Lagrangian Relaxation

Karl Stratos

1 Max-Min Inequality

Let $f : A \times B \to \mathbb{R}$ be any function where $A \subseteq \mathbb{R}^m$, $B \subseteq \mathbb{R}^n$, and the smallest upper bound $\min_{a \in A} \max_{b \in B} f(a, b)$ exists. The max-min inequality says it is at least as large as the largest lower bound.

Lemma 1.1 (Max-min inequality)**.**

$$\max_{b \in B} \min_{a \in A} f(a, b) \leq \min_{a \in A} \max_{b \in B} f(a, b)$$

We say $(a^*, b^*) \in A \times B$ is a **saddle point** if $f(a^*, b^*) = \min_{a \in A} f(a, b^*) = \max_{b \in B} f(a^*, b)$. The following lemma says the existence of a saddle point is equivalent to the "strong" max-min property of f, that is the largest lower bound equals the smallest upper bound.

Lemma 1.2 (Strong max-min property)**.** $(a^*, b^*) \in A \times B$ is a saddle point iff

$$\min_{a \in A} f(a, b^*) = \max_{b \in B} \min_{a \in A} f(a, b) = \min_{a \in A} \max_{b \in B} f(a, b) = \max_{b \in B} f(a^*, b)$$

2 Lagrangian and Duality

We consider minimizing $f : \mathbb{R}^d \to \mathbb{R}$ subject to m inequality constraints and r equality constraints. Let $h : \mathbb{R}^d \to \mathbb{R}^m$ and $l : \mathbb{R}^d \to \mathbb{R}^r$ define a set of **primal feasible** points $\mathcal{P} := \left\{x \in \mathbb{R}^d : h(x) \leq 0_m, \, l(x) = 0_r\right\}$. The **primal problem** is

$$f^* := \min_{x \in \mathcal{P}} f(x) \tag{1}$$

We assume f^* exists. Let $\mathcal{D} := \{(u, v) \in \mathbb{R}^m \times \mathbb{R}^r : u \geq 0_m\}$ denote the set of **dual feasible** points. Each dual feasible point $\lambda \in \mathcal{D}$ corresponds to $m + r$ appropriate weights for penalizing constraint violation. The **Lagrangian** $L : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}$ is

$$L(x, (u, v)) := f(x) + \langle u, h(x) \rangle + \langle v, l(x) \rangle$$

Lemma 2.1 (Unconstrained primal via the Lagrangian)**.**

$$f^* = \min_{x \in \mathbb{R}^d} \max_{\lambda \in \mathcal{D}} L(x, \lambda)$$

We now switch the order of optimization. The **dual problem** is

$$g^* := \max_{\lambda \in \mathcal{D}} \min_{x \in \mathbb{R}^d} L(x, \lambda) \tag{2}$$

By defining the **Lagrangian dual function** $g(\lambda) := \min_{x \in \mathbb{R}^d} L(x, \lambda)$ we can write the dual problem more simply as $g^* = \max_{\lambda \in \mathcal{D}} g(\lambda)$. Note that g is concave since it is a point-wise minimum of affine functions. For simple functions (e.g., affine) we can derive an explicit form of the dual problem (Appendix B). The following result known as "weak duality" is an immediate application of the max-min inequality.

Corollary 2.2 (Weak duality)**.** $f^* \geq g^*$

Definition 2.1. We say $(x^*, \lambda^*) \in \mathbb{R}^d \times \mathcal{D}$ is a **strong optimal pair** if all of the following hold:

1. **Primal optimal**: $x^* \in \mathcal{P}$ and $f^* = f(x^*)$
2. **Dual optimal**: $g^* = g(\lambda^*)$
3. **Strong duality**: $f(x^*) = g(\lambda^*)$

For (x, λ) to be a strong optimal pair, they must satisfy a number of stringent requirements. First, $x \in \mathbb{R}^d$ has to be not only primal feasible (i.e., $h(x) \leq 0_m$ and $l(x) = 0_r$) but also primal optimal: $f^* = f(x)$. Second, $\lambda \in \mathcal{D}$ has to be dual optimal: $g^* = g(\lambda)$. Finally, the primal and dual optimum that they achieve must be the same: $f^* = f(x) = g(\lambda) = g^*$. Because of this last requirement, we cannot guarantee the existence of a strong optimal pair. Specifically, there exists a strong optimal pair iff $f^* = g^*$ in the given the problem.

Theorem 2.3 (Saddle point theorem)**.** If $(x^*, \lambda^*) \in \mathbb{R}^d \times \mathcal{D}$ is a saddle point of the Lagrangian, then it is a strong optimal pair.

Lemma 2.4 (Certificate of optimality)**.** Let $(x^*, \lambda^*) \in \mathcal{P} \times \mathcal{D}$. If $f(x^*) = g(\lambda^*)$, then (x^*, λ^*) is a saddle point of the Lagrangian.

3 KKT Conditions

Lemma 3.1 (Necessary optimality conditions)**.** If $(x^*, \lambda^*) \in \mathbb{R}^d \times \mathcal{D}$ is a strong optimal pair where $\lambda^* = (u^*, v^*)$,

1. **Stationarity**: 0_d is a subgradient of $L(x, \lambda^*)$ with respect to x at $x = x^*$.
2. **Complementary slackness**: $u_i^* h_i(x^*) = 0$ for each $i = 1 \dots m$.

Note that if the Lagrangian is differentiable in x, then the stationarity condition simply says $\nabla_x L(x, \lambda^*) = 0_d$. This motivates the following definition:

Definition 3.1. Let $x \in \mathbb{R}^d$, $u \in \mathbb{R}^m$, and $v \in \mathbb{R}^r$. Write $\lambda = (u, v)$. We say (x, λ) satisfies the **KKT conditions** if

1. **Primal feasibility**: $x \in \mathcal{P}$
2. **Dual feasibility**: $\lambda \in \mathcal{D}$
3. **Stationarity**: $0_d \in \left.\frac{\partial L(x', \lambda)}{\partial x'}\right|_{x'=x}$
4. **Complementary slackness**: $u_i h_i(x) = 0$ for each $i = 1 \dots m$

Lemma 3.1 says a strong optimal pair (x^*, λ^*) satisfies the KKT conditions. The converse, that (x, λ) satisfying the KKT conditions is a strong optimal pair, does not hold in general and requires an additional *convexity* condition.

Lemma 3.2 (Sufficient optimality conditions)**.** Suppose $(x^*, \lambda^*) \in \mathcal{P} \times \mathcal{D}$ satisfies the KKT conditions. If f, $h_1 \dots h_m$, and $l_1 \dots l_r$ are convex, then (x^*, λ^*) is a saddle point of the Lagrangian. In particular, it is a strong optimal pair (Theorem 2.3).

Theorem 3.3 (Slater's condition)**.** If

1. f and $h_1 \dots h_m$ are convex; $l_1 \dots l_r$ are affine, and
2. There exists a strictly primal feasible point, namely $x \in \mathbb{R}^d$ such that $h(x) < 0_m$ and $l(x) = 0_r$,

then strong duality holds: $f^* = g^*$.

The above theorem, given without proof, is a sufficient condition for strong duality in convex-affine programs known as **Slater's condition**. Since strong duality guarantees the existence of a strong optimal pair, we summarize our results in the following theorem (also using strong duality in linear programs, Theorem B.1).

Theorem 3.4 (Lagrangian relaxation)**.** Let $f, h_1 \dots h_m, l_1 \dots l_r$ be real-valued functions over $\mathbb{R}^d$ where $f, h_1 \dots h_m$ are convex and $l_1 \dots l_r$ are affine. Assume that either

1. $f, h_1 \dots h_m$ are also affine, or
2. There is at least one $x \in \mathbb{R}^d$ such that $h_i(x) < 0$ and $l_j(x) = 0$ for all i, j.

Then a strong optimal pair exists. Furthermore, $(x^*, \lambda^*) \in \mathbb{R}^d \times \mathcal{D}$ is a strong optimal pair iff it satisfies the KKT conditions.

3.1 Constraint qualifications

The conditions 1 and 2 in the theorem are examples of constraint qualifications. They are *sufficient* conditions for strong duality. In particular, strong duality may hold without a constraint qualification. If $f, h_1 \dots h_m, l_1 \dots l_r$ are convex, any point that satisfies the KKT conditions is a strong optimal pair by Lemma 3.2. Hence one strategy when $f, h_1 \dots h_m, l_1 \dots l_r$ are convex is to search for a point that satisfies the KKT conditions even if the conditions in the theorem are not strictly met: if we are able to find it, we have a strong optimal pair. This is especially convenient if there are no inequality constraints.

3.2 No Inequality Constraints

Suppose $m = 0$ so that the primal problem is $f^* = \min_{x \in \mathbb{R}^d : l(x) = 0_r} f(x)$ and the Lagrangian is $L(x, v) = f(x) + \langle v, l(x) \rangle$. If f, l are also differentiable, the KKT conditions on $(x, v) \in \mathbb{R}^d \times \mathbb{R}^r$ simplify to just $l(x) = 0_r$ (primal feasibility) and $\nabla_x L(x, v) = 0_d$ (stationarity). Note that we can express primal feasibility as $\nabla_v L(x, v) = 0_r$. Thus finding $(x^*, v^*) \in \mathbb{R}^d \times \mathbb{R}^r$ satisfying KKT conditions is equivalent to solving a system of $d + r$ linear equations: $\nabla_x L(x, v) = 0_d$ and $\nabla_v L(x, v) = 0_r$. If the equations are linearly independent the solution will be unique; otherwise (i.e., underdetermined) there will be infinitely many solutions. If $f, l_1 \dots l_r$ are additionally convex, the KKT conditions are sufficient for x^* to be the solution of the primal problem (Lemma 3.2). Since this allows for an explicit derivation of a solution of the primal problem, it is useful to convert inequality constraints to equality constraints if possible, for instance by arguing that any optimal solution must make the inequalities tight (Section 4.3).

4 Applications

4.1 Relaxing Inequality Constraints in Convex Programs

Lemma 4.1. Let $f, q_1 \dots q_m$ be convex functions over $\mathbb{R}^d$. Pick $\lambda^* \geq 0_m$. If $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x) + \langle \lambda^*, q(x) \rangle$, there exists some $t \in \mathbb{R}^m$ such that $x^* \in \arg\min_{x \in \mathbb{R}^d :\, q(x) \leq t} f(x)$.

Lemma 4.2. Let $f, q_1 \dots q_m$ be convex functions over $\mathbb{R}^d$. Pick $t \in \mathbb{R}^m$ and assume $\left\{x \in \mathbb{R}^d : q(x) < t\right\}$ is nonempty. If $x^* \in \arg\min_{x \in \mathbb{R}^d :\, q(x) \leq t} f(x)$, there exists some $\lambda^* \geq 0_m$ such that $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x) + \langle \lambda^*, q(x) \rangle$.

Since $||\cdot||_p : \mathbb{R}^d \to \mathbb{R}$ is convex for every $p \geq 1$ (including $p = \infty$ which gives the uniform norm $||x||_\infty := \lim_{p \to \infty} ||x||_p = \max_{i=1}^d |x_i|$), and the set $\{x \in \mathbb{R}^d : ||x||_p < t\}$ is nonempty for any $t > 0$, we get the following standard result on l_p regularization.

Corollary 4.3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be any convex function and $p \geq 1$. The following statements are equivalent:

- $x^* \in \mathbb{R}^d$ minimizes $f(x) + \lambda^* ||x||_p$ for some $\lambda^* \geq 0$.
- $x^* \in \mathbb{R}^d$ minimizes $f(x)$ subject to the constraint $||x||_p \leq t$ for some $t > 0$.

4.2 SVM Dual

Let $X \in \mathbb{R}^{N \times d}$ and $y \in \{\pm 1\}^N$ where the i-th row $x_i \in \mathbb{R}^d$ of X represents a d-dimensional input vector and y_i is the corresponding binary label. Given some $C > 0$, the primal SVM problem is

$$w^*, \xi^* = \underset{w \in \mathbb{R}^d,\, \xi \in \mathbb{R}^N :\, y \odot Xw \geq 1_N - \xi,\, \xi \geq 0_N}{\arg\min} \frac{1}{2} ||w||^2 + C \langle 1_N, \xi \rangle$$

where we omit the bias parameter without loss of generality by assuming that the first dimension of an input vector is always 1. Thanks to the slack variables, the primal problem is strictly feasible. Thus by Theorem 3.4 there exist (dual-feasible) $\lambda^*, \mu^* \geq 0_N$ that, together with (primal-feasible) w^*, ξ^*, form a strong optimal pair and satisfy the KKT conditions. The Lagrangian is

$$L(w, \xi, \lambda, \mu) = \frac{1}{2} ||w||^2 + C \langle 1_N, \xi \rangle + \langle \lambda, 1_N - \xi - y \odot Xw \rangle - \langle \mu, \xi \rangle$$

Since L is convex in $w \in \mathbb{R}^d$ and $\xi \in \mathbb{R}^N$ and

$$\begin{aligned} \nabla_w L(w, \xi, \lambda, \mu) = 0_d \quad &\Leftrightarrow \quad w = X^\top (\lambda \odot y) \\ \nabla_\xi L(w, \xi, \lambda, \mu) = 0_N \quad &\Leftrightarrow \quad \mu = C1_N - \lambda \end{aligned}$$

the Lagrangian dual function is

$$g(\lambda, \mu) := \min_{w \in \mathbb{R}^d,\, \xi \in \mathbb{R}^N} L(w, \xi, \lambda, \mu) = \langle \lambda, 1_N \rangle - \frac{1}{2} (\lambda \odot y)^\top X X^\top (\lambda \odot y)$$

While g is just a function of λ, the constraint $\mu \geq 0$ remains in the dual problem. Since this is equivalent to $\lambda \leq C1_N$ in the Lagrangian dual, the dual problem is

$$\lambda^* = \operatorname*{arg\,max}_{0_N \leq \lambda \leq C1_N} \langle \lambda, 1_N \rangle - \frac{1}{2} (\lambda \odot y)^\top X X^\top (\lambda \odot y)$$

By complementary slackness,

$$\begin{aligned} \lambda_i^* \left(1 - \xi_i^* - y_i \langle w^*, x_i \rangle\right) = 0 \quad &\Leftrightarrow \quad \lambda_i^* = 0 \quad \vee \quad y_i \langle w^*, x_i \rangle = 1 - \xi_i^* \\ &\Rightarrow \quad \lambda_i^* = 0 \quad \vee \quad y_i \langle w^*, x_i \rangle \leq 1 \end{aligned}$$

for all $i = 1 \dots N$. Thus $\lambda_i^* > 0$ implies $y_i \langle w^*, x_i \rangle \leq 1$ (the converse is not necessarily true): in this case x_i is called a support vector since $w^* = \sum_{i=1:\ \lambda_i^* > 0}^N \lambda_i^* y_i x_i$. Furthermore, suppose $\lambda_i^* < C$. Note again by complementary slackness,

$$\mu_i^* \xi_i^* = 0 \quad \Leftrightarrow \quad \mu_i^* = 0 \quad \vee \quad \xi_i^* = 0$$

for all $i = 1 \dots N$. Thus $\mu_i^* = C - \lambda_i^* > 0$ and we have $\xi_i^* = 0$. Combining these observations, we see that if $0 < \lambda_i^* < C$, then $y_i \langle w^*, x_i \rangle = 1$.

4.3 Local Update with Linear Approximation

Consider minimizing a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. We can optimize it iteratively by starting from some random point in $\mathbb{R}^d$ and at each non-stationary point x' (i.e., $\nabla f(x') \neq 0_d$) optimizing a linear approximation of f around that point: $f(x) \approx f(x') + \nabla f(x')^\top (x - x')$. Since this is only accurate for a local region around x', we enforce a locality constraint on x with a Q-norm $||u||_Q := \sqrt{u^\top Q u}$ for some $d \times d$ positive definite matrix $Q \succ 0$. This gives the primal problem: for some $C > 0$

$$x^* = \operatorname*{arg\,min}_{x \in \mathbb{R}^d:\ \frac{1}{2} ||x - x'||_Q^2 \leq C} \nabla f(x')^\top (x - x')$$

We simplify the problem by searching for the difference vector $\delta := x - x'$:

$$\delta^* = \operatorname*{arg\,min}_{\delta \in \mathbb{R}^d:\ \frac{1}{2} ||\delta||_Q^2 \leq C} \nabla f(x')^\top \delta$$

from which we recover $x^* = x' + \delta^*$. Suppose $\frac{1}{2} ||\delta^*||_Q^2 = C' < C$. Then $d^+ := \sqrt{C/C'} \delta^*$ satisfies $\frac{1}{2} ||\delta^+||_Q^2 = C$ and $\nabla f(x')^\top \delta^+ = \sqrt{C/C'} \nabla f(x')^\top \delta^* < \nabla f(x')^\top \delta^*$ using the fact that $\nabla f(x')^\top \delta^* < 0$ and $\sqrt{C/C'} > 1$. Thus

$$\delta^* = \operatorname*{arg\,min}_{\delta \in \mathbb{R}^d:\ \frac{1}{2} ||\delta||_Q^2 = C} \nabla f(x')^\top \delta$$

The Lagrangian is

$$L(\delta, \lambda) = \nabla f(x')^\top \delta + \lambda \left(\frac{1}{2} ||\delta||_Q^2 - C \right)$$

Since this is a linear objective with a convex equality constraint, it is sufficient to solve the following system of $d+1$ linear equations to find a strong optimal pair (δ^*, λ^*) if it exists (Section 3.2):

$$\begin{aligned} \nabla_\delta L(\delta, \lambda) = 0_d \quad &\Leftrightarrow \quad \delta = -\frac{1}{\lambda} Q^{-1} \nabla f(x') \\ \nabla_\lambda L(\delta, \lambda) = 0 \quad &\Leftrightarrow \quad \frac{1}{2} ||\delta||_Q^2 = C \end{aligned}$$

Taking the squared Q-norm of the $\delta = -\frac{1}{\lambda} Q^{-1} \nabla f(x')$ and setting it to $2C$ gives

$$\lambda^* = \sqrt{\frac{||\nabla f(x')||_{Q^{-1}}^2}{2C}} \qquad \delta^* = -\sqrt{\frac{2C}{||\nabla f(x')||_{Q^{-1}}^2}} Q^{-1} \nabla f(x')$$

Note that we have the certificate of optimality with the optimum value $-\sqrt{2C}\, ||\nabla f(x')||_{Q^{-1}}$ and thus strong duality holds (Lemma 2.4), even though none of the constraint qualifications in Theorem 3.4 is met (because the equality constraint is not affine). Hence the solution to the original problem is

$$x^* = x' - \eta Q^{-1} \nabla f(x')$$

where $\eta = \sqrt{2C / ||\nabla f(x')||_{Q^{-1}}^2}$. Gradient descent, Newton's method, and natural gradient are special cases of this solution with Q equal to the identity, Hessian, and Fisher information matrix.

A Proofs

Proof of Lemma 1.1. For any fixed $(a', b') \in A \times B$,

$$\min_{a \in A} f(a, b') \leq f(a', b')$$

Since this holds for all $b' \in B$, the following also holds:

$$\max_{b \in B} \min_{a \in A} f(a, b) \leq \max_{b \in B} f(a', b)$$

The LHS is now constant. Since this holds for all $a' \in A$, the following also holds:

$$\max_{b \in B} \min_{a \in A} f(a, b) \leq \min_{a \in A} \max_{b \in B} f(a, b)$$

■

Proof of Lemma 1.2. If (a^*, b^*) is a saddle point,

$$\max_{b \in B} \min_{a \in A} f(a, b) \geq \min_{a \in A} f(a, b^*) = \max_{b \in B} f(a^*, b) \geq \min_{a \in A} \max_{b \in B} f(a, b)$$

Since the other direction holds by the max-min inequality, we have an equality. Now suppose the equation holds for some $(a^*, b^*) \in A \times B$, repeated here for convenience:

$$\min_{a \in A} f(a, b^*) = \max_{b \in B} \min_{a \in A} f(a, b) = \min_{a \in A} \max_{b \in B} f(a, b) = \max_{b \in B} f(a^*, b)$$

Then we have

$$f(a^*, b^*) \leq \max_{b \in B} f(a^*, b) = \min_{a \in A} f(a, b^*) \leq f(a^*, b^*)$$

so the inequalities are equalities and (a^*, b^*) is a saddle point. ■

Proof of Lemma 2.1. For any $x \in \mathbb{R}^d$,

$$\begin{aligned} \max_{(u,v) \in \mathcal{D}} L(x, (u, v)) &= \max_{u \in \mathbb{R}^m :\, u \geq 0_m,\, v \in \mathbb{R}^r} f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j l_j(x) \\ &= \begin{cases} f(x) & \text{if } x \in \mathcal{P} \\ \infty & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

where we recall $\mathcal{P} := \left\{x \in \mathbb{R}^d :\ h(x) \leq 0_m,\ l(x) = 0_r\right\}$. Thus

$$\min_{x \in \mathbb{R}^d} \max_{(u,v) \in \mathcal{D}} L(x, (u, v)) = \min_{x \in \mathcal{P}} f(x) = f^*$$

■

Proof of Theorem 2.3. We must show that (1) $x^* \in \mathcal{P}$ and $f^* = f(x^*)$, (2) $g^* = g(\lambda^*)$, and (3) $f^* = g^*$. By the strong max-min property lemma (Lemma 1.2), we have

$$\min_{x \in \mathbb{R}^d} L(x, \lambda^*) = \max_{\lambda \in \mathcal{D}} \min_{x \in \mathbb{R}^d} L(x, \lambda) = \min_{x \in \mathbb{R}^d} \max_{\lambda \in \mathcal{D}} L(x, \lambda) = \max_{\lambda \in \mathcal{D}} L(x^*, \lambda)$$

The middle equality $g^* = f^*$ gives the third claim. The first equality can be written as $g(\lambda^*) = \max_{\lambda \in \mathcal{D}} g(\lambda)$ and gives the second claim. For the first claim, by the same observation in (3)

$$\max_{\lambda \in \mathcal{D}} L(x, \lambda) = \begin{cases} f(x) & \text{if } x \in \mathcal{P} \\ \infty & \text{otherwise} \end{cases}$$

Thus the last equality can be written as $\min_{x \in \mathcal{P}} f(x) = f(x^*)$. This gives the first claim. ■

Proof of Lemma 2.4.

$$f(x^*) \geq f^* \geq g^* \geq g(\lambda^*) = f(x^*)$$

where the last equality is by premise. Thus all the inequalities are equalities and we have

$$\max_{\lambda \in \mathcal{D}} L(x^*, \lambda) = \min_{x \in \mathbb{R}^d} \max_{\lambda \in \mathcal{D}} L(x, \lambda) = \max_{\lambda \in \mathcal{D}} \min_{x \in \mathbb{R}^d} L(x, \lambda) = \min_{x \in \mathbb{R}^d} L(x, \lambda^*)$$

where for the first term we use the fact $\max_{\lambda \in \mathcal{D}} L(x^*, \lambda) = f(x^*)$ since $x^* \in \mathcal{P}$ (see again (3)). By the strong max-min property lemma (Lemma 1.2), this means (x^*, λ^*) is a saddle point of $L : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}$. ■

Proof of Lemma 3.1. Recall the notation $\lambda^* = (u^*, v^*)$. Note that

$$\begin{aligned}
f(x^*) &= g(u^*, v^*) && \text{(strong duality)} \\
&= \min_{x \in \mathbb{R}^d} f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* l_j(x) \\
&\leq f(x^*) + \sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* l_j(x^*) \\
&\leq f(x^*) && \text{(feasibility of } x^* \text{ and } \lambda^*)
\end{aligned}$$

Therefore the inequalities are equalities. From the third equality (previously first inequality) we have that $x^* \in \arg\min_{x \in \mathbb{R}^d} L(x, \lambda^*)$. Since this is an unconstrained problem it must be a stationary point of $L(x, \lambda^*)$: that is the set of subgradients of $L(x, \lambda^*)$ with respect to x at $x = x^*$ must include 0_d. This gives the stationarity condition. From the final equality (previously second inequality) we have that

$$\sum_{i=1}^m u_i^* h_i(x^*) = -\sum_{j=1}^r v_j^* l_j(x^*) = 0$$

using the fact that $l_j(x^*) = 0$. Since $u_i^* \geq 0$ and $h_i(x^*) \leq 0$, $u_i^* h_i(x^*) \leq 0$ for all i. Thus it must be that $u_i^* h_i(x^*) = 0$ for all i, giving the complementary slackness condition. ■

Proof of Lemma 3.2. Writing $\lambda^* = (u^*, v^*)$,

$$\begin{aligned}
g(u^*, v^*) &= \min_{x \in \mathbb{R}^d} f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* l_j(x) \\
&= f(x^*) + \sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* l_j(x^*) && \text{(convexity of } f, h_i, l_j \text{; stationarity of } x^*) \\
&= f(x^*) && \text{(feasibility of } x^*, \lambda^* \text{; comp. slackness } u_i^* h_i(x^*) = 0)
\end{aligned}$$

Therefore we have the certificate of optimality (Lemma 2.4) and (x^*, λ^*) is a saddle point of the Lagrangian. ■

Proof of Theorem 3.4. Since Slater's condition is satisfied by premise, strong duality holds (Theorem 3.3) and thus a strong optimal pair exists. The KKT conditions are neccessary for $(x^*, \lambda^*) \in \mathbb{R}^d \times \mathcal{D}$ to be a strong optimal pair (Lemma 3.1). The KKT conditions are also sufficient in this case since f, $h_1 \dots h_m$, and $l_1 \dots l_r$ are convex (Lemma 3.2). ■

Proof of Lemma 4.1. Let $t = q(x^*)$ and consider the primal problem

$$\min_{x \in \mathbb{R}^d :\ q(x) \leq q(x^*)} f(x)$$

Note that x^* is primal feasible. The Lagrangian is $L(x, \lambda) = f(x) + \langle \lambda, q(x) - q(x^*) \rangle$. Since x^* minimizes $f(x) + \langle \lambda^*, q(x) \rangle$ over $\mathbb{R}^d$ we must have 0_d in

$$\frac{\partial}{\partial x} \left(f(x) + \langle \lambda^*, q(x) \rangle \right) \Big|_{x = x^*} = \frac{\partial}{\partial x} L(x, \lambda^*) \Big|_{x = x^*}$$

thus we have the stationarity condition. $\lambda^* \geq 0_m$ is dual feasible by premise. Let $h(x) = q(x) - q(x^*)$ and note that $h(x^*) = 0$, thus $\lambda_i^* h_i(x^*) = 0$ for all $i = 1 \dots m$ and we have the complementary slackness condition. Hence the feasible pair (x^*, λ^*) satisfies the KKT conditions and is a strong optimal pair (Lemma 3.2). In particular, x^* solves the primal problem. ■

Proof of Lemma 4.2. By premise Slater's condition is met and thus strong duality holds (Theorem 3.3). Strong duality says that there is some dual feasible point $\lambda^* \geq 0_m$ such that $f(x^*) = g(\lambda^*)$. By Lemma 2.4 (x^*, λ^*) is a saddle point of the Lagrangian $L(x, \lambda) = f(x) + \langle \lambda, q(x) - t \rangle$, in particular (Lemma 1.2)

$$\min_{x \in \mathbb{R}^d} \max_{\lambda \geq 0_m} L(x, \lambda) = \max_{\lambda \geq 0_m} \min_{x \in \mathbb{R}^d} L(x, \lambda) = \min_{x \in \mathbb{R}^d} L(x, \lambda^*)$$

Therefore

$$x^* \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d \colon q(x) \leq t} f(x) = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \max_{\lambda \geq 0_m} L(x, \lambda) = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} L(x, \lambda^*) = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} f(x) + \langle \lambda^*, q(x) \rangle$$

where the first equality follows from Lemma 2.1. ■

B Linear Programs

If the objective f and the constraints $h_1 \dots h_m, l_1 \dots l_r$ are affine, the constrained optimization problem is called a **linear program**. In this case, we can express the objective as $f(x) = c^\top x + \alpha$ for some $c \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$. As for a constraint, we have the following possibilities where $a \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$ are some constants:

1. $a^\top x + \beta = 0$: We keep it as is.
2. $a^\top x + \beta \leq 0$: We can turn it into $a^\top x + \beta + s = 0$ where $s \geq 0$ is a new nonnegative ("slack") variable.

Additionally, for every $k = 1 \dots d$, we can replace $x_k \in \mathbb{R}$ with $x_k^+ - x_k^-$ where $x_k^+, x_k^- \geq 0$. Finally, we can turn every equality constraint of form $a^\top x + \beta = 0$ into $a^\top x + \beta \geq 0$ and $-a^\top x - \beta \geq 0$. As a result, we can always write a linear program in the so-called canonical form for some $c \in \mathbb{R}^d$, $A \in \mathbb{R}^{m \times d}$, and $b \in \mathbb{R}^m$:

$$f^* := \min_{x \in \mathbb{R}^d \colon x \geq 0_d,\ Ax \geq b} c^\top x \qquad \text{(linear program in canonical form)}$$

The Lagrangian and the Lagrangian dual function are

$$L(x, (u, u')) = c^\top x + u^\top (b - Ax) - (u')^\top x \qquad \forall x \in \mathbb{R}^d,\ u \geq 0_m,\ u' \geq 0_d$$
$$g(u, u') = \max_{x \in \mathbb{R}^d} c^\top x + u^\top (b - Ax) - (u')^\top x \qquad \forall u \geq 0_m,\ u' \geq 0_d$$

Given any $u \geq 0_m$ and $u' \geq 0_d$, a maximizer $x \in \mathbb{R}^d$ of the Lagrangian must satisfy

$$\frac{\partial L(x, (u, u'))}{\partial x} = c - A^\top u - u' = 0_d$$

This implies the dual problem

$$g^* := \max_{u \geq 0_m,\ u' \geq 0_d} g(u, u') = \max_{u \geq 0_m,\ u' \geq 0_d \colon c = A^\top u + u'} b^\top u = \max_{u \geq 0_m \colon A^\top u \leq c} b^\top u$$

where the last equality follows by treating $u' \geq 0_d$ as a slack variable. This has a natural interpretation. There are d items with costs $c \in \mathbb{R}^d$ and m tasks with benefits $b \in \mathbb{R}^m$ (cost/benefit can be negative). We have $x \geq 0_d$ items and $u \geq 0_m$ utilities (in the m tasks). We assume the following linear relationship: $x \in \mathbb{R}^d$ items imply $Ax \in \mathbb{R}^m$ benefits; $u \in \mathbb{R}^m$ utilites imply $A^\top u \in \mathbb{R}^d$ costs. In the primal, we minimize the total cost $c^\top x$ while ensuring benefits $Ax \geq b$. In the dual, we maximize the total benefit $b^\top u$ while limiting costs $A^\top u \leq c$.

There is a large body of work on solving linear programs. An important result is that strong duality holds for linear programs. We omit the proof, but the result can be shown by contradiction (using Farkas' lemma) or by construction (using the simplex algorithm).

Theorem B.1 (Strong duality in linear programs)**.** Given any $c \in \mathbb{R}^d$, $A \in \mathbb{R}^{m \times d}$, and $b \in \mathbb{R}^m$, let

$$f^* = \min_{x \in \mathbb{R}^d \colon x \geq 0_d,\ Ax \geq b} c^\top x \qquad g^* = \max_{u \geq 0_m \colon A^\top u \leq c} b^\top u$$

If either solution exists, the other also exists and $f^* = g^*$.