# SVM Dual Problem

He He

Slides based on Lecture 4c from David Rosenberg's course material.

CDS, NYU

Feb 23, 2021

# SVM as a Quadratic Program

- The SVM optimization problem is equivalent to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & -\xi_i \leqslant 0 \quad \text{for } i = 1,\ldots,n \\ & \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \quad \text{for } i = 1,\ldots,n \end{array}$$

- Differentiable objective function
- $n + d + 1$ unknowns and $2n$ affine constraints.
- A **quadratic program** that can be solved by any off-the-shelf QP solver.
- Let's learn more by examining the dual.

# Why Do We Care About the Dual?

# The Lagrangian

The general [inequality-constrained] optimization problem is:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0, \quad i = 1, \ldots, m \end{aligned}$$

## Definition

The **Lagrangian** for this optimization problem is

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

- $\lambda_i$'s are called **Lagrange multipliers** (also called the **dual variables**).
- Weighted sum of the objective and constraint functions
- Hard constraints $\rightarrow$ soft constraints

# Lagrange Dual Function

**Definition**

The **Lagrange dual function** is

$$g(\lambda) = \inf_x L(x,\lambda) = \inf_x \left( f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right)$$

- $g(\lambda)$ is concave (why?)
- **Lower bound property**: if $\lambda \succeq 0$, $g(\lambda) \leqslant p^*$ where $p^*$ is the optimal value of the optimization problem.
- $g(\lambda)$ can be $-\infty$ (uninformative lower bound)

# The Primal and the Dual

- For any **primal form** optimization problem,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leqslant 0, \;\; i = 1, \ldots, m, \end{array}$$

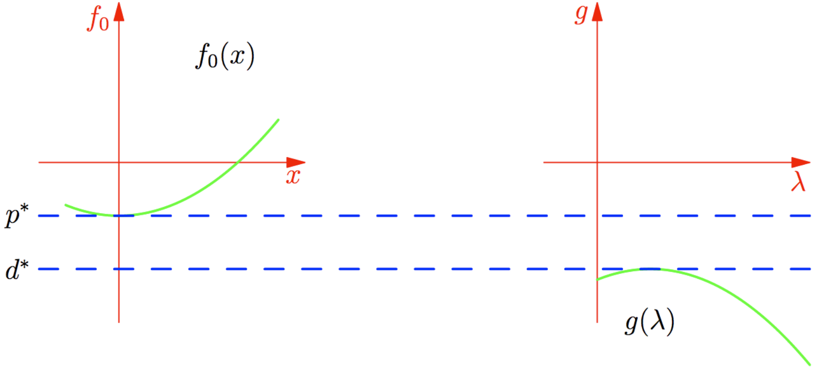
  there is a recipe for constructing a corresponding **Lagrangian dual problem:**

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda_i \geqslant 0, \;\; i = 1, \ldots, m, \end{array}$$

- The dual problem is always a convex optimization problem.
- The dual variables often have interesting and relevant interpretations.
- The dual variables provide certificate for optimality.

# Weak Duality

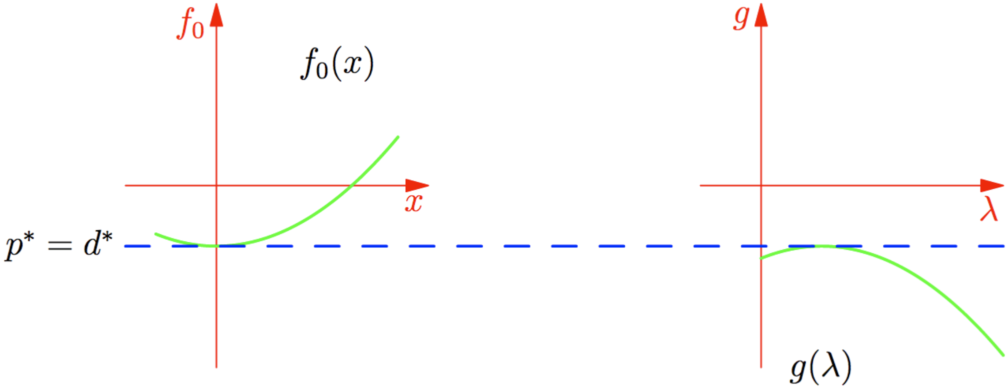
We always have **weak duality:** $p^* \geqslant d^*$.




Plot courtesy of Brett Bernstein.

# Strong Duality

For some problems, we have **strong duality**: $p^* = d^*$.




For convex problems, strong duality is fairly typical.

Plot courtesy of Brett Bernstein.

# Complementary Slackness

- Assume strong duality. Let $x^*$ be primal optimal and $\lambda^*$ be dual optimal. Then:

$$
\begin{aligned}
f_0(x^*) &= g(\lambda^*) = \inf_x L(x, \lambda^*) \quad \text{(strong duality and definition)} \\
&\leqslant L(x^*, \lambda^*) \\
&= f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) \\
&\leqslant f_0(x^*).
\end{aligned}
$$

Each term in sum $\sum_{i=1} \lambda_i^* f_i(x^*)$ must actually be 0. That is

$$
\lambda_i > 0 \implies f_i(x^*) = 0 \quad \text{and} \quad f_i(x^*) < 0 \implies \lambda_i = 0 \quad \forall i
$$

This condition is known as **complementary slackness**.

# The SVM Dual Problem

# SVM Lagrange Multipliers

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & -\xi_i \leqslant 0 \quad \text{for } i = 1,\ldots,n \\ & \left(1 - y_i\left[w^Tx_i + b\right]\right) - \xi_i \leqslant 0 \quad \text{for } i = 1,\ldots,n \end{array}$$

| Lagrange Multiplier | Constraint |
|---|---|
| $\lambda_i$ | -$\xi_i \leqslant 0$ |
| $\alpha_i$ | $\left(1 - y_i\left[w^Tx_i + b\right]\right) - \xi_i \leqslant 0$ |

$$L(w,b,\xi,\alpha,\lambda) = \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left[w^Tx_i + b\right] - \xi_i\right) + \sum_{i=1}^{n}\lambda_i\left(-\xi_i\right)$$

Dual optimum value: $d^* = \sup_{\alpha,\lambda\succeq 0}\inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda)$

# Strong Duality by Slater's Constraint Qualification

The SVM optimization problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & -\xi_i \leqslant 0 \text{ for } i = 1, \ldots, n \\ & \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1, \ldots, n \end{aligned}$$

Slater's constraint qualification:

- Convex problem + affine constraints $\Longrightarrow$ strong duality iff problem is feasible
- Do we have a feasible point?
- For SVM, we have strong duality.

# SVM Dual Function: First Order Conditions

Lagrange dual function is the inf over primal variables of $L$:

$$
\begin{aligned}
g(\alpha,\lambda) &= \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda) \\
&= \inf_{w,b,\xi}\left[\frac{1}{2}w^Tw+\sum_{i=1}^n \xi_i\left(\frac{c}{n}-\alpha_i-\lambda_i\right)+\sum_{i=1}^n \alpha_i\left(1-y_i\left[w^Tx_i+b\right]\right)\right]
\end{aligned}
$$

$$\partial_w L = 0 \iff w-\sum_{i=1}^n \alpha_i y_i x_i = 0 \iff \boxed{w=\sum_{i=1}^n \alpha_i y_i x_i}$$

$$\partial_b L = 0 \iff -\sum_{i=1}^n \alpha_i y_i = 0 \iff \boxed{\sum_{i=1}^n \alpha_i y_i = 0}$$

$$\partial_{\xi_i} L = 0 \iff \frac{c}{n}-\alpha_i-\lambda_i = 0 \iff \boxed{\alpha_i+\lambda_i=\frac{c}{n}}$$

# SVM Dual Function

- Substituting these conditions back into $L$, the second term disappears.
- First and third terms become

$$
\begin{aligned}
\frac{1}{2} w^T w &= \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\
\sum_{i=1}^{n} \alpha_i \left(1 - y_i \left[ w^T x_i + b \right] \right) &= \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0}.
\end{aligned}
$$

- Putting it together, the dual function is

$$
g(\alpha, \lambda) = \begin{cases} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i & \substack{\sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i + \lambda_i = \frac{c}{n}, \text{ all } i} \\ -\infty & \text{otherwise.} \end{cases}
$$

# SVM Dual Problem

- The **dual function** is

$$g(\alpha,\lambda) = \begin{cases} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j x_j^T x_i & \begin{array}{c}\sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i + \lambda_i = \frac{c}{n}, \text{ all } i\end{array} \\ -\infty & \text{otherwise.} \end{cases}$$

- The **dual problem** is $\sup_{\alpha,\lambda\succeq 0} g(\alpha,\lambda)$:

$$\begin{array}{ll} \sup_{\alpha,\lambda} & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j x_j^T x_i \\ \text{s.t.} & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i + \lambda_i = \frac{c}{n} \quad \alpha_i, \lambda_i \geqslant 0, \; i = 1, \ldots, n \end{array}$$

# Insights from the Dual Problem

# KKT Conditions

For convex problems, if Slater's condition is satisfied, then **KKT conditions** provide necessary and sufficient conditions for the optimal solution.

- Primal feasibility: $f_i(x) \leqslant 0 \quad \forall i$
- Dual feasibility: $\lambda \succeq 0$
- Complementary slackness: $\lambda_i f_i(x) = 0$
- First-order condition:

$$\frac{\partial}{\partial x} L(x, \lambda) = 0$$

# The SVM Dual Solution

- We found the SVM dual problem can be written as:

$$
\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \;\; i = 1, \ldots, n.
\end{aligned}
$$

- Given solution $\alpha^*$ to dual, primal solution is $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$.
- The solution is in the space spanned by the inputs.
- Note $\alpha_i^* \in [0, \frac{c}{n}]$. So $c$ controls max weight on each example. (**Robustness**!)
  - What's the relation between $c$ and regularization?

# Complementary Slackness Conditions

- Recall our primal constraints and Lagrange multipliers:

| Lagrange Multiplier | Constraint |
|---|---|
| $\lambda_i$ | $-\xi_i \leqslant 0$ |
| $\alpha_i$ | $(1 - y_i f(x_i)) - \xi_i \leqslant 0$ |

- Recall first order condition $\nabla_{\xi_i} L = 0$ gave us $\lambda_i^* = \frac{c}{n} - \alpha_i^*$.
- By strong duality, we must have **complementary slackness**:

$$\alpha_i^* \left(1 - y_i f^*(x_i) - \xi_i^*\right) = 0$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

# Consequences of Complementary Slackness

By strong duality, we must have **complementary slackness**.

$$\alpha_i^*\left(1 - y_i f^*(x_i) - \xi_i^*\right) = 0$$
$$\left(\frac{c}{n} - \alpha_i^*\right)\xi_i^* = 0$$

Recall "**slack variable**" $\xi_i^* = \max\left(0, 1 - y_i f^*(x_i)\right)$ is the hinge loss on $(x_i, y_i)$.

- If $y_i f^*(x_i) > 1$ then the margin loss is $\xi_i^* = 0$, and we get $\alpha_i^* = 0$.
- If $y_i f^*(x_i) < 1$ then the margin loss is $\xi_i^* > 0$, so $\alpha_i^* = \frac{c}{n}$.
- If $\alpha_i^* = 0$, then $\xi_i^* = 0$, which implies no loss, so $y_i f^*(x) \geqslant 1$.
- If $\alpha_i^* \in \left(0, \frac{c}{n}\right)$, then $\xi_i^* = 0$, which implies $1 - y_i f^*(x_i) = 0$.

# Complementary Slackness Results: Summary

If $\alpha^*$ is a solution to the dual problem, then primal solution is

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \quad \text{where} \alpha_i^* \in [0, \frac{c}{n}].$$

Relation between margin and example weights ($\alpha_i$'s):

$$\begin{aligned}
\alpha_i^* = 0 \quad &\implies \quad y_i f^*(x_i) \geqslant 1 \\
\alpha_i^* \in \left(0, \frac{c}{n}\right) \quad &\implies \quad y_i f^*(x_i) = 1 \\
\alpha_i^* = \frac{c}{n} \quad &\implies \quad y_i f^*(x_i) \leqslant 1
\end{aligned}$$

$$\begin{aligned}
y_i f^*(x_i) < 1 \quad &\implies \quad \alpha_i^* = \frac{c}{n} \\
y_i f^*(x_i) = 1 \quad &\implies \quad \alpha_i^* \in \left[0, \frac{c}{n}\right] \\
y_i f^*(x_i) > 1 \quad &\implies \quad \alpha_i^* = 0
\end{aligned}$$

# Support Vectors

- If $\alpha^*$ is a solution to the dual problem, then primal solution is

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

with $\alpha_i^* \in [0, \frac{c}{n}]$.

- The $x_i$'s corresponding to $\alpha_i^* > 0$ are called **support vectors.**
- Few margin errors or "on the margin" examples $\implies$ sparsity in input examples.

# The Bias Term: $b$

- For our SVM primal, the complementary slackness conditions are:

$$\alpha_i^* \left(1 - y_i \left[x_i^T w^* + b\right] - \xi_i^*\right) = 0 \tag{1}$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0 \tag{2}$$

- Suppose there's an $i$ such that $\alpha_i^* \in \left(0, \frac{c}{n}\right)$.
- (2) implies $\xi_i^* = 0$.
- (1) implies

$$\begin{aligned} & y_i \left[x_i^T w^* + b^*\right] = 1 \\ \iff \quad & x_i^T w^* + b^* = y_i \text{ (use } y_i \in \{-1, 1\}) \\ \iff \quad & \boxed{b^* = y_i - x_i^T w^*} \end{aligned}$$

# The Bias Term: $b$

- We get the same $b^*$ for any choice of $i$ with $\alpha_i^* \in (0, \frac{c}{n})$

$$b^* = y_i - x_i^T w^*$$

- With numerical error, more robust to average over all eligible $i$'s:

$$b^* = \text{mean}\left\{ y_i - x_i^T w^* \mid \alpha_i^* \in \left(0, \frac{c}{n}\right) \right\}.$$

- If there are no $\alpha_i^* \in (0, \frac{c}{n})$?
  - Then we have a **degenerate SVM training problem**[1] $(w^* = 0)$.

[1]See Rifkin et al.'s "A Note on Support Vector Machine Degeneracy", an MIT AI Lab Technical Report.

# Teaser for Kernelization

# Dual Problem: Dependence on $x$ through inner products

- SVM Dual Problem:

$$
\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \;\; i = 1, \ldots, n.
\end{aligned}
$$

- Note that all dependence on inputs $x_i$ and $x_j$ is through their inner product: $\langle x_j, x_i \rangle = x_j^T x_i$.
- We can replace $x_j^T x_i$ by other products...
- This is a "kernelized" objective function.