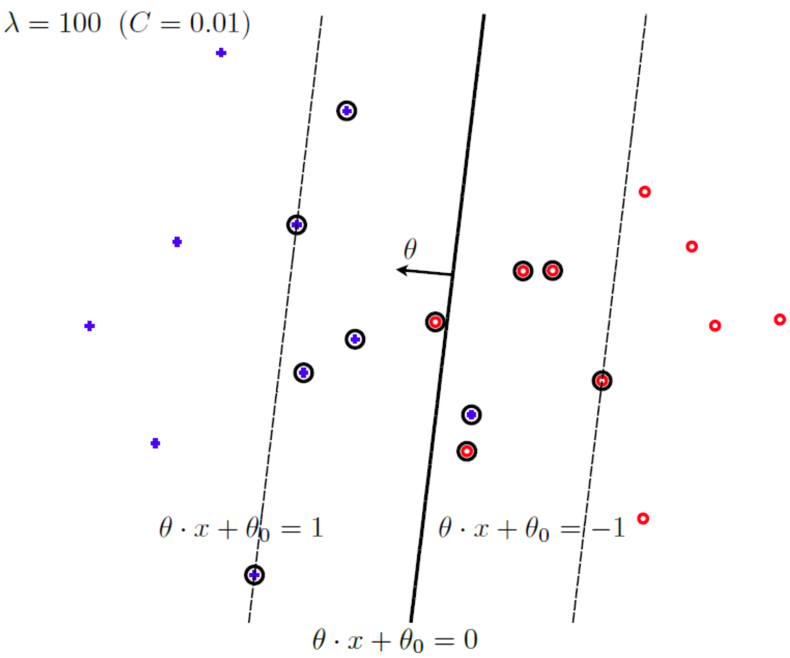




SVM WITH ERRORS

Not Linearly Separable.





SVM WITH ERRORS

Dual.

$$\text{maximize} \quad \sum_{(x,y)} \alpha_{x,y} - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, y y' (x^\top x')$$

$$\text{subject to} \quad 1/\lambda \geq \alpha_{x,y} \geq 0 \text{ for all } (x,y)$$

$$\sum_{(x,y)} \alpha_{x,y}\, y = 0$$

Putting limits on what the adversary can do.

There are many efficient solvers for quadratic problems with box constraints.