Geometry of Duality: Duality Gap and Convexity

- With reference to Figure 20, if the set $\mathcal{I}$ is not convex, there could be a **gap** between the $z-$intercept $(\mathbf{0}, \alpha_1)$ of the best supporting hyperplane $\mathcal{H}_{\lambda_1,\alpha_1}$ and the closest point $(\mathbf{0}, \delta_1)$ of $\mathcal{I}$ on the $z-$axis (solution to the primal).
- For non-convex $\mathcal{I}$, we can never prove in zero duality gap in general.
- Homework (Quiz 1, Problem 1): Write dual for constrained problem $\min_x \ f(x) = 5x^2 + 6x^3 - x^4$ on the closed interval $[-2, 10]$. Does it have a duality gap?

Geometry of Duality: Duality Gap and Convexity

- With reference to Figure 20, if the set $\mathcal{I}$ is not convex, there could be a **gap** between the $z-$intercept $(\mathbf{0}, \alpha_1)$ of the best supporting hyperplane $\mathcal{H}_{\lambda_1,\alpha_1}$ and the closest point $(\mathbf{0}, \delta_1)$ of $\mathcal{I}$ on the $z-$axis (solution to the primal).
- For non-convex $\mathcal{I}$, we can never prove in zero duality gap in general.
- Homework (Quiz 1, Problem 1): Write dual for constrained problem $\min_x f(x) = 5x^2 + 6x^3 - x^4$ on the closed interval $[-2, 10]$. Does it have a duality gap?
- For well-behaved convex functions (as in the case of linear programming),

Geometry of Duality: Duality Gap and Convexity

- With reference to Figure 20, if the set $\mathcal{I}$ is not convex, there could be a **gap** between the $z-$intercept $(\mathbf{0}, \alpha_1)$ of the best supporting hyperplane $\mathcal{H}_{\lambda_1,\alpha_1}$ and the closest point $(\mathbf{0}, \delta_1)$ of $\mathcal{I}$ on the $z-$axis (solution to the primal).
- For non-convex $\mathcal{I}$, we can never prove in zero duality gap in general.
- Homework (Quiz 1, Problem 1): Write dual for constrained problem $\min_x \; f(x) = 5x^2 + 6x^3 - x^4$ on the closed interval $[-2, 10]$. Does it have a duality gap?
- For well-behaved convex functions (as in the case of linear programming), there are no duality gaps. Figure 31 illustrates the case of a well-behaved convex program.

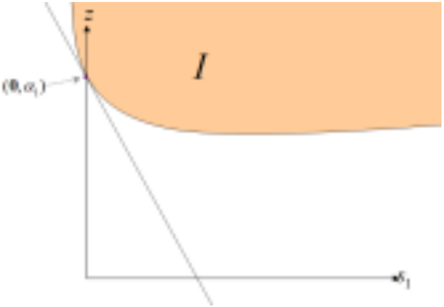

Figure 29: Example of the convex set $\mathcal{I}$ and hyperplanes $\mathcal{H}_{\lambda,\alpha}$ for a single constrained well-behaved convex program.

Non-convexity and Duality Gap: H/W Illustration

Find all the points of local and global minima and maxima as well as any saddle point(s) of the function $f(x) = 5x^2 + 6x^3 - x^4$ on the closed interval $[-2, 10]$.

Solution:

Setting the derivative of $f(x)$ to 0, we first find all the critical points.

$$f'(x) = 10x + 18x^2 - 4x^3 = 0$$

Factorizing

$$2(5 - x)(1 + 2x)x = 0$$

The critical points of this function are $-1/2, 0$ and 5.

Differentiating once more

$$f''(x) = 10 + 36x - 12x^2 = 0$$

Non-convexity and Duality Gap: H/W Illustration

One can easily verify that $f''(-1/2) < 0$, $f''(0) > 10$ and $f''(5) < 0$. Thus, we have atleast a local maximum (concave region) at $-1/2$ and 5 and a local minimum (convex region) at 0. As for global maximum, we can simply evaluate the function at the three critical points as well as at the extreme points and report. $f(-2) = -44$, $f(-1/2) = 0$, $f(5) = 250$, $f(10) = -3500$. Thus, we have a global (and local) maximum at 5, global minimum at 10 and local maximum at $-1/2$ and local minimum at 0.

Non-convexity and Duality Gap: H/W Illustration

- To derive its dual (for $\lambda_1, \lambda_2 \geq 0$):

$$L^*(\lambda_1, \lambda_2) = \min_{x \in [-2,10]} 5x^2 + 6x^3 - x^4 + (x+2)\lambda_1 + (-x+10)\lambda_2$$

  Setting derivative wrt x to be 0 $2(5-x)(1+2x)x = \lambda_2 - \lambda_1$

- Plotting L^* (and/or the first order necessary condition for different values of λ_1 and λ_2), we find that the min in the interval $[-2, 10]$ is always either at -2 or at 10 (based on nature of λ_1 and λ_2)

$$L^*(\lambda_1, \lambda_2) = min(20+48-16+8\lambda_2, 500+6000-10000+12\lambda_1) = min(52+8\lambda_2, -3500+12\lambda_1)$$

- That is, if $3\lambda_1 - 2\lambda_2 > 3552/4 = 888$ then $L^*(\lambda_1, \lambda_2) = -3500 + 12\lambda_1$ else $L^*(\lambda_1, \lambda_2) = 52 + 8\lambda_2$.
- L^* is piecewise linear and concave. Dual optimization problem is about maximizing L^* wrt $\lambda_1, \lambda_2 \geq 0$. We expect duality gap.

Geometry of Duality: Duality Gap even with Convexity

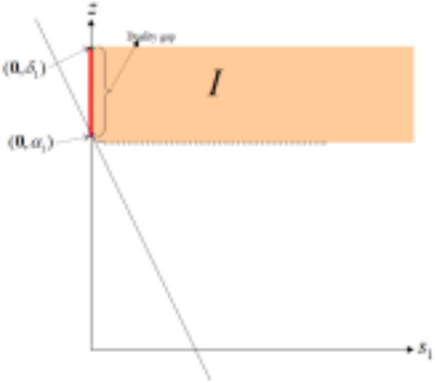

Figure 30: Example of the convex set $\mathcal{I}$ and hyperplanes $\mathcal{H}_{\lambda,\alpha}$ for a single constrained semi-definite program.

Recap: Max cut as an SDP

- And even when the set $\mathcal{I}$ is convex, bizzaire things can happen; for example, in the case of semi-definite programming, the set $\mathcal{I}$, though convex, is not at all well-behaved and this yields a large duality gap, as shown in Figure 30.
- In fact, the set $\mathcal{I}$ is **open from below (the dotted boundary) for a semi-definite program**. We could create very simple problems with convex $\mathcal{I}$, for which there are duality gaps.

On the other hand, non-convex problems such as SVD can have 0 duality gap

Bringing Things Together: Zero Duality Gap, Differentiability, **Necessity of KKT Conditions**

Zero Duality Gap, Differentiability ⇒ KKT Conditions (contd.)

- Consider the following general optimization problem.

$$
\begin{array}{lll}
\min\limits_{\mathbf{x}\in\mathcal{D}} & f(\mathbf{x}) & \\
\text{subject to} & g_i(\mathbf{x}) \le 0, & i = 1,\ldots,m \\
 & h_j(\mathbf{x}) = 0, & j = 1,\ldots,p \\
\text{variable } \mathbf{x} = (x_1,\ldots,x_n) & &
\end{array}
$$

- Suppose that the primal and dual optimal values for the above problem are attained and equal, that is, strong duality holds. Let $\widehat{\mathbf{x}}$ be a primal optimal and $(\widehat{\lambda},\widehat{\mu})$ be a dual optimal point ($\widehat{\lambda} \in \Re^m, \widehat{\mu} \in \Re^p$). **Thus....**

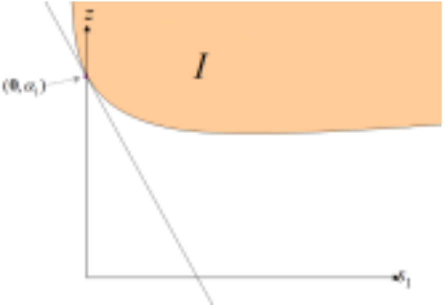

Figure 31: Example of the convex set $\mathcal{I}$ and hyperplanes $\mathcal{H}_{\lambda,\alpha}$ for a single constrained well-behaved convex program.

Zero Duality Gap, Differentiability ⇒ KKT Conditions (contd.)

$$\begin{aligned}
f(\widehat{\mathbf{x}}) &= L^*(\widehat{\lambda}, \widehat{\mu}) \\
&= \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) + \widehat{\lambda}^T \mathbf{g}(\mathbf{x}) + \widehat{\mu}^T \mathbf{h}(\mathbf{x}) \\
&\le f(\widehat{\mathbf{x}}) + \widehat{\lambda}^T \mathbf{g}(\widehat{\mathbf{x}}) + \widehat{\mu}^T \mathbf{h}(\widehat{\mathbf{x}}) \\
&\le f(\widehat{\mathbf{x}})
\end{aligned}$$

- The last inequality follows from the fact that $\widehat{\lambda} \geq \mathbf{0}$, $\mathbf{g}(\widehat{\mathbf{x}}) \leq \mathbf{0}$, and $\mathbf{h}(\widehat{\mathbf{x}}) = \mathbf{0}$.
- We can therefore conclude that the two inequalities must be equalities

Zero Duality Gap, Differentiability ⇒ KKT Conditions (contd.)

$$\begin{aligned} f(\widehat{\mathbf{x}}) &= L^*(\widehat{\lambda}, \widehat{\mu}) \\ &= \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) + \widehat{\lambda}^T\mathbf{g}(\mathbf{x}) + \widehat{\mu}^T\mathbf{h}(\mathbf{x}) \\ &\le f(\widehat{\mathbf{x}}) + \widehat{\lambda}^T\mathbf{g}(\widehat{\mathbf{x}}) + \widehat{\mu}^T\mathbf{h}(\widehat{\mathbf{x}}) \\ &\le f(\widehat{\mathbf{x}}) \end{aligned}$$

- The last inequality follows from the fact that $\widehat{\lambda} \ge \mathbf{0}$, $\mathbf{g}(\widehat{\mathbf{x}}) \le \mathbf{0}$, and $\mathbf{h}(\widehat{\mathbf{x}}) = \mathbf{0}$.
- We can therefore conclude that the two inequalities in this chain must hold with equality.
- Conclusions from this chain of equalities (continued on next slide):

Zero Duality Gap, Differentiability ⇒ KKT Conditions (contd.)

$$f(\widehat{\mathbf{x}}) \quad = f(\widehat{\mathbf{x}}) + \widehat{\lambda}^T \mathbf{g}(\widehat{\mathbf{x}}) + \widehat{\mu}^T \mathbf{h}(\widehat{\mathbf{x}})$$

1. That $\widehat{\mathbf{x}}$ is a minimizer for $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ over $\mathbf{x} \in \mathcal{D}$. In particular, if the functions f, $g_1, g_2, \ldots, g_m$ and $h_1, h_2, \ldots, h_p$ are differentiable (and therefore have open domains),

gradient wrt x at x-hat must vanish

Zero Duality Gap, Differentiability ⇒ KKT Conditions (contd.)

$$f(\widehat{\mathbf{x}}) \quad = f(\widehat{\mathbf{x}}) + \widehat{\lambda}^T \mathbf{g}(\widehat{\mathbf{x}}) + \widehat{\mu}^T \mathbf{h}(\widehat{\mathbf{x}})$$

1. That $\widehat{\mathbf{x}}$ is a minimizer for $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ over $\mathbf{x} \in \mathcal{D}$. In particular, if the functions f, $g_1, g_2, \ldots, g_m$ and $h_1, h_2, \ldots, h_p$ are differentiable (and therefore have open domains), the gradient of $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ must vanish at $\widehat{\mathbf{x}}$, since any point of global optimum must be a point of local optimum. That is, $\nabla f(\widehat{\mathbf{x}}) + \sum_{i=1}^{m} \widehat{\lambda}_i \nabla g_i(\widehat{\mathbf{x}}) + \sum_{j=1}^{p} \widehat{\mu}_j \nabla h_j(\widehat{\mathbf{x}}) = \mathbf{0}$
2. That $\widehat{\lambda}^T \mathbf{g}(\widehat{\mathbf{x}}) = \sum_{i=1}^{n} \widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) = 0$:

Rest are straightforward (Primal and dual feasibility conditions)

Zero Duality Gap, Differentiability ⇒ KKT Conditions (contd.)

$$f(\widehat{\mathbf{x}}) \quad = f(\widehat{\mathbf{x}}) + \widehat{\lambda}^T\mathbf{g}(\widehat{\mathbf{x}}) + \widehat{\mu}^T\mathbf{h}(\widehat{\mathbf{x}})$$

1. That $\widehat{\mathbf{x}}$ is a minimizer for $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ over $\mathbf{x} \in \mathcal{D}$. In particular, if the functions f, $g_1, g_2, \ldots, g_m$ and $h_1, h_2, \ldots, h_p$ are differentiable (and therefore have open domains), the gradient of $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ must vanish at $\widehat{\mathbf{x}}$, since any point of global optimum must be a point of local optimum. That is, $\nabla f(\widehat{\mathbf{x}}) + \sum_{i=1}^{m} \widehat{\lambda}_i \nabla g_i(\widehat{\mathbf{x}}) + \sum_{j=1}^{p} \widehat{\mu}_j \nabla h_j(\widehat{\mathbf{x}}) = \mathbf{0}$
2. That $\widehat{\lambda}^T\mathbf{g}(\widehat{\mathbf{x}}) = \sum_{i=1}^{n} \widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) = 0$: Since each term in this sum is nonpositive, we conclude that

Zero Duality Gap, Differentiability ⇒ KKT Conditions (contd.)

$$f(\widehat{\mathbf{x}}) \quad = f(\widehat{\mathbf{x}}) + \widehat{\lambda}^T\mathbf{g}(\widehat{\mathbf{x}}) + \widehat{\mu}^T\mathbf{h}(\widehat{\mathbf{x}})$$

1. That $\widehat{\mathbf{x}}$ is a minimizer for $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ over $\mathbf{x} \in \mathcal{D}$. In particular, if the functions f, $g_1, g_2, \ldots, g_m$ and $h_1, h_2, \ldots, h_p$ are differentiable (and therefore have open domains), the gradient of $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ must vanish at $\widehat{\mathbf{x}}$, since any point of global optimum must be a point of local optimum. That is, $\nabla f(\widehat{\mathbf{x}}) + \sum_{i=1}^{m} \widehat{\lambda}_i \nabla g_i(\widehat{\mathbf{x}}) + \sum_{j=1}^{p} \widehat{\mu}_j \nabla h_j(\widehat{\mathbf{x}}) = \mathbf{0}$
2. That $\widehat{\lambda}^T\mathbf{g}(\widehat{\mathbf{x}}) = \sum_{i=1}^{n} \widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) = 0$: Since each term in this sum is nonpositive, we conclude that $\widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) = 0$ *for* $i = 1, 2, \ldots, m$. This condition is called *complementary slackness* and is a necessary condition for strong duality.
   - Complementary slackness implies that the i^{th} optimal lagrange multiplier is 0 unless the i^{th} inequality constraint is active at the optimum. That is,

$$\begin{aligned} \widehat{\lambda}_i > 0 \quad &\Rightarrow \quad g_i(\widehat{\mathbf{x}}) = 0 \\ g_i(\widehat{\mathbf{x}}) < 0 \quad &\Rightarrow \quad \widehat{\lambda}_i = 0 \end{aligned}$$

Zero Duality Gap, Differentiability ⇒ KKT Conditions (contd.)

Putting together these conditions along with the feasibility conditions for any primal solution and dual solution, we can state the following Karush-Kuhn-Tucker (KKT) necessary conditions for zero duality gap:

$$\begin{array}{lrcll}
(1) & \nabla f(\widehat{\mathbf{x}}) + \sum_{i=1}^{m} \widehat{\lambda}_i \nabla g_i(\widehat{\mathbf{x}}) + \sum_{j=1}^{p} \widehat{\mu}_j \nabla h_j(\widehat{\mathbf{x}}) & = & \mathbf{0} & \\
(2) & g_i(\widehat{\mathbf{x}}) & \leq & 0 & i = 1, 2, \ldots, m \\
(3) & \widehat{\lambda}_i & \geq & 0 & i = 1, 2, \ldots, m \\
(4) & \widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) & = & 0 & i = 1, 2, \ldots, m \\
(5) & h_j(\widehat{\mathbf{x}}) & = & 0 & j = 1, 2, \ldots, p
\end{array} \qquad (84)$$

Bringing Things Together: **Sufficiency of KKT Conditions**, Convexity, Differentiability, Zero Duality Gap

KKT Conditions, Convexity and Differentiability ⇒ Zero Duality Gap

Theorem

If the function f is convex, g_i are convex and h_j are affine, then KKT conditions in (84) are necessary and sufficient conditions for zero duality gap.

Proof: The necessity part has already been proved; here we only prove the sufficiency part. The conditions (2) and (5) in (84) ensure that $\widehat{\mathbf{x}}$ is primal feasible. Since $\lambda \geq \mathbf{0}$, $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ is convex in $\mathbf{x}$. Based on condition (1) in (84) and sufficient condition for global minimum of a convex function, we can infer that $\widehat{\mathbf{x}}$ minimizes L(x,lambda-hat,mu-hat)
since gradient vanishing is a
sufficient condition for global min
of a convex function

KKT Conditions, Convexity and Differentiability ⇒ Zero Duality Gap

Theorem

If the function f is convex, g_i are convex and h_j are affine, then KKT conditions in (84) are necessary and sufficient conditions for zero duality gap.

Proof: The necessity part has already been proved; here we only prove the sufficiency part. The conditions (2) and (5) in (84) ensure that $\widehat{\mathbf{x}}$ is primal feasible. Since $\lambda \geq \mathbf{0}$, $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ is convex in $\mathbf{x}$. Based on condition (1) in (84) and sufficient condition for global minimum of a convex function, we can infer that $\widehat{\mathbf{x}}$ minimizes $\underline{L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})}$. We can thus conclude that

KKT Conditions, Convexity and Differentiability ⇒ Zero Duality Gap

Theorem

If the function f is convex, g_i are convex and h_j are affine, then KKT conditions in (84) are necessary and sufficient conditions for zero duality gap.

Proof: The necessity part has already been proved; here we only prove the sufficiency part. The conditions (2) and (5) in (84) ensure that $\widehat{\mathbf{x}}$ is primal feasible. Since $\lambda \geq \mathbf{0}$, $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$ is convex in $\mathbf{x}$. Based on condition (1) in (84) and sufficient condition for global minimum of a convex function, we can infer that $\widehat{\mathbf{x}}$ minimizes $L(\mathbf{x}, \widehat{\lambda}, \widehat{\mu})$. We can thus conclude that

$$\boxed{\begin{aligned} L^*(\widehat{\lambda}, \widehat{\mu}) &= f(\widehat{\mathbf{x}}) + \widehat{\lambda}^T \mathbf{g}(\widehat{\mathbf{x}}) + \widehat{\mu}^T \mathbf{h}(\widehat{\mathbf{x}}) \\ &= f(\widehat{\mathbf{x}}) \end{aligned}}$$

In the equality above, we use $\underline{h_j(\widehat{\mathbf{x}}) = 0 \text{ and } \widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) = 0.}$

KKT Conditions, Convexity and Differentiability $\Rightarrow$ Zero Duality Gap (contd.)

Further, given the relation between d^* and $L^*(\widehat{\lambda}, \widehat{\mu})$ and between $f(\widehat{\mathbf{x}})$ and p^*

KKT Conditions, Convexity and Differentiability ⇒ Zero Duality Gap (contd.)

Further, given the relation between d^* and $L^*(\widehat{\lambda}, \widehat{\mu})$ and between $f(\widehat{\mathbf{x}})$ and p^*

$$\underline{d^* \geq L^*(\widehat{\lambda}, \widehat{\mu}) = f(\widehat{\mathbf{x}}) \geq p^*}$$

The weak duality theorem however states that

d* <= p*

Putting together these inequalities yields x-hat as point of primal optimality and lambda-hat, mu-hat as point of dual optimality

KKT Conditions, Convexity and Differentiability ⇒ Zero Duality Gap (contd.)

Further, given the relation between d^* and $L^*(\widehat{\lambda}, \widehat{\mu})$ and between $f(\widehat{\mathbf{x}})$ and p^*

$$d^* \geq L^*(\widehat{\lambda}, \widehat{\mu}) = f(\widehat{\mathbf{x}}) \geq p^*$$

The weak duality theorem however states that $p^* \geq d^*$. This implies that

$$d^* = L^*(\widehat{\lambda}, \widehat{\mu}) = f(\widehat{\mathbf{x}}) = p^*$$

This shows that $\widehat{\mathbf{x}}$ and $(\widehat{\lambda}, \widehat{\mu})$ correspond to the primal and dual optimals respectively and the problem therefore has zero duality gap. □

KKT Conditions, Convexity and Differentiability $\Rightarrow$ Zero Duality Gap (contd.)

- In summary, for any convex optimization problem with differentiable objective and constraint functions, any points that satisfy the KKT conditions are primal and dual optimal, and have zero duality gap.
- The KKT conditions play a very important role in optimization. In some rare cases, it is possible to solve the optimization problems by finding a solution to the KKT conditions analytically.
- Many algorithms for convex optimization are conceived as, or can be interpreted as, methods for solving the KKT conditions.

Often difference between f(x^k) (>= p*) and L(lambda^k,mu^k) (<= d*) is used as an estimate for how far we are from the optimal solution (where we expect zero duality gap)

Eg: Dual Ascent algo, ADMM, Interior point methods

Bringing Things Together: Convexity, **Constaint Qualifications**, Zero Duality Gap

Even with convex functions and constraints, there are other sufficient conditions for strong duality!

Convexity, Slater's Constraint Qualification ⇒ Zero Duality Gap

Slater's sufficient condition is in terms of strict feasibility

Recap the result on Weak Duality

$$p^* = \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) \geq \min_{\lambda\geq\mathbf{0}} L^*(\lambda) = d^*$$

The weak duality theorem has some important implications.

(Eg: linear program with trivial constraints)

- If the primal problem is unbounded below, that is, $p^* = -\infty$, we must have $d^* = -\infty$, which means that the Lagrange dual problem is infeasible.

(Eg: Again trivial LP)

- Conversely, if the dual problem is unbounded above, that is, $d^* = \infty$, we must have $p^* = \infty$, which is equivalent to saying that the primal problem is infeasible. The difference, $p^* - d^*$ is called the duality gap.
- In many hard combinatorial optimization problems with duality gaps, we get good dual solutions, which tell us that we are guaranteed of being some k % within the optimal solution to the primal, for some satisfactorily low values of k. This is one of the powerful uses of duality theory; constructing bounds for optimization problems.

Convexity, Slater's Constraint Qualification ⇒ Zero Duality Gap (contd.)

Under what other conditions can one assert that strong duality ($d^* = p^*$) holds?

- It usually holds for convex problems but there are exceptions to that - one of the most typical being that of the semi-definite optimization problem. The semi-definite program (SDP) is defined, with the linear matrix inequality constraint as follows:

$$\begin{aligned} \min_{\mathbf{x}\in\Re^n} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & x_1A_1 + \ldots + x_nA_n + G \underline{\preceq} 0 \\ & A\mathbf{x} = \mathbf{b} \end{aligned} \tag{85}$$

linear objective with semi-definite conic constraint

- Sufficient conditions for strong duality in convex problems are called *constraint qualifications*. One of the most useful sufficient conditions for strong duality is called the *Slater's constraint qualification* **(requires separating hyperplane theorem).**

In terms of the interior of the feasible region being non-empty
Proof comes from strong separation hyper plane theorem

Convexity, Slater's Constraint Qualification ⇒ Zero Duality Gap (contd.)

Definition

[Slater's constraint qualification]: For a convex problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathcal{D}} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & A\mathbf{x} = \mathbf{b} \\ \text{variable } \mathbf{x} = (x_1, \ldots, x_n) & \end{array} \tag{86}$$

strong duality holds (that is $d^* = p^*$) if it is *strictly feasible*. That is,

$$\boxed{\exists \mathbf{x} \in int(\mathcal{D}) : \quad g_i(\mathbf{x}) < 0 \quad i = 1, 2, \ldots, m \quad A\mathbf{x} = \mathbf{b}}$$

However, if any g_i is linear, it need not hold with strict inequality.

Separating hyperplane theorem (also see additional optional notes)

If $\mathcal{C}$ and $\mathcal{D}$ are disjoint convex sets, *i.e.*, $\mathcal{C} \cap \mathcal{D} = \phi$, then there exists $\mathbf{a} \neq \mathbf{0}$, with a $b \in \Re$ such that
$\mathbf{a}^T\mathbf{x} \leq \mathbf{b}$ for $\mathbf{x} \in \mathcal{C}$,
$\mathbf{a}^T\mathbf{x} \geq \mathbf{b}$ for $\mathbf{x} \in \mathcal{D}$.
That is, the hyperplane $\left\{\mathbf{x}|\mathbf{a}^T\mathbf{x} = \mathbf{b}\right\}$ separates $\mathcal{C}$ and $\mathcal{D}$.

- The seperating hyperplane need not be unique though.
- Strict separation requires additional assumptions (e.g., C is closed, D is a singleton).

Farka's lemma, Theorem of alternatives

Simpler Case of Slater's Constraint Qualification with Analysis (also see additional optional notes)

- Strong Duality for Conic Programs (CP) as generalization of Linear Programs (LP) illustrated through additional concepts such as (i) Proper Cones, (ii) use of Farkas' Lemma (theorem of alternatives as in complementary slackness) in 2015 offering of this course[12]
- Trajectory was: Linear program (LP) $\Longrightarrow$ weak duality $\Longrightarrow$ Dual LP $\Longrightarrow$ Generalized inequality $\Longrightarrow$ Proper cones $\Longrightarrow$ Conic Program (CP) $\Longrightarrow$ Weak duality for CP $\Longrightarrow$ Dual for CP using dual cone (Semi-definite program and LP are special cases of Conic Programs)
- Detailed discussion on Strong Duality for Conic Programs presented by Nemirovski[13]

[12] Read until lecture 10 on 13.2.2015 of `https://www.cse.iitb.ac.in/~cs709/2015a/calendar.html`.

[13] `http://www2.isye.gatech.edu/~nemirovs/ICMNemirovski.pdf`

Convexity, KKT Conditions, Slater's Constraint Qualification, Zero Duality Gap

Table 1 summarizes some optimization problems, their duals and conditions for strong duality.

Problem type	**Objective Function**	**Constraints**	$L^*(\lambda)$	**Dual constraints**	**Strong duality**
Linear Program	$\mathbf{c}^T\mathbf{x}$	$A\mathbf{x} \leq \mathbf{b}$	$-\mathbf{b}^T\lambda$	$A^T\lambda + \mathbf{c} = \mathbf{0}$ $\lambda \geq \mathbf{0}$	Feasible primal and dual
Quadratic Program	$\frac{1}{2}\mathbf{x}^TQ\mathbf{x} + \mathbf{c}^T\mathbf{x}$ for $Q \in \mathcal{S}^n_{++}$	$A\mathbf{x} \leq \mathbf{b}$	$-\frac{1}{2}\left(\mathbf{c} - A^T\lambda\right)^T Q^{-1}\left(\mathbf{c} - A^T\lambda\right) + \mathbf{b}^T\lambda$	$\lambda \geq \mathbf{0}$	Always
Entropy maximization	$x_i\sum_{i=1}^n \ln x_i$	$A\mathbf{x} \leq \mathbf{b}$ $\mathbf{x}^T\mathbf{1} = 1$	$-\mathbf{b}^T\lambda - \mu - e^{-\mu-1}\sum_{i=1}^n e^{-\mathbf{a}_i^T\lambda}$ $\mathbf{a}_i$ is the i^{th} column of A	$\lambda \geq \mathbf{0}$	Primal constraints are satisfied.

Table 1: Examples of functions and their duals.

1) Read extra notes on Derivation of Duals for Support Vector Classification and Regression (Strong Duality): https://www.cse.iitb.ac.in/~cs709/notes/enotes/svr-kkt-dual-derivation.pdf
2) See algorithms for SVM primal and dual at
http://pages.cs.wisc.edu/~swright/talks/sjw-complearning.pdf

Homework: The general SLBQP problem

- We define the general SLBQP (**S**ingle **L**inear equality constrained **B**ounded **Q**uadratic **P**rogram) problem as

$$\min f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{c}^T\mathbf{x}$$

  s.t.

  - $l_i \leq x_i \leq u_i, \forall i$
  - $\mathbf{a}^T\mathbf{x} = b$

  These constraints form the non-empty closed convex set $\mathcal{C}$
- What about the dual function of SLBQP? Will the duality gap be zero?
- Projection methods can solve bounded constrained optimization problems with large changes in the working set of constraints at each iteration.
- Are their other algorithms motivated by Lagrange Duality Theory?

We have already discussed projected and proximal gradient descent capable of handling constraints. Any others?

- We will now look at other methods
  - Inspired by dual
  - Consider constraints (start with simple linear constraints $A\mathbf{x} = \mathbf{b}$)
- Interior point methods
  - Make use of barrier function (such as logarithmic barrier; recall variant of Linear Program discussed in last lecture), and
  - Convergence analyzed through gap using dual ($\frac{m}{t}$)

Dual Ascent and ADMM

Dual ascent

- Consider

$$\min_{\mathbf{x}} f(\mathbf{x})$$

$$\text{s.t. } A\mathbf{x} = b$$

- We have
  - $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda^\top (A\mathbf{x} - b)$
  - $L^*(\lambda) = \inf_{\mathbf{x}} L(\mathbf{x}, \lambda)$
    (under strong duality, infimum is attained)
  - Recapping definition of the convex conjugate function $f^*(\mathbf{h}) = \sup_{\mathbf{x}} \mathbf{h}^T\mathbf{x} - f(\mathbf{x})$

    $L^*(\lambda) = \inf_{\mathbf{x}} L(\mathbf{x}, \lambda) =$

- Consider

$$\min_{\mathbf{x}} f(\mathbf{x})$$

$$\text{s.t. } A\mathbf{x} = b$$

- We have
  - $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda^\top (A\mathbf{x} - b)$
  - $L^*(\lambda) = \inf_{\mathbf{x}} L(\mathbf{x}, \lambda)$
    (under strong duality, infimum is attained)
  - Recapping definition of the convex conjugate function $f^*(\mathbf{h}) = \sup_{\mathbf{x}} \mathbf{h}^T\mathbf{x} - f(\mathbf{x})$

    $L^*(\lambda) = \inf_{\mathbf{x}} L(\mathbf{x}, \lambda) = \underline{-f^*(-A^T\lambda) - \mathbf{b}^T\lambda}$

Idea of dual ascent

1. Initialize $\lambda^{(0)}$
2. Iteratively
    1. $\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}} L(\mathbf{x}, \lambda^k)$
    2. Gradient ascent for dual maximization problem: $d^* = \max_{\lambda \geq 0} L^*(\lambda)$...approximated as

[14] There are other algorithms such as cutting plane algorithm that also work for non-differentiable dual.

Idea of dual ascent

1. Initialize $\lambda^{(0)}$
2. Iteratively
   1. $\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}} L(\mathbf{x}, \lambda^k)$ we have discussed several strategies for x including proxir
   2. Gradient ascent for dual maximization problem: $d^* = \max_{\lambda \geq 0} L^*(\lambda)$...approximated as
      - ★ $d^* = \max_{\lambda \geq 0} L(\mathbf{x}^{k+1}, \lambda)$
      - ★ $\lambda^{k+1} = \lambda^k + t^k \partial_\lambda \left(f(\mathbf{x}^{k+1}) + \lambda^\top (A\mathbf{x}^{k+1} - b) \right)$

        (sub)gradient ascent

[14] There are other algorithms such as cutting plane algorithm that also work for non-differentiable dual.

Idea of dual ascent

1. Initialize $\lambda^{(0)}$
2. Iteratively
    1. $\mathbf{x}^{k+1} = \underset{\mathbf{x}}{\text{argmin}}\, L(\mathbf{x}, \lambda^k)$
    2. Gradient ascent for dual maximization problem: $d^* = \max_{\lambda \geq 0} L^*(\lambda)$...approximated as
        - ⋆ $d^* = \max_{\lambda \geq 0} L(\mathbf{x}^{k+1}, \lambda)$
        - ⋆ $\lambda^{k+1} = \lambda^k + t^k \partial_\lambda \left(f(\mathbf{x}^{k+1}) + \lambda^\top (A\mathbf{x}^{k+1} - b) \right)$
          $= \lambda^k + t^k (A\mathbf{x}^{k+1} - b)$
        - ⋆ Leads to convergence (under assumptions of strong convexity etc) even if the Lagrange dual $L^*(\lambda)$ is non-differentiable[14].

[14] There are other algorithms such as cutting plane algorithm that also work for non-differentiable dual.

- If λ converges to $\lambda^* = \underset{\lambda}{\text{argmax}}\, L^*(\lambda)$ and strong duality holds, *i.e.*

$$\min_{\mathbf{x}} f(\mathbf{x}) = \max_{\lambda \geq 0} L^*(\lambda)$$

$$\text{s.t. } A\mathbf{x} = b$$

  then,

$$\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmin}}\, L(\mathbf{x}, \lambda^*)$$

- If f is **strongly convex with constant** m, and you ensure $t^k \leq m$, then convergence rate is $O\left(\frac{1}{k}\right)$.

Dual decomposition

- $f(\mathbf{x})$ is decomposable into v blocks of variables (such as in Machine Learning, with decomposition over examples)

Dual decomposition

- $f(\mathbf{x})$ is decomposable into v blocks of variables (such as in Machine Learning, with decomposition over examples)

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_v} \sum_{i=1}^{v} f_i(\mathbf{x}_i)$$

$$\text{s.t. } A\mathbf{x} = b$$

- Let $A = [A_1, A_2 \ldots A_i .. A_v]$ be a matrix of v blocks of columns of A corresponding to the blocks $\mathbf{x}_i$.

$$\underbrace{\begin{bmatrix} A_{11} & A_{i1} & A_{v1} \\ A_{12} & A_{i2} & A_{v2} \\ A_{1p} & A_{ip} & A_{vp} \end{bmatrix}}_{p \text{ Linear constraints}} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_i \\ \mathbf{x}_v \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{v} A_{i1}\mathbf{x}_i \\ \sum_{i=1}^{v} A_{i2}\mathbf{x}_i \\ \sum_{i=1}^{v} A_{ip}\mathbf{x}_i \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_p \end{bmatrix}$$

Dual decomposition (contd.)

- Thus: $f(\mathbf{x}) = \sum_{i=1}^{v} f_i(\mathbf{x}_i)$ and $\sum_{i=1}^{v} A_i\mathbf{x}_i = \mathbf{b}$
- Using this, simplify the first iterative step of dual ascent as $\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}) + {\lambda^k}^\top (A\mathbf{x} - b)$