# Constrained Optimization[1]

In general, a constrained optimization problem can be written as

$$\hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathcal{C}} \Psi(\boldsymbol{x}), \tag{1}$$

with constraint set $\mathcal{C} \subseteq \mathbb{R}^{n_p}$. Here, we focus on constraint sets consisting of equality and/or inequality constraints:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^{n_p}} \ & \Psi(\boldsymbol{x}), \\ \text{s.t. } & g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, n_i, \\ & h_i(\boldsymbol{x}) = 0, \quad i = 1, \ldots, n_e. \end{aligned} \tag{2}$$

"Greater-than-or-equal" inequalities can be included via negation. Many of the constraints common to image reconstruction problems can be expressed as equality or inequality constraints:

1. Nonnegativity constraints. Enforcing $\boldsymbol{x}$ in the nonnegative orthant is equivalent to introducing $n_p$ inequality constraints of the form $g_i(\boldsymbol{x}) = -x_i$, for $i = 1, \ldots, n_p$. Then, $n_i = n_p$.
2. Box constraints. Enforcing that $\boldsymbol{x}$ lies in the closed box $\prod_{i=1}^{n_p} [l_i, u_i]$ corresponds to $2n_p$ inequality constraints: $g_i(\boldsymbol{x}) = l_i - x_i$, and $g_{i+n_p}(\boldsymbol{x}) = x_i - u_i$, for $i = 1, \ldots, n_p$.
3. It is sometimes useful to enforce $\boldsymbol{x}$ belonging to the intersection of $n_e$ hyperplanes, $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$, where $\boldsymbol{C} \in \mathbb{R}^{n_e \times n_p}$, and $\boldsymbol{d} \in \mathbb{R}^{n_e}$. This constraint is equivalent to enforcing $h_i(\boldsymbol{x}) = [\boldsymbol{C}\boldsymbol{x}]_i - d_i$, for $i = 1, \ldots, n_e$.

These constraints have real practical applications. Nonnegativity constraints are useful when the image signal represents an actual quantity or count that should not fall below zero. In X-ray CT, or PET, we have observed our image signal corresponds to absorption of photons, so a negative value would have no physical meaning. Equality constraints also have many uses, including forcing pixels outside the object (based on prior information like a reference image) to zero, and applications like inpainting and superresolution, if the acquired data/pixels should remain unchanged.

## 1 Penalty Method for Constrained Optimization

The penalty method replaces equality and inequality constraints by adding a regularizer-like function penalizing choices of $\boldsymbol{x}$ that violate the constraints to the objective function. The constrained problem in Eq. (2) is equivalent to the unconstrained problem

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x}) + \sum_{i=1}^{n_i} I_+(g_i(\boldsymbol{x})) + \sum_{i=1}^{n_e} I_\pm(h_i(\boldsymbol{x})), \tag{3}$$

where $I_+(u)$ and $I_\pm(u)$ are indicator functions for the positive and nonzero reals, respectively:

$$I_+(u) = \begin{cases} 0, & u \leq 0; \\ \infty, & u > 0; \end{cases} \qquad I_\pm(u) = \begin{cases} 0, & u = 0; \\ \infty, & u \neq 0. \end{cases}$$

Usually, these functions are not used directly; instead "barrier" functions approximating these are used. The logarithmic barrier is one common choice for the inequality constraint:

$$I_+(u) \approx \begin{cases} -\log(-u), & u \leq 0; \\ \infty, & u > 0. \end{cases}$$

Such barrier functions are convex, so when $\Psi(\boldsymbol{x})$ is convex, and $\{g_i(\boldsymbol{x})\}$ and $\{h_i(\boldsymbol{x})\}$ are affine, the overall problem in Eq. (3) is also convex. The objective function is infinite when $\boldsymbol{x}$ violates any of the inequality or equality constraints and is equal to $\Psi(\boldsymbol{x})$ when $\boldsymbol{x}$ satisfies all the constraints. However, the sharpness of the indicator and logarithmic barrier functions complicates the optimization.

Instead, using a quadratic penalty yields

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} Q(\boldsymbol{x}; \mu) \triangleq \Psi(\boldsymbol{x}) + \frac{\mu}{2} \sum_{i=1}^{n_i} \| \max\{g_i(\boldsymbol{x}), 0\}\|_2^2 + \frac{\mu}{2} \sum_{i=1}^{n_e} \|h_i(\boldsymbol{x})\|_2^2, \tag{4}$$

with penalty parameter $\mu > 0$ chosen to control the closeness of the approximation to the indicator function. Other penalty functions, smooth and nonsmooth, are possible.

As $\mu \to \infty$, the quadratic penalty converges pointwise to the indicator function, and the problem in Eq. (4) approximates the constrained problem. Generally, penalty methods follow a "continuation schedule" for increasing $\mu$: from an initial value $\mu_0$, $\mu$ is increased by a scheduled factor each time the unconstrained problem approximately converges for that particular $\mu$. Convergence proofs exist for cases where the unconstrained problems are solved exactly or approximately. However, such convergence proofs guarantee only convergence to critical points of $\| \max\{g_i(\boldsymbol{x}), 0\}\|^2$ and $\|h_i(\boldsymbol{x})\|^2$, for all $i$, and not all critical points are feasible in general. When all constraints are linear or affine, and linearly independent, a frequent special case, all critical points are feasible, so the solution converges to an optimum of the original constrained problem as $\mu \to \infty$.

While implementing a penalty method is relatively straightforward, ensuring convergence requires repeatedly solving the objective function $Q(\boldsymbol{x}; \mu)$ for larger and larger $\mu$. Since the penalized problem generally becomes ill-conditioned as $\mu$ increases, approximate solvers become inefficient computationally for later iterations of the penalized method. In addition, the continuation schedule requires careful tuning to ensure $\mu$ increases at a rate that accelerates convergence without introducing instability. The augmented Lagrangian method enables convergence to the solution of the constrained problem with a finite choice of $\mu$, by combining the penalty method with Lagrange multipliers, alleviating these difficulties.

## 2 The Lagrange Dual Problem

Consider again the constrained optimization problem in Eq. 2. For convenience, we collect the inequality and equality constraints and write $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$, and $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$. The Lagrangian of this problem is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq \Psi(\boldsymbol{x}) + \boldsymbol{\lambda}'\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\nu}'\boldsymbol{h}(\boldsymbol{x}), \tag{5}$$

where $\boldsymbol{\lambda} \geq \boldsymbol{0}$ and $\boldsymbol{\nu}$ are called Lagrange multipliers or dual variables. The Lagrange dual function is defined as

$$\ell(\boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}). \tag{6}$$

For any feasible point $\boldsymbol{x}$, and $\boldsymbol{\lambda} \geq \boldsymbol{0}$, $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \Psi(\boldsymbol{x})$, so $\ell(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \Psi(\hat{\boldsymbol{x}})$ for all $\boldsymbol{\lambda} \geq \boldsymbol{0}$ and $\boldsymbol{\nu}$.

In addition, it is easy to see that $\ell(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is a concave function of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$, regardless of the convexity of the original problem: for any $\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1 \geq \mathbf{0}$, $\boldsymbol{\nu}_0, \boldsymbol{\nu}_1$, and $\theta \in (0,1)$:

$$\begin{aligned}
\ell(\theta\boldsymbol{\lambda}_0 + (1-\theta)\boldsymbol{\lambda}_1, \theta\boldsymbol{\nu}_0 + (1-\theta)\boldsymbol{\nu}_1) &= \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \theta\boldsymbol{\lambda}_0 + (1-\theta)\boldsymbol{\lambda}_1, \theta\boldsymbol{\nu}_0 + (1-\theta)\boldsymbol{\nu}_1) \\
&= \inf_{\boldsymbol{x}} \theta\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_0, \boldsymbol{\nu}_0) + (1-\theta)\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_1, \boldsymbol{\nu}_1) \\
&\geq \theta(\inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_0, \boldsymbol{\nu}_0)) + (1-\theta)(\inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_1, \boldsymbol{\nu}_1)) \\
&= \theta\ell(\boldsymbol{\lambda}_0, \boldsymbol{\nu}_0) + (1-\theta)\ell(\boldsymbol{\lambda}_1, \boldsymbol{\nu}_1).
\end{aligned}$$

Because the original (primal) problem is bounded below for all feasible $\boldsymbol{x}$ by the Lagrange dual function, regardless of $\boldsymbol{\lambda} \geq 0$, $\boldsymbol{\nu}$, we can maximize the Lagrange dual function over all $\boldsymbol{\lambda} \geq \mathbf{0}$ and $\boldsymbol{\nu}$, and this maximum also is a lower bound on the minimal primal objective function value. We call this maximization the Lagrange dual problem,

$$\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}}\} \in \arg\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \ell(\boldsymbol{\lambda}, \boldsymbol{\nu}). \tag{7}$$

In general, this lower bound $\mathcal{L}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}})$ is not necessarily tight, and we have "weak duality." For some problems, the maximum value of the Lagrange dual function equals the minimum feasible value of $\Psi(\boldsymbol{x})$, and we have "strong duality."

In general, necessary conditions for optimality of $\hat{\boldsymbol{x}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}}$ include:

1. $\boldsymbol{g}(\hat{\boldsymbol{x}}) \leq \mathbf{0}$.
2. $\hat{\boldsymbol{\lambda}} \geq \mathbf{0}$.
3. $\boldsymbol{h}(\hat{\boldsymbol{x}}) = \mathbf{0}$.

When strong duality is satisfied, we also have the "complementary slackness" condition: $\hat{\lambda}_i g_i(\hat{\boldsymbol{x}}) = 0$ for all $i = 1, \ldots, n_i$. When the Lagrangian for such a problem is differentiable, $\nabla_{\boldsymbol{x}} \mathcal{L}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\nu}}) = \mathbf{0}$. Altogether, these are called the Karush-Kuhn-Tucker (KKT) conditions.

### 2.1 Inpainting example

Consider the problem

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{R}\boldsymbol{x}\|_2^2, \quad \text{s.t.} \quad \boldsymbol{S}\boldsymbol{x} = \boldsymbol{y},$$

where the matrix $\boldsymbol{S}$ specifies a mask identifying known image pixels, $(\boldsymbol{R}'\boldsymbol{R})^+$ is the Moore-Penrose pseudoinverse (take SVD and invert nonzero singular values), and $\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'$ is invertible. The Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\nu}) = \tfrac{1}{2}\|\boldsymbol{R}\boldsymbol{x}\|_2^2 + \boldsymbol{\nu}'(\boldsymbol{S}\boldsymbol{x} - \boldsymbol{y}).$$

Differentiating the Lagrangian and solving for $\boldsymbol{x}$ yields

$$\begin{aligned}
\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\nu}) &= \boldsymbol{R}'\boldsymbol{R}\boldsymbol{x} + \boldsymbol{S}'\boldsymbol{\nu} = \mathbf{0}; \\
\boldsymbol{x} &= -(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'\boldsymbol{\nu}.
\end{aligned}$$

Plugging back into the Lagrangian yields the Lagrange dual function:

$$\begin{aligned}
\ell(\boldsymbol{\nu}) &= \tfrac{1}{2}\|\boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'\boldsymbol{\nu}\|_2^2 + \boldsymbol{\nu}'(-\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'\boldsymbol{\nu} - \boldsymbol{y}) \\
&= -\tfrac{1}{2}\boldsymbol{\nu}'\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'\boldsymbol{\nu} - \boldsymbol{\nu}'\boldsymbol{y}.
\end{aligned}$$

The Moore-Penrose pseudoinverse satisfies $(\boldsymbol{R}'\boldsymbol{R})^+(\boldsymbol{R}'\boldsymbol{R})(\boldsymbol{R}'\boldsymbol{R})^+ = (\boldsymbol{R}'\boldsymbol{R})^+$. Also note this function is concave in $\boldsymbol{\nu}$. Solving the Lagrange dual problem by differentiating $\ell(\boldsymbol{\nu})$ and solving for $\hat{\boldsymbol{\nu}}$ yields

$$\begin{aligned}\nabla_{\boldsymbol{\nu}}\ell(\boldsymbol{\nu}) &= -\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'\hat{\boldsymbol{\nu}} - \boldsymbol{y} = \boldsymbol{0}\\ \hat{\boldsymbol{\nu}} &= -(\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}')^{-1}\boldsymbol{y}.\end{aligned}$$

Plugging $\hat{\boldsymbol{\nu}}$ back into our expression for $\boldsymbol{x}$ results in

$$\hat{\boldsymbol{x}} = (\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'(\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}')^{-1}\boldsymbol{y}.$$

Our solution $\hat{\boldsymbol{x}}, \hat{\boldsymbol{\nu}}$ can be used to verify strong duality for this problem:

$$\begin{aligned}\Psi(\hat{\boldsymbol{x}}) &= \tfrac{1}{2}\|\boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'(\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}')^{-1}\boldsymbol{y}\|_2^2\\ &= \tfrac{1}{2}\boldsymbol{y}'(\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}')^{-1}\boldsymbol{y};\\ \ell(\hat{\boldsymbol{\nu}}) &= -\tfrac{1}{2}\boldsymbol{y}'(\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}')^{-1}\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}'(\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}')^{-1}\boldsymbol{y} + \boldsymbol{y}'(\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}')^{-1}\boldsymbol{y}\\ &= \tfrac{1}{2}\boldsymbol{y}'(\boldsymbol{S}(\boldsymbol{R}'\boldsymbol{R})^+\boldsymbol{S}')^{-1}\boldsymbol{y}\\ &= \Psi(\hat{\boldsymbol{x}}).\end{aligned}$$

## 3 Augmented Lagrangian Method

When strong duality is satisfied, direct solution of the Lagrange dual problem results in dual variables that are optimal for the primal problem. However, the dual problem is not always easy to solve.

Instead, we introduce the augmented Lagrangian approach that combines the Lagrangian with a quadratic penalty function. Let us start by considering only equality constraints:

$$\mathcal{L}_A(\boldsymbol{x}, \boldsymbol{\nu}; \mu) \triangleq \Psi(\boldsymbol{x}) + \boldsymbol{\nu}'(\boldsymbol{h}(\boldsymbol{x})) + \tfrac{\mu}{2}\|\boldsymbol{h}(\boldsymbol{x})\|_2^2, \tag{8}$$

where $\mu > 0$ is the penalty parameter previously introduced in Eq. (4). The augmented Lagrangian method (sometimes called method of multipliers) repeats two steps until convergence:

$$\boldsymbol{x}^{(k)} = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x}) + \boldsymbol{\nu}^{(k)'}(\boldsymbol{h}(\boldsymbol{x})) + \tfrac{\mu}{2}\|\boldsymbol{h}(\boldsymbol{x})\|_2^2; \tag{9}$$

$$\boldsymbol{\nu}^{(k+1)} = \boldsymbol{\nu}^{(k)} + \mu\boldsymbol{h}(\boldsymbol{x}^{(k)}). \tag{10}$$

Some intuition about this method can be gained by assuming differentiability and finding the gradient of the solution for the $k$th iteration:

$$\nabla_{\boldsymbol{x}}\mathcal{L}_A(\boldsymbol{x}^{(k)}, \boldsymbol{\nu}^{(k)}; \mu) = \nabla_{\boldsymbol{x}}\Psi(\boldsymbol{x}^{(k)}) + (\nabla_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x}^{(k)}))'\boldsymbol{\nu}^{(k)} + \mu(\nabla_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x}^{(k)}))'\boldsymbol{h}(\boldsymbol{x}^{(k)}) = \boldsymbol{0}.$$

Assuming strong duality, this equality compares to the KKT optimality condition

$$\nabla_{\boldsymbol{x}}\mathcal{L}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\nu}}) = \nabla_{\boldsymbol{x}}\Psi(\hat{\boldsymbol{x}}) + (\nabla_{\boldsymbol{x}}\boldsymbol{h}(\hat{\boldsymbol{x}}))'\hat{\boldsymbol{\nu}} = \boldsymbol{0},$$

when $\boldsymbol{x}^{(k)} \approx \hat{\boldsymbol{x}}$, and $\hat{\boldsymbol{\nu}} \approx \boldsymbol{\nu}^{(k)} + \mu\boldsymbol{h}(\boldsymbol{x}^{(k)})$. This relationship yields the augmented Lagrangian update rule in Eq. (10). This concept of "adding back the error" contrasts with the penalty method described earlier. For the penalty method, we need $\mu\boldsymbol{h}(\boldsymbol{x}^{(k)})$ to converge to $\hat{\boldsymbol{\nu}}$ to ensure optimality of the solution. To achieve this convergence, we must increase $\mu$ indefinitely, while ensuring $\boldsymbol{h}(\boldsymbol{x}^{(k)})$ gets closer to zero. In the augmented Lagrangian framework, convergence can be achieved by getting $\boldsymbol{\nu}^{(k)} \to \hat{\boldsymbol{\nu}}$, regardless of our choice of $\mu$. This fundamental difference allows us to establish convergence of the augmented Lagrangian method for a range of $\mu \geq \bar{\mu}$, where $\bar{\mu} < \infty$. Sufficient for the convergence of this method are linear independence of the equality constraints, and $\nabla^2_{\boldsymbol{x}\boldsymbol{x}'} \mathcal{L}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\nu}}) \succ \mathbf{0}$. Then, $\hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}} \mathcal{L}_A(\boldsymbol{x}, \hat{\boldsymbol{\nu}}; \mu)$. Furthermore, there exists a range of $\boldsymbol{\nu}^{(k)}$ sufficiently close to $\hat{\boldsymbol{\nu}}$, such that the solution of $\arg\min_{\boldsymbol{x}} \mathcal{L}_A(\boldsymbol{x}, \boldsymbol{\nu}^{(k)}; \mu)$ also is close to $\hat{\boldsymbol{x}}$, and $\boldsymbol{\nu}^{(k)} \to \hat{\boldsymbol{\nu}}$, $\boldsymbol{x}^{(k)} \to \hat{\boldsymbol{x}}$.

### 3.1 Scaled dual variable formulation

Completing the square yields an alternative formulation of the augmented Lagrangian function that may be convenient:

$$\mathcal{L}_A(\boldsymbol{x}, \boldsymbol{\nu}; \mu) = \Psi(\boldsymbol{x}) + \tfrac{\mu}{2}\|\boldsymbol{h}(\boldsymbol{x}) + \tfrac{1}{\mu}\boldsymbol{\nu}\|_2^2 - \tfrac{1}{2\mu}\|\boldsymbol{\nu}\|^2, \tag{11}$$

where the last term does not depend on $\boldsymbol{x}$. Defining the scaled dual vector $\boldsymbol{b} = \boldsymbol{\nu}/\mu$, we have

$$\boldsymbol{x}^{(k)} = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x}) + \tfrac{\mu}{2}\|\boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{b}^{(k)}\|_2^2; \tag{12}$$

$$\boldsymbol{b}^{(k+1)} = \boldsymbol{b}^{(k)} + \boldsymbol{h}(\boldsymbol{x}^{(k)}). \tag{13}$$

### 3.2 Augmented Lagrangian for inequality constraints

To extend the equality-constrained formulation of the augmented Lagrangian method to inequality constraints, write the augmented Lagrangian including the inequality constraint penalty function $\|\max\{\mathbf{0}, \boldsymbol{g}(\boldsymbol{x})\}\|_2^2$:

$$\mathcal{L}_A(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}; \mu) = \Psi(\boldsymbol{x}) + \boldsymbol{\lambda}'\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\nu}'\boldsymbol{h}(\boldsymbol{x}) + \tfrac{\mu}{2}\|\max\{\mathbf{0}, \boldsymbol{g}(\boldsymbol{x})\}\|_2^2 + \tfrac{\mu}{2}\|\boldsymbol{h}(\boldsymbol{x})\|_2^2. \tag{14}$$

The dual variable update rule derived for the equality constraint becomes $\boldsymbol{\lambda}^{(k+1)} = \max\{\mathbf{0}, \boldsymbol{\lambda}^{(k)} + \mu\boldsymbol{g}(\boldsymbol{x}^{(k)})\}$, for the inequality constraint, enforcing the nonnegativity constraint on $\boldsymbol{\lambda}$ in the Lagrange dual problem.

## 4 Active Set Method

The active set method is another straightforward approach to solving constrained optimization problems with inequality constraints. To describe the method, let us first revisit the problem in (2). For a given feasible $\boldsymbol{x}$, inequality constraints $g_i(\boldsymbol{x}) \leq 0$ for $i = 1, \ldots, n_i$. An inequality constraint is called "active" if $g_i(\boldsymbol{x}) = 0$, "inactive" if $g_i(\boldsymbol{x}) < 0$. Equality constraints are always "active." Such constraints are considered active because they limit the minimum achievable value of the primal objective function $\Phi(\boldsymbol{x})$. If the constraint were eliminated, the objective function $\Phi(\boldsymbol{x})$ could be decreased further. Therefore, an equivalent optimization given an active set (set of active constraints) would be to minimize $\Phi(\boldsymbol{x})$, constraining the active set of constraints with equality, and ignoring the inactive constraints.

The active set method is:

1. Start from feasible $\boldsymbol{x}^{(0)}$.

2. Evaluate $g_i(\boldsymbol{x})$, $h_i(\boldsymbol{x})$ for all $i$. Divide constraints into active and inactive sets.

3. Minimize $\Phi(\boldsymbol{x})$ constraining the active set with equality (including active inequalities), and ignoring the inactive set.

4. From the solution, compute the Lagrange multipliers $\lambda_i$ for the active inequality constraints. Remove the active inequality constraints with $\lambda_i < 0$.

5. Evaluate inactive set and add back those inequality constraints that are violated.

6. Repeat this process until convergence.

The active set method assumes a procedure available to at least improve the solution of the equality constrained problem with each repetition. To help simplify the solution process, sequential linear programming (SLP) and sequential quadratic programming (SQP) use first-order and second-order approximations, respectively, of the constraint functions $g_i(\boldsymbol{x})$ and $h_i(\boldsymbol{x})$, in each iteration.

### 4.1 Example: $\ell_1 - \ell_2$ Minimization

To demonstrate, let's solve the 2D problem of finding the closest point $\boldsymbol{x}$ to a point $(1, -1)$ that has 1-norm less than or equal to one.

$$\begin{aligned} \arg\min_{x_1,x_2} \ & (x_1-1)^2 + (x_2+1)^2, \\ \text{s.t. } & x_1 + x_2 \le 1, \\ & -x_1 + x_2 \le 1, \\ & -x_1 - x_2 \le 1, \\ & x_1 - x_2 \le 1. \end{aligned} \tag{15}$$

To start, we need to identify a feasible point. Let's start with $\boldsymbol{x} = (1; 0)$. This point satisfies the first and last constraints with equality, so the active set are $\{x_1 + x_2 = 1, x_1 - x_2 = 1\}$. Since the feasible set with these two equality constraints consists of just $\boldsymbol{x} = (1; 0)$, we have our solution. What remains is to find the Lagrange multipliers:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = (x_1-1)^2 + (x_2+1)^2 + \lambda_1(x_1 + x_2 - 1) + \lambda_4(x_1 - x_2 - 1)$$

has solution $\boldsymbol{x} = (1; -1) - \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\boldsymbol{\lambda}$. And $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\boldsymbol{\lambda} = (0; -2)$ yields $\lambda_1 = -1$ and $\lambda_2 = 1$. This means that the first constraint should be removed from the active set, and the second retained.

The active set is now just $\{x_1 - x_2 = 1\}$; the Lagrangian is now

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = (x_1-1)^2 + (x_2+1)^2 + \lambda_4(x_1 - x_2 - 1).$$

The gradient is $2(\boldsymbol{x} - (1; -1)) + (1; -1)\lambda_4 = \mathbf{0}$, which yields $\boldsymbol{x} = (1 - \lambda_4/2)(1; -1)$. Substituting into the equality constraint yields $\lambda_4 = 1$, and $\boldsymbol{x} = (1/2; -1/2)$.

At this stage, we note that the active set remain active, and no inactive inequalities are violated, so the active set remains the same. Therefore, the method has converged. The solution $x_1 = 1/2$, $x_2 = -1/2$ is consistent with the geometric interpretation of the problem.