Lagrange Duality and KKT Conditions

ECE 4200
02/21/20 Recitation

Mathematical Optimization

- All learning is some optimization problem
  -> Stick to canonical form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \dots, m \end{array}$$

- $x = (x_1, x_2, \dots, x_n)$ – opt. variables ; x^*
- $f_0 : R^n \to R$ – *objective function*
- $f_i : R^n \to R$ – *constraint function*

Optimization Example

- Well familiar with: *regression*
  - *Least squares*

$$\min. \quad ||y - X\beta||_2^2$$

  - *Add some constraints*

$$\begin{aligned} \min. \quad & ||y - X\beta||_2^2 \\ \text{s.t.} \quad & ||\beta||_2^2 \leq \lambda \end{aligned}$$

Why convex optimization?

- Can’t solve most OPs
  - E.g. NP Hard, even high polynomial time too slow

- Convex OPs
  - (Generally) No analytic solution
  - Efficient algorithms to find (global) solution

What is Convex Optimization?

- OP with *convex* objective and constraint functions

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

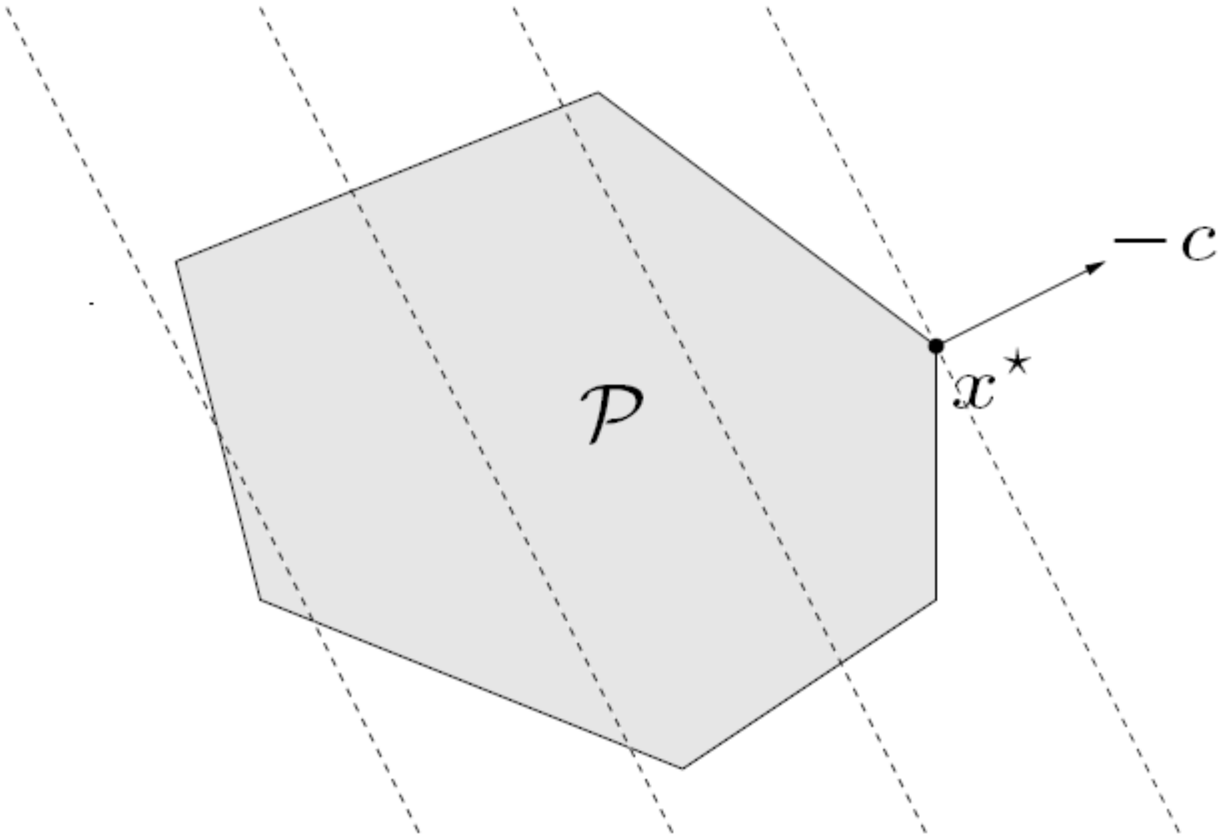
- f_0 , … , f_m are *convex* = convex OP that has an efficient solution!

Some common convex OPs

- LP: affine objective function, affine constraints

$$\begin{array}{ll} \text{minimize} & c^T x + d \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array}$$

LP Visualization



Quadratic Program

- QP: Quadratic objective, affine constraints

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P x + q^T x + r \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array}$$

- LP is special case
- Many SVM problems result in QP, regression

SVM Optimization

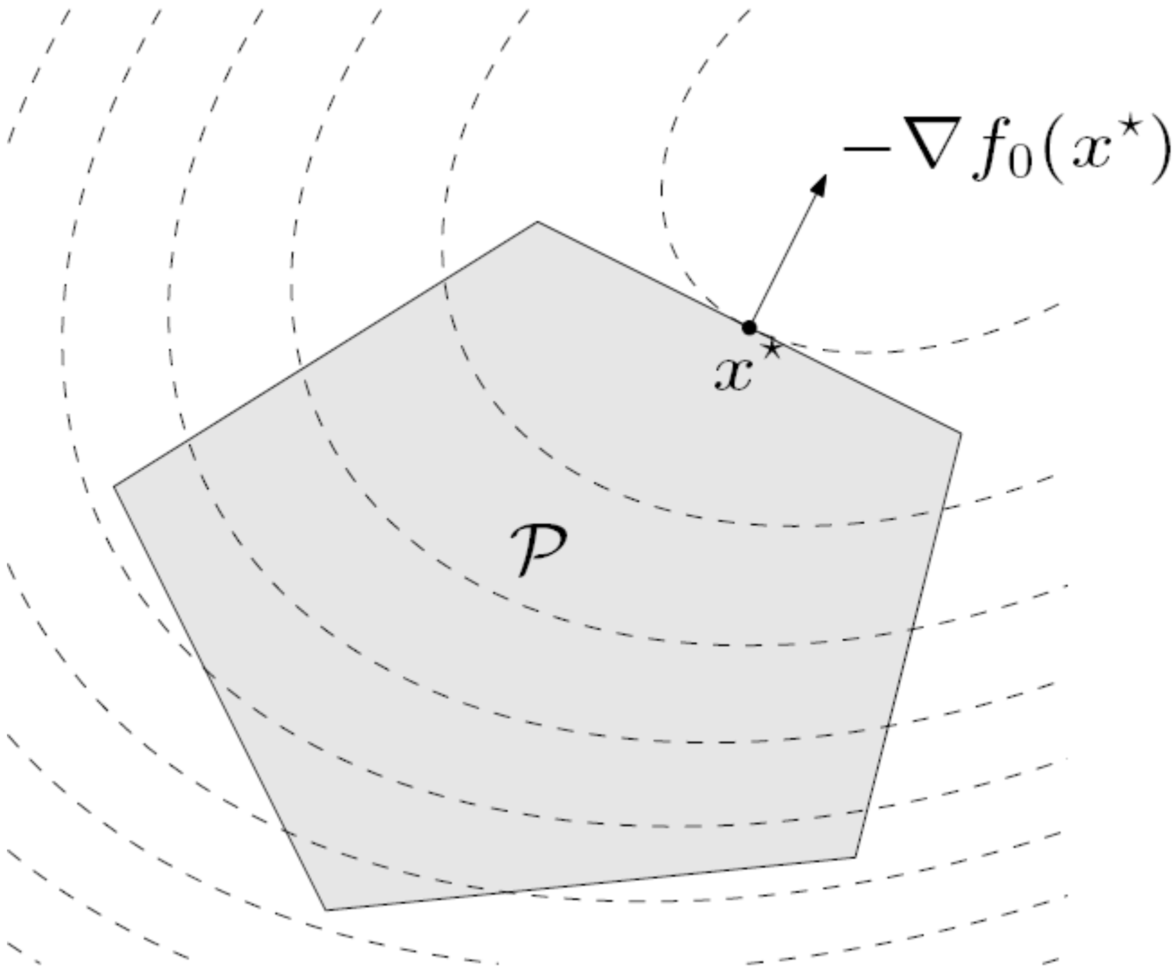
Minimize over $\overline{w}, t$:

$$\| \overline{w} \|_2^2$$

Subject to:

$$y_i(\overline{w} \cdot \bar{X}_i - t) \geq 1$$

QP Visualization



Lagrange

- Standard form:

$$
\begin{array}{ll}
\text{minimize} & f_0(x) \\
\text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots, m \\
& h_i(x) = 0, \quad i = 1, \dots, p
\end{array}
$$

- Lagrange L:

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

Lagrange Dual Function

$$
\begin{aligned}
g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\
&= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)
\end{aligned}
$$

- Lagrange Dual found by minimizing L with respect to primal variables

Lagrange Dual Function

- Lagrange Dual provides lower bound on objective value at solution

lower bound property: if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^\star$

proof: if $\tilde{x}$ is feasible and $\lambda \succeq 0$, then

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\tilde{x}$ gives $p^\star \geq g(\lambda, \nu)$

Lagrange Dual Problem

- Why not make the lower bound best possible?
- Dual problem:

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- *Always* convex opt. problem (even when primal is non-convex)
- **Weak Duality:** d* <= p* (have already seen this)

Lagrange Dual Function

- Note: Lagrange dual function is the point-wise infimum of family of affine functions of (lambda, nu)
- Thus, g is concave even if problem is not convex

Example: SVM

Minimize over $\vec{w}, t$:

$$\| \vec{w} \|_2^2$$

Subject to:

$$\forall i, 1 - y_i(\vec{w} \cdot \vec{x_i} - t) \leq 0$$

Lagrange L:

$$\mathcal{L}(\vec{w}, t, \boldsymbol{\alpha}) = \frac{1}{2} \cdot \| \vec{w} \|_2^2 + \sum_{1 \leq i \leq m} \alpha_j \left(1 - y_i(\vec{w} \cdot \vec{x_i} - t)\right)$$

Lagrange Dual function:

$$g(\boldsymbol{\alpha}) = \mathrm{Inf}_{\vec{w}, t} \mathcal{L}(\vec{w}, t, \boldsymbol{\alpha})$$

Dual Problem:

maximize $g(\boldsymbol{\alpha})$

subject to $\boldsymbol{\alpha} \succcurlyeq \mathbf{0}$

Optimization problem

$$\mathcal{L}(\vec{w}, t, \boldsymbol{\alpha}) = \frac{1}{2} \cdot \| \vec{w} \|_2^2 + \sum_{1 \le i \le n} \alpha_i \big(1 - y_i(\vec{w} \cdot \vec{x_i} - t)\big)$$

Primal: $\min_{\vec{w}, t} \max_{\boldsymbol{\alpha} \ge 0} \mathcal{L}(\vec{w}, t, \boldsymbol{\alpha})$ (p*)

Dual: $\max_{\boldsymbol{\alpha} \ge 0} \min_{\vec{w}, t} \mathcal{L}(\vec{w}, t, \boldsymbol{\alpha})$ (d*)

Strong Duality

- If d* = p*, strong duality holds
- Does not hold in general
- *Slater's Theorem*: If convex problem, and strictly feasible point exists, then strong duality holds! (proof too involved, omit here)
- *=> For convex problems, can use dual problem to find solution*

Complementary Slackness

- When strong duality holds

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

(definition)

(since constraints satisfied at x*)

- Sandwiched between $f_0(x)$, last 2 inequalities are equalities!

Complementary Slackness

$$f_0(x^\star) \quad = \quad f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star h_i(x^\star)$$

- Which means: $\sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) = 0$
- Since each term is non-positive, we have *complementary slackness:*

$$\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m.$$

- Whenever constraint is non-active, corresponding multiplier is zero

Complementary Slackness

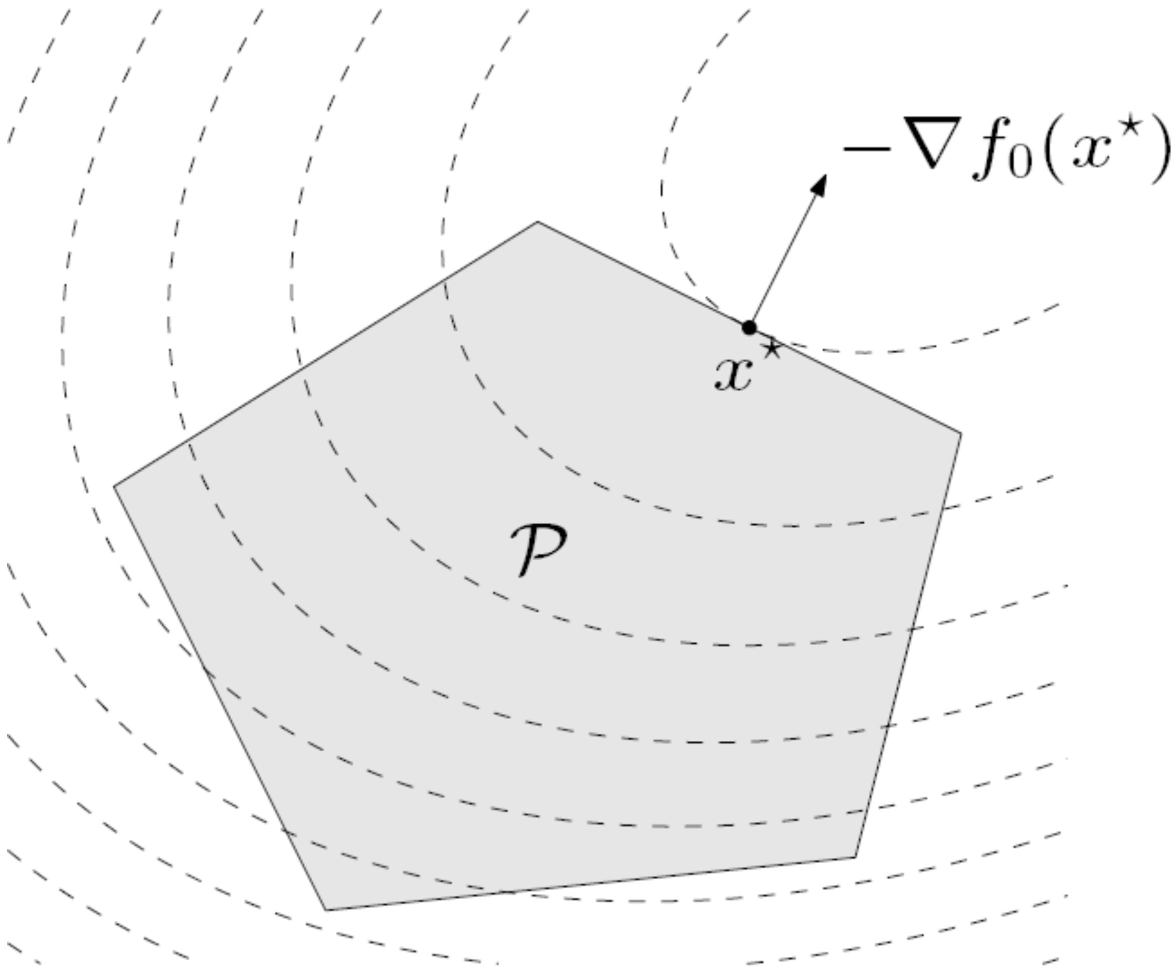
$$\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m.$$

- This can also be described by

$$\lambda_i^\star > 0 \implies f_i(x^\star) = 0,$$
$$f_i(x^\star) < 0 \implies \lambda_i^\star = 0.$$

- Since usually only a few active constraints at solution (see geometry), the dual variable lambda is often *sparse*

QP Visualization

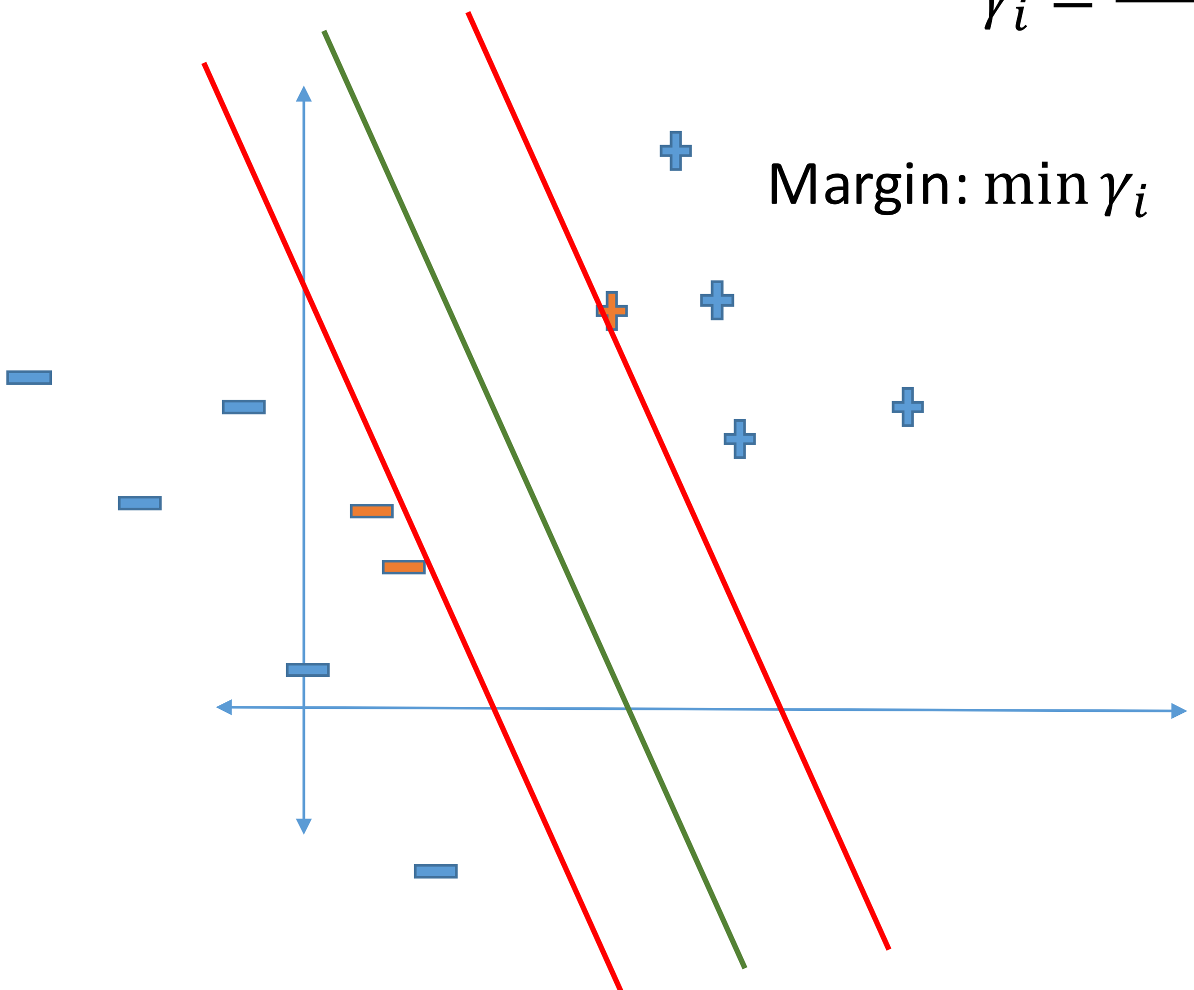


Complementary Slackness

$$\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m.$$

- As we will see, this is why support vector machines result in solution with only key support vectors
  - These come from the dual problem, constraints correspond to points, and complementary slackness ensures only the “active” points are kept

Assume data linearly separable



Distance of the point $\overrightarrow{x_i}$:

$$\gamma_i = \frac{y_i(\overrightarrow{w} \cdot \overrightarrow{x_i} - t)}{\| \overrightarrow{w} \|}$$

Margin: $\min \gamma_i$

KKT Conditions

- The KKT conditions are then just what we call that set of conditions required at the solution (basically list what we know)
- KKT conditions play important role
  - Can sometimes be used to find solution analytically
  - Otherwise can think of many methods as ways of solving KKT conditions

KKT Conditions

- Again given strong duality and assuming differentiable, since $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$ gradient must be 0 at x*

$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star \nabla h_i(x^\star) = 0.$$

- Thus, putting it all together, for non-convex problems we have

KKT Conditions – non-convex

$$
\begin{aligned}
f_i(x^\star) &\leq 0, \quad i = 1, \ldots, m \\
h_i(x^\star) &= 0, \quad i = 1, \ldots, p \\
\lambda_i^\star &\geq 0, \quad i = 1, \ldots, m \\
\lambda_i^\star f_i(x^\star) &= 0, \quad i = 1, \ldots, m \\
\nabla f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^p \nu_i^\star \nabla h_i(x^\star) &= 0,
\end{aligned}
$$

- Necessary conditions

KKT Conditions – convex

$$
\begin{array}{rcll}
f_i(\tilde{x}) & \leq & 0, & i = 1, \ldots, m \\
h_i(\tilde{x}) & = & 0, & i = 1, \ldots, p \\
\tilde{\lambda}_i & \geq & 0, & i = 1, \ldots, m \\
\tilde{\lambda}_i f_i(\tilde{x}) & = & 0, & i = 1, \ldots, m \\
\nabla f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{i=1}^{p} \tilde{\nu}_i \nabla h_i(\tilde{x}) & = & 0, &
\end{array}
$$

then $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu})$ are primal and dual optimal, with zero duality gap.

- Also sufficient conditions:

$$
\begin{aligned}
g(\tilde{\lambda}, \tilde{\nu}) &= L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \\
&= f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(\tilde{x}) + \sum_{i=1}^{p} \tilde{\nu}_i h_i(\tilde{x}) \\
&= f_0(\tilde{x}),
\end{aligned}
$$

(according to the last equation)

Optimization problem

$$\mathcal{L}(\vec{w}, t, \boldsymbol{\alpha}) = \frac{1}{2} \cdot \| \vec{w} \|_2^2 + \sum_{1 \le i \le m} \alpha_i \big(1 - y_i(\vec{w} \cdot \vec{x_i} - t)\big)$$

Take the gradient equal to zero:

$$\nabla_{\vec{w}} \mathcal{L}(\vec{w}, t, \boldsymbol{\alpha}) = 0 \Rightarrow \vec{w} = \sum_{1 \le i \le m} \alpha_i y_i \vec{x_i}.$$

Linear combinations

Optimization problem

Take the gradient equal to zero:

$$\frac{\partial \mathcal{L}(\vec{w}, t, \boldsymbol{\alpha})}{\partial t} = 0$$

gives,

$$\sum_{1 \le i \le m} \alpha_i y_i = 0.$$

Dual Problem

We substitute these values of $\vec{w}, t$.

Maximize:

$$\sum_{1 \le i \le n} \alpha_i - \sum_{1 \le i,j \le n} \alpha_i \alpha_j y_i y_j \left(\vec{x_i} \cdot \vec{x_j}\right)$$

Subject to

$$\sum \alpha_i y_i = 0, \alpha_i \ge 0.$$

Reference

Slides are credited to:

http://www.ittc.ku.edu/bioinfo_seminar/ppt/Summer09/ConvexOpt_July09.ppt

More resources:

1. Stanford Convex Optimization (Boyd & Vandenberghe):

http://ee364a.stanford.edu/lectures/duality.pdf

2. Slater's condition (Wiki):

https://en.wikipedia.org/wiki/Slater%27s_condition