EE306: Signals and Systems II

Lecture 12

The dual problem

* The Lagrange dual problem: maximize $g(\underline{\lambda},\underline{\nu})$ subject to $\lambda \geq 0$

finds the best lower bound on p^*.

- It is a convex optimization pb. $g(\underline{\lambda}^*,\underline{\nu}^*)=d^*$.

→ Weak duality: $d^* \leq p^*$, always holds.

→ Strong duality: $d^* = p^*$, holds for a convex problem: minimize $f_0(\underline{x})$

subject to $f_i(\underline{x}) \leq 0$, $\underline{A}\,\underline{x}=\underline{b}$

if $\exists\, \underline{x} \in \text{int}(D)$ s.t. $f_i(\underline{x})<0$, $\underline{A}\,\underline{x}=\underline{b}$.

These are called Slater's condition.

Ex1 Find the dual problem for the LP primal problem: minimize $\underline{c}^T\underline{x}$ subject to $\underline{A}\,\underline{x} \leq \underline{b}$.

Sol. $L(\underline{x},\underline{\lambda}) = \underline{c}^T\underline{x} + \underline{\lambda}^T(\underline{A}\,\underline{x}-\underline{b}) = -\underline{b}^T\underline{\lambda} + (\underline{c}+\underline{A}^T\underline{\lambda})^T\underline{x}$

$$\Rightarrow g(\underline{\lambda}) = \begin{cases} -\underline{b}^T\underline{\lambda}, & \underline{A}^T\underline{\lambda}+\underline{c}=\underline{0} \\ -\infty, & \text{o.w.} \end{cases}$$

Dual problem: maximize $-\underline{b}^T\underline{\lambda}$

subject to $\underline{A}^T\underline{\lambda}+\underline{c}=\underline{0}$, $\underline{\lambda} \geq \underline{0}$

→ Slater's cond. $d^*=p^*$ if $\exists\, \underline{x}$ s.t. $\underline{A}\,\underline{x}<\underline{b}$.

Karush-Kuhn-Tucker (KKT) conditions

* These are the following 4 conditions:

1. Primal constraints $f_i(\underline{x}) \leq 0$, $i = 1, 2, \cdots, m$
   , $h_i(\underline{x}) = 0$, $i = 1, 2, \cdots, p$.

2. Dual constraints $\underline{\lambda} \geq \underline{0}$.

3. Complementary slackness $\lambda_i f_i(\underline{x}) = 0$, $i = 1, 2, \cdots, m$.

4. Vanishing gradient of Lagrangian w.r.t. $\underline{x}$

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

→ If $\exists\ \underline{x}, \underline{\lambda}, \underline{\nu}$ that satisfy KKT conditions for a convex problem, then they are $\underline{x}^*, \underline{\lambda}^*, \underline{\nu}^*$ and $p^* = f_0(\underline{x}^*) = g(\underline{\lambda}^*, \underline{\nu}^*) = d^*$.

→ Why complementary slackness?

$$d^* = g(\underline{\lambda}^*, \underline{\nu}^*) = \inf_{\underline{x}} \left[f_0(\underline{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\underline{x}) + \sum_{i=1}^{p} \nu_i^* h_i(\underline{x}) \right]$$

$$\leq f_0(\underline{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\underline{x}^*) + \sum_{i=1}^{p} \nu_i^* \underbrace{\cancel{h_i(\underline{x}^*)}}_{0}$$

$$\leq f_0(\underline{x}^*) = p^*.$$

Therefore, for $d^* = p^*$, $\lambda_i^* f_i(\underline{x}^*) = 0$, $i = 1, 2, \cdots, m$.

In other words, $\lambda_i^* > 0 \Rightarrow f_i(\underline{x}^*) = 0$,
$f_i(\underline{x}^*) < 0 \Rightarrow \lambda_i^* = 0$.

Ex 2 Water-filling: We have n channels with signal (noise) powers x_i (α_i), $i = 1, 2, \ldots, n$. We want to maximize the sum rate $\sum_{i=1}^{n} \ln\left(1 + \frac{x_i}{\alpha_i}\right)$, where $\sum_{i=1}^{n} x_i = 1$, $\alpha_i > 0 \ \forall i$.

Sol. Clearly $x_i \geq 0$. Thus, we can re-formulate the optimization problem as: minimize $-\sum_{i=1}^{n} \ln(x_i + \alpha_i)$

subject to $\underline{x} \succeq \underline{0}$, $\underline{1}^T \underline{x} = 1$.

$$L(\underline{x}, \underline{\lambda}, \underline{\nu}) = -\sum_{i=1}^{n} \ln(x_i + \alpha_i) - \sum_{i=1}^{n} \lambda_i x_i + \nu\left(\sum_{i=1}^{n} x_i - 1\right) \Rightarrow$$

$$\nabla_{\underline{x}} L(\underline{x}, \underline{\lambda}, \underline{\nu}) = -\left[\frac{1}{x_1 + \alpha_1} \ \frac{1}{x_2 + \alpha_2} \ \cdots \ \frac{1}{x_n + \alpha_n}\right]^T - \lambda^T + \nu \underline{1}^T.$$

$\Rightarrow \nabla_{\underline{x}} L(\underline{x}, \underline{\lambda}, \underline{\nu}) = 0$ leads to

$$\frac{1}{x_i + \alpha_i} + \lambda_i = \nu, \ \forall i \in \{1, 2, \ldots, n\}.$$

Thus, $\underline{x}$ is optimal if and only if

$\underline{x} \succeq \underline{0}$, $\underline{1}^T \underline{x} = 1$, and $\exists \ \underline{\lambda} \in \mathbb{R}^n$, $\nu \in \mathbb{R}$ such that

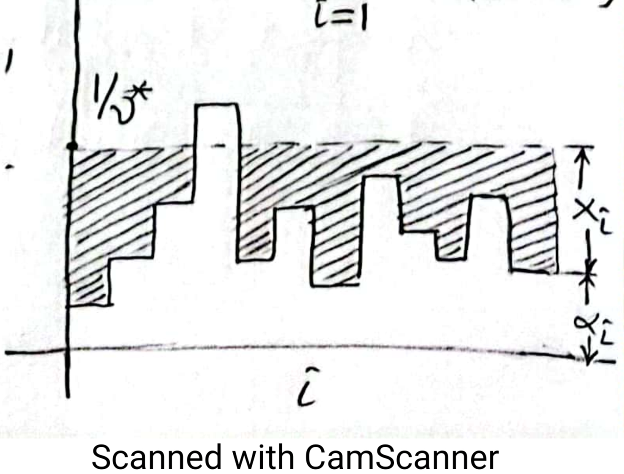
$$\underline{\lambda} \succeq \underline{0}, \ \underbrace{\lambda_i x_i = 0}, \ \frac{1}{x_i + \alpha_i} + \lambda_i = \nu, \ \forall i.$$

$\rightarrow$ Complementary slackness:

$\lambda_i = 0 \Rightarrow x_i = \frac{1}{\nu} - \alpha_i$, $\nu < \frac{1}{\alpha_i}$,

$x_i = 0 \Rightarrow \lambda_i = \nu - \frac{1}{\alpha_i}$, $\nu \geq \frac{1}{\alpha_i}$.

$$\underline{1}^T \underline{x} = 1 = \sum_{i=1}^{n} \max\left(\frac{1}{\nu} - \alpha_i, 0\right)$$



Q How can we interpret this optimization result?

Scanned with CamScanner

Descent methods (minimization)

* $\underline{x}^{(k+1)} = \underline{x}^{(k)} + t^{(k)} \Delta \underline{x}^{(k)}$ with $f(\underline{x}^{(k+1)}) < f(\underline{x}^{(k)})$.

Other notation: $\underline{x}^{+} = \underline{x} + t\Delta \underline{x}$, $\underline{x} := \underline{x} + t\Delta \underline{x}$.

→ $\Delta \underline{x}$ is the step or search direction.
t is the step size or step length.

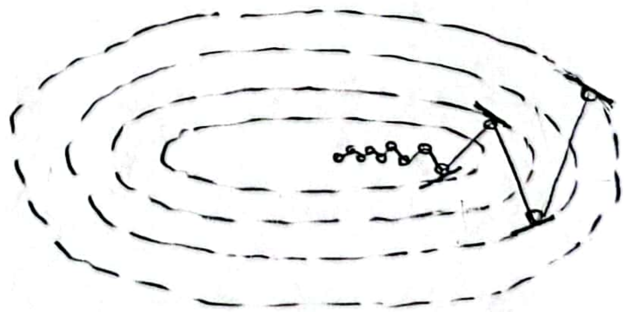
→ For convex functions $f(\underline{x}^{+}) < f(\underline{x})$ implies
$\nabla_{\underline{x}} f(\underline{x})^T \Delta \underline{x} < 0 \Rightarrow \Delta \underline{x}$ is the descent direction.

Gradient descent algorithm

* It is a general descent method with $\Delta \underline{x} = -\nabla_{\underline{x}} f(\underline{x})$.

<u>given</u> a starting point $x \in D$.
<u>repeat</u> <u>1.</u> $\Delta \underline{x} := -\nabla_{\underline{x}} f(\underline{x})$.
<u>2.</u> Choose step size t (line search).
<u>3.</u> Update $\underline{x} := \underline{x} + t\Delta \underline{x}$.
<u>until</u> stopping criterion is satisfied
e.g., $\| f(\underline{x}^{+}) - f(\underline{x}) \|_2 < \varepsilon$ or $\| \nabla_{\underline{x}} f(\underline{x}) \|_2 < \varepsilon$.

* Other descent algorithms:
Steepest descent.
Stochastic gradient descent.



Gradient descent