Lagrange Multipliers & Duality of SVMs

Rishabh Iyer
University of Texas at Dallas

Most Slides are from Nick Rouzzi! Thanks Nick!!

Lecture 5

Lagrange Multipliers & Duality of SVMs

[separability]

Rishabh Iyer

University of Texas at Dallas

Most Slides are from Nick Rouzzi! Thanks Nick!!

The Strategy So Far...

Regression — Perceptron/SVM

- Choose hypothesis space

Regression: $y \in \mathbb{R}$, $f(x) = w^T x + b$.

Perceptron/SVM: $y \in [-1,+1]$, $f(x) = \mathrm{sign}(w^T x + b)$

- Construct loss function (ideally convex)

$$L(y, f(x,w,b)) = [y - f(x,w,b)]^2 \;/\; L(y, f(x,w,b)) = \max(0, -y f(x,w,b))$$

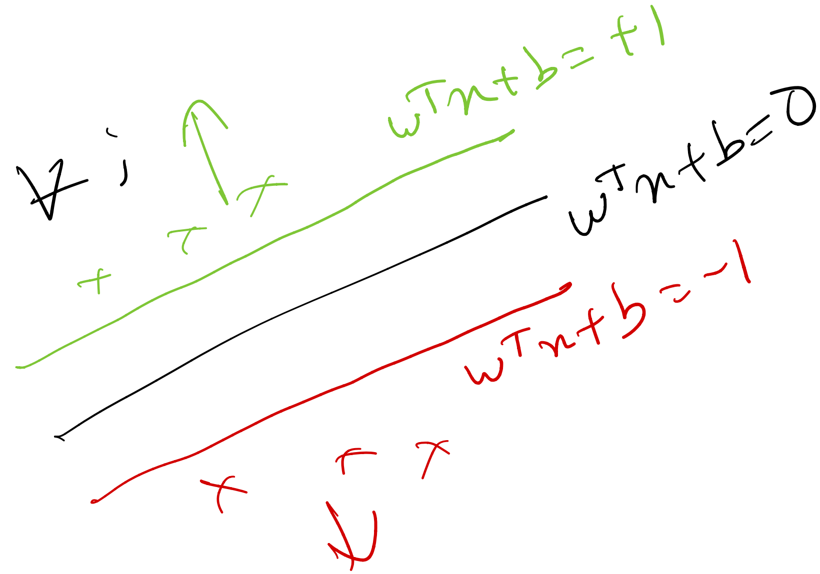
- Minimize loss to “learn” correct parameters

$$\sum_{i=1}^{M} L(y_i, f(x_i, w, b))$$

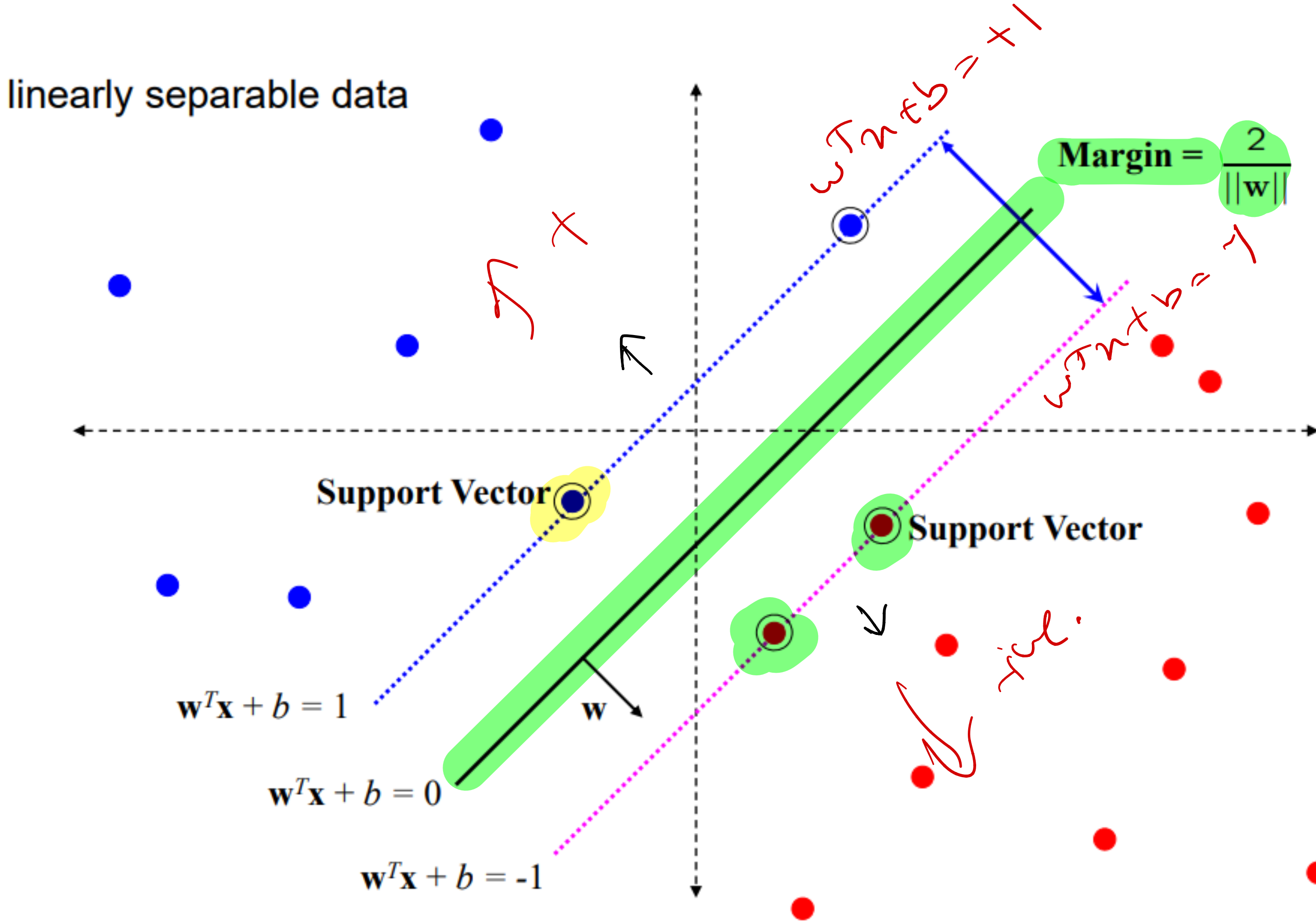
SVM

$$\min \frac{1}{2}\|w\|^2$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1, \ \forall i$$



Recap: SVM



linearly separable data
Margin = $\frac{2}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$
$\mathbf{w}$

SVM Optimization Problem

- Recall: The SVM optimization problem:

$$\min_{w,b} ||w||^2$$

such that

$$y^{(i)}(w^T x^{(i)} + b) \geq 1, \text{for all } i$$

- This is a standard quadratic programming problem
  - Falls into the class of convex optimization problems
  - Can be solved with many specialized optimization tools (e.g., quadprog() in MATLAB)

Constrained Optimization

A mathematical detour, we'll come back to SVMs soon!

$x \to$ opt variable.

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

Obj

$x = \begin{bmatrix} w \\ b \end{bmatrix}$

subject to:

Ineq Constraints

Constraints

$$f_i(x) \le 0, \qquad i = 1, \ldots, m$$
$$h_i(x) = 0, \qquad i = 1, \ldots, p$$

eq. constraints

$\min \frac{1}{2} \|w\|^2$

s.t. $y_i(w^T x_i + b) \ge 1$

$\Rightarrow \underbrace{1 - y_i(w^T x_i + b)}_{f_i(x)} \le 0$, $i = 1 \ldots M$

(w, b)

Constrained Optimization

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

f_0 is not necessarily convex

subject to:

$$f_i(x) \leq 0, \qquad i = 1, \ldots, m$$
$$h_i(x) = 0, \qquad i = 1, \ldots, p$$

General Optimization

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

Constraints do not need to be linear

$$f_i(x) \leq 0, \quad i = 1, \dots, m$$
$$h_i(x) = 0, \quad i = 1, \dots, p$$

Example

$$\overbrace{\min_{x \in \mathbb{R}^2} x_1 \log x_1 + x_2 \log x_2}^{f_0(x)}$$

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

subject to:

$$x_1 + x_2 = 1$$

$\leftarrow h(x) = 0$

$x_1 + x_2 - 1 = 0$

$$x_1 \geq 0$$

$$x_2 \geq 0$$

$f_1(x) \leq 0$

$-x_1 \leq 0$

$f_2(x) \leq 0$

$-x_2 \leq 0$

$f_1(x) = \begin{bmatrix} -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$= -x_1$

Example

$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

$f(x) = x \log x.$

$f'(x) = \log x + 1$

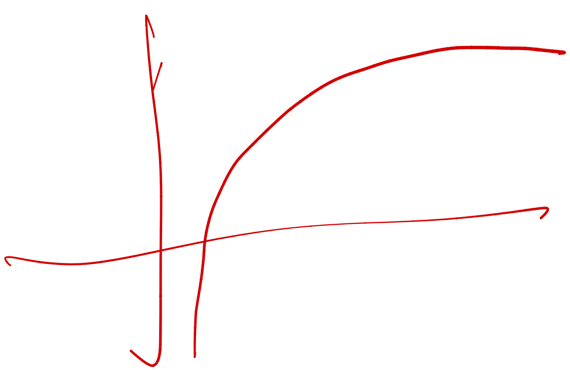
$f''(x) = \frac{1}{x}.$

subject to:

$$1 - x_1 - x_2 = 0$$
$$-x_1 \leq 0$$
$$-x_2 \leq 0$$

$f''(x) \geq 0$

$\frac{1}{x} \geq 0$



Lagrangian

min max
w λ,ν

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

Constraints : $\lambda \geq 0$

- Incorporate constraints into a new objective function
- $\lambda \geq 0$ and ν are vectors of ***Lagrange multipliers***
- The Lagrange multipliers can be thought of as enforcing soft constraints

$$\boxed{\begin{array}{ll} \min & f_0(x) \\ \text{s.t} & f_i(x) \leq 0 \\ & h_i(x) = 0 \end{array}}$$

λ_i $[i = 1:m]$

ν_i $[i = 1:p]$

Example

$$\min_{x\in\mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2 \; = f_0(x)$$

subject to:

$$1 - x_1 - x_2 = 0 \qquad h(x) \qquad \nu_1$$
$$-x_1 \le 0 \qquad f_1(x) \qquad \lambda_1$$
$$-x_2 \le 0 \qquad f_2(x) \qquad \lambda_2.$$

$$L(\overbrace{x_1, x_2}, \overbrace{\nu_1}^{h}, \overbrace{\lambda_1, \lambda_2})$$
$$= \underbrace{x_1 \log x_1 + x_2 \log x_2}_{f_0(x)} + \nu_1 \cdot \underbrace{(1 - x_1 - x_2)}_{h(x)} - \lambda_1 \underbrace{x_1}_{f_1(x)} - \underbrace{\lambda_2 x_2}_{f_2(x)}$$

Duality

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- Construct a dual function by minimizing the Lagrangian over the primal variables

Dual variables

$$\underbrace{g(\lambda, \nu)}_{} = \overset{\min}{\inf_x} L(x, \lambda, \nu)$$

- $g(\lambda, \nu) = -\infty$ whenever the Lagrangian is not bounded from below for a fixed λ and ν

$$\min \; f_0(x)$$
$$\text{s.t.} \quad h_i(x) = 0, \; i = 1:m$$
$$f_i(x) \leq 0 \quad i = 1:p$$

Primal

$$\max_{\lambda, \nu} \; g(\lambda, \nu)$$
$$\text{s.t.} \quad \lambda \geq 0$$
$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$$

Dual

Example

$$\min_{x \in \mathbb{R}^2} \underbrace{x_1 \log x_1 + x_2 \log x_2}_{f_0(x)}$$

subject to:

$$1 - x_1 - x_2 = 0 \leftarrow \nu_1$$
$$-x_1 \leq 0 \leftarrow \lambda_1$$
$$-x_2 \leq 0 \leftarrow \lambda_2$$

$$L(x, \lambda, \nu) = x_1 \log x_1 + x_2 \log x_2 + \nu_1(1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2$$

$$g(\underline{\lambda}, \underline{\nu}) = \min_x L(x, \lambda, \nu)$$

$$\frac{dL}{dx_1} = \log x_1 + 1 - \nu_1 - \lambda_1 = 0 \Rightarrow x_1 = \exp(\nu_1 + \lambda_1 - 1)$$

$$\log x_2 + 1 - \nu_1 - \lambda_2 = 0 \Rightarrow x_2 = \exp(\nu_1 + \lambda_2 - 1)$$

$$\frac{dL}{dx_2} = \log x_2 + 1$$

Example

$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

Primal

$$1 - x_1 - x_2 = 0$$
$$-x_1 \leq 0$$
$$-x_2 \leq 0$$

max $g(\lambda, v)$

s.t $\lambda \geq 0$

Dual.

$$g(\lambda, v) = (v_1 + \lambda_1 - 1)\exp(v_1 + \lambda_1 - 1) + (v_1 + \lambda_2 - 1)\exp(v_1 + \lambda_2 - 1)$$
$$+ v_1\left(1 - \exp(v_1 + \lambda_1 - 1) - \exp(v_1 + \lambda_2 - 1)\right)$$
$$- \lambda_1 \exp(v_1 + \lambda_1 - 1) - \lambda_2\left(\exp(v_1 + \lambda_2 - 1)\right)$$

$$= v_1 - \exp(v_1 + \lambda_1 - 1) - \exp(v_1 + \lambda_2 - 1)$$

$$\max\left(v_1 - \exp(v_1 + \lambda_1 - 1) - \exp(v_1 + \lambda_2 - 1)\right)$$

s.t $\lambda_1 \geq 0, \lambda_2 \geq 0$

1. Write down the constrained formulation

$$\min f_0(x)$$
$$\text{s.t.} \quad f_i(x) \le 0, \quad i = 1:m$$
$$h_i(x) = 0, \quad i = 1:p$$

2. $$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)$$

3. $$g(v, \lambda) = \min_x \Big[\underbrace{f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)}_{L(x, \lambda, v)} \Big]$$

$$\frac{\partial L}{\partial \underline{x}} = \underbrace{\frac{\partial f_0}{\partial x} + \sum_{i=1}^{m} \lambda_i \frac{\partial f_i}{\partial x} + \sum_{i=1}^{p} v_i \frac{\partial h_i}{\partial x} = 0}$$

$$\to x = F(\lambda, v)$$

The Primal Problem

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$f_i(x) \leq 0, \qquad i = 1, \dots, m$$
$$h_i(x) = 0, \qquad i = 1, \dots, p$$

Equivalently,

min max

$$\inf_x \underbrace{\left[\sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu) \right]}_{f_0(x)}$$

Why are these equivalent?

The Primal Problem

$$\min_{x\in\mathbb{R}^n} f_0(x)$$

$f_i(x) = 2$

$+2\lambda$

∞

subject to:

$\max_{\lambda \geq 0}$

$$f_i(x) \leq 0, \qquad i = 1, \dots, m$$
$$h_i(x) = 0, \qquad i = 1, \dots, p$$

$f_i(x) = -2$

-2λ

Equivalently,

$f_0(x)$

$$\inf_{x} \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$$

$f_i(x) \leq 0$

$h_i(x) = 0$

≤ 0

$$\sup_{\lambda \geq 0, \nu} \left[f_0(x) + \sum_{i=1}^{m} \lambda_i \underbrace{f_i(x)}_{\leq 0} + \sum_{i=1}^{p} \nu_i \underbrace{h_i(x)}_{0} \right] = \infty$$

whenever x violates the constraints

$$\boxed{\min_{w,b} \frac{1}{2}\|w\|^2 \qquad f_0(w,b)}$$

$$\text{s.t} \quad y_i(w^T x_i + b) \geq 1$$

$$\boxed{1 - y_i(w^T x_i + b) \leq 0 \qquad f_i(w,b) \leq 0}$$

Primal.

$\Downarrow$ Lagrangian.

$$L(w,b,\lambda) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{M} \lambda_i \left(1 - y_i(w^T x_i + b)\right)$$

$$\underbrace{g(\lambda)} = \min_{\underbrace{w,b}} \Big[\underbrace{\frac{1}{2}\|w\|^2 + \sum_{i=1}^{M} \lambda_i \left(1 - y_i(w^T x_i + b)\right)}_{L(\lambda, w, b)} \Big]$$

$$\frac{\partial L}{\partial w} = 0, \quad \frac{\partial L}{\partial b} = 0$$

$$w = F(\lambda), \; b = F(\lambda)$$

The Dual Problem

concave.

$$\sup_{\lambda \geq 0, \nu} g(\lambda, \nu)$$

$L(x, \lambda, \nu)$

Equivalently,

$$\sup_{\lambda \geq 0, \nu} \inf_x L(x, \lambda, \nu)$$

- The dual problem is always concave, even if the primal problem is not convex
  - For each x, $L(x, \lambda, \nu)$ is a linear function in λ and ν
  - Minimum (or infimum) of linear functions is concave!

Primal vs. Dual

λ^*, ν^*

$$\underbrace{\sup_{\lambda\ge 0,\nu} \overbrace{\inf_x L(x,\lambda,\nu)}^{g(\lambda,\nu)}}_{\text{Dual}} \le \underbrace{\inf_x \overbrace{\sup_{\lambda\ge 0,\nu} L(x,\lambda,\nu)}^{f_0(x)}}_{\text{Primal}} \leftarrow x^*$$

$$L(x,\lambda,\nu) = f_0(x) + \underbrace{\sum_{i=1}^{m} \underbrace{\lambda_i}_{\ge 0} \underbrace{f_i(x)}_{\le 0}}_{\le 0} + \sum_{i=1}^{p} \nu_i \underbrace{h_i(x)}_{=0}$$

- Why?
  - (i) $g(\lambda,\nu) \le L(x,\lambda,\nu)$ for all x
  - (ii) $L(x',\lambda,\nu) \le f_0(x')$ for any feasible x', $\lambda \ge 0$
    - x is feasible if it satisfies all of the constraints
  - Let x^* be the optimal solution to the primal problem and $\lambda \ge 0$

$$\text{Dual} \rightarrow \underbrace{g(\lambda,\nu) \le L(x^*,\lambda,\nu)}_{\text{(i)}} \overbrace{\le f_0(x^*)}^{\text{(ii)}} \rightarrow \text{Primal}$$

$$g(\lambda^*,\nu^*) \le f_0(x^*)$$

$$\min f_0(x)$$

$$\text{s.t.} \quad f_i(x) \leq 0, \; i = 1:m$$

$$h_i(x) \geq 0 \quad i = 1..p$$

$$\max g(\lambda, v)$$

$$\lambda \geq 0$$

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)$$

$$g(\lambda) = \max_{x} L(x, \lambda, v)$$

$$x = \underline{\underline{F(\lambda, v)}}$$

Example: Solving the Dual Problem

$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$
$$-x_1 \leq 0$$
$$-x_2 \leq 0$$

$\geq \quad \max_{\lambda, \nu} g(\lambda, \nu)$ s.t. $\lambda \geq 0$

$$\max_{\lambda, \nu} g(\lambda, \nu) = \max_{\lambda, \nu} \nu_1 - \underbrace{\exp(\nu_1 + \lambda_1 - 1)}_{\text{dec}(\lambda_1)} - \underbrace{\exp(\nu_1 + \lambda_2 - 1)}_{\text{dec}(\lambda_2)}$$

$\lambda \geq 0$

max attained at $(\lambda_1, \lambda_2) = (0, 0)$

$$\max_{\nu} \nu_1 - 2\exp(\nu_1 - 1) = g(\lambda, \nu)$$

$$\frac{dg}{d\nu} = 0 \Rightarrow 1 - 2\exp(\nu_1 - 1) = 0$$

$$\Rightarrow \nu_1 = \log(0.5) + 1$$

More Examples

- Minimize $x^2 + y^2$ subject to $x + y \geq 1$

- Given a point $z \in \mathbb{R}^n$ and a hyperplane $w^T x + b = 0$, find the projection of the point z onto the hyperplane

Duality

- Under certain conditions, the two optimization problems are equivalent

Dual   Primal

$$\sup_{\lambda\geq 0,\nu}\underbrace{\inf_{x} L(x,\lambda,\nu)}_{g(\lambda,\nu)} \underset{\leq}{=} \inf_{x}\underbrace{\sup_{\lambda\geq 0,\nu} L(x,\lambda,\nu)}_{f(x)}$$

  - This is called strong duality

- If the inequality is strict, then we say that there is a duality gap

  $<$

  - Size of gap measured by the difference between the two sides of the inequality

Slater's Condition

For any optimization problem of the form

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

Convex set.

$$f_i(x) \leq 0, \qquad i = 1, \dots, m$$

$$Ax = b$$

Affine.

where $f_0, \dots, f_m$ are convex functions, strong duality holds if there exists an x such that

$$f_i(x) < 0, \qquad i = 1, \dots, m$$

$$Ax = b$$

Dual SVM

$w \in \mathbb{R}^n$

$x \in \mathbb{R}^M$

convex ✓

$$\min_{w} \frac{1}{2}\|w\|^2 \quad \leftarrow f_0(w,b)$$

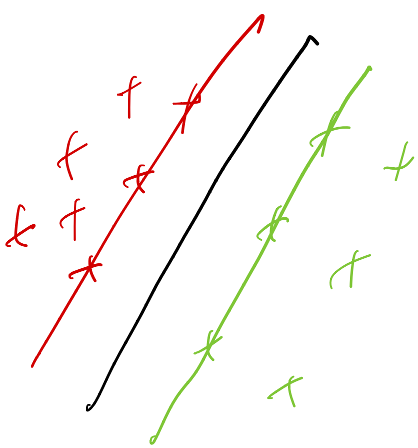
such that

convex ✓

$$\underbrace{y_i\left(w^T x^{(i)} + b\right) \geq 1,}\text{for all } i \quad \leftarrow f_i(w,b) \leq 0$$

$$1 - y_i(w^T x^i + b) \leq 0$$

- Note that Slater's condition holds as long as the data is linearly separable



$$1 - y_i(w^T x^i + b) \leq 0, \quad \forall\, i = 1:M$$

$M = \#$ Train Examples

Dual SVM

$$L(w,b,\lambda) = \underbrace{\frac{1}{2}w^T w + \sum_i \lambda_i (1 - y_i(w^T x^{(i)} + b))}$$

Convex in w, so take derivatives to form the dual $\lambda \in \mathbb{R}^M$

$$\frac{\partial L}{\partial w_k} = w_k + \sum_i -\lambda_i y_i x_k^{(i)} = 0$$

$$\frac{\partial L}{\partial b} = \sum_i -\lambda_i y_i = 0 \Rightarrow \sum_i \lambda_i y_i = 0$$

$$w = \sum_{i=1}^{m} \lambda_i y_i x^{(i)}$$

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{M} \lambda_i \left(1 - y_i (w^T x^i + b)\right)$$

$$w = \sum_{i=1}^{M} \lambda_i y_i x^{(i)}$$

$$\frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \lambda_i \lambda_j y_i y_j x^{(i)T} x^{(j)} + \sum_{i=1}^{M} \lambda_i \left(1 - y_i \sum_{j=1}^{M} \lambda_j y_j x^{(j)T} x^{(i)} - y_i b\right)$$

$$- \sum_{i=1}^{M} \sum_{j=1}^{M} y_i y_j \lambda_i \lambda_j x^{(j)T} x^{(i)} + \sum_{i=1}^{M} \lambda_i + 0$$

$$=$$

Dual SVM

$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_i \lambda_i (1 - y_i (w^T x^{(i)} + b))$$

Convex in w, so take derivatives to form the dual

$$w = \sum_i \lambda_i y_i x^{(i)}$$

$$\sum_i \lambda_i y_i = 0$$

Dual SVM

$$\max_{\lambda \geq 0} -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)^T} x^{(j)} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

- By strong duality, solving this problem is equivalent to solving the primal problem
    - Given the optimal λ, we can easily construct w (b can be found by complementary slackness…)

Complementary Slackness

- Suppose that there is zero duality gap
- Let x^* be an optimum of the primal and (λ^*, ν^*) be an optimum of the dual

$$
\begin{aligned}
f_0(x^*) &= g(\lambda^*, \nu^*) \\
&= \inf_x \left[f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right] \\
&\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \\
&= f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) \\
&\leq f_0(x^*)
\end{aligned}
$$

Complementary Slackness

- This means that

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

- As $\lambda \geq 0$ and $f_i(x_i^*) \leq 0$, this can only happen if $\lambda_i^* f_i(x^*) = 0$ for all i
- Put another way,
  - If $f_i(x^*) < 0$ (i.e., the constraint is not tight), then $\lambda_i^* = 0$
  - If $\lambda_i^* > 0$, then $f_i(x^*) = 0$
  - ONLY applies when there is no duality gap

Dual SVM (Obtaining b)

$$\max_{\lambda \geq 0} -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)^T} x^{(j)} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

- By complementary slackness, $\lambda_i^* > 0$ means that $x^{(i)}$ is a support vector (can then solve for b using w)
- In particular,

$$b = y_i \ - w.x_i$$

for any i where $\lambda_i > 0$

Dual SVM

$$\max_{\lambda \geq 0} -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)^T} x^{(j)} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

- Takes $O(n^2)$ time just to evaluate the objective function
  - Active area of research to try to speed this up