Introduction to convex optimization II

Sergio García[1]

The University of Edinburgh, UK

June 2018

[1]Based on lecture notes by Dr Paresh Date, Brunel University

A re-cap and an outline

- We have been looking at a class of convex optimization problems (convex objective, convex inequality constraints, affine equality constraints).
- We looked at some special types of practically relevant problems (LP, QP).

A re-cap and an outline

- We have been looking at a class of convex optimization problems (convex objective, convex inequality constraints, affine equality constraints).
- We looked at some special types of practically relevant problems (LP, QP).
- Now, we will first look at one more special type of problem (semidefinite program, or SDP).

A re-cap and an outline

- We have been looking at a class of convex optimization problems (convex objective, convex inequality constraints, affine equality constraints).
- We looked at some special types of practically relevant problems (LP, QP).
- Now, we will first look at one more special type of problem (semidefinite program, or SDP).
- Then we will move on to look at conditions for a point to achieve the optimum, followed by methods to solve *generic* convex problems (both unconstrained problems and constrained problems).

Semidefinite programming problems

- SDP has the form

$$\text{minimize} \quad \mathbf{c}^\top \mathbf{x} \tag{1}$$

$$\text{subject to} \quad x_1 F_1 + x_2 F_2 + \cdots + x_n F_n + G \leq 0, \tag{2}$$

  where F_i, G are symmetric matrices, x_i are decision variables.

- The inequality here is called an affine matrix inequality (AMI, also referred to as an LMI) and indicates that the eigenvalues of the matrix on the left hand side are non-positive.

Semidefinite programming problems

- SDP has the form

$$\text{minimize} \quad \mathbf{c}^\top \mathbf{x} \tag{1}$$

$$\text{subject to} \quad x_1 F_1 + x_2 F_2 + \cdots + x_n F_n + G \leq 0, \tag{2}$$

  where F_i, G are symmetric matrices, x_i are decision variables.

- The inequality here is called an affine matrix inequality (AMI, also referred to as an LMI) and indicates that the eigenvalues of the matrix on the left hand side are non-positive.
- Recall: the maximum eigenvalue of a symmetric matrix M is a convex function of M; hence (2) is convex in $\mathbf{x}$.
- If the matrices F_i and G are diagonal, this problem reduces to an LP.
- SDP can also be expressed in terms of symmetric, matrix valued decision variables.

SDP and Schur complement

- The following standard result from linear algebra is useful for converting certain non-standard convex problems into SDP:

$$M := \begin{bmatrix} M_1 & M_3 \\ M_3^\top & M_2 \end{bmatrix} \geq 0,\ M_2 > 0 \Leftrightarrow M_1 - M_3 M_2^{-1} M_3^\top \geq 0.$$

- $M_1(\mathbf{x})$, $M_2(\mathbf{x})$ and $M_3(\mathbf{x})$ are affine (symmetric matrix-valued) functions of $\mathbf{x}$. Note that $M_3 M_2^{-1} M_3^\top$ can be a very complicated expression in $\mathbf{x}$.

SDP and Schur complement

- The following standard result from linear algebra is useful for converting certain non-standard convex problems into SDP:

$$M := \begin{bmatrix} M_1 & M_3 \\ M_3^\top & M_2 \end{bmatrix} \geq 0,\ M_2 > 0 \Leftrightarrow M_1 - M_3 M_2^{-1} M_3^\top \geq 0.$$

- $M_1(\mathbf{x})$, $M_2(\mathbf{x})$ and $M_3(\mathbf{x})$ are affine (symmetric matrix-valued) functions of $\mathbf{x}$. Note that $M_3 M_2^{-1} M_3^\top$ can be a very complicated expression in $\mathbf{x}$.
- The block matrix $M_1 - M_3 M_2^{-1} M_3^\top$ is called the *Schur complement* of M_2 in M.
- This result can also be used to show that QP is a special case of SDP.

QP as a special case of SDP

- Using Schur complement,

$$\begin{bmatrix} t - (\mathbf{q}^\top \mathbf{x} + r) & \mathbf{x}^\top \\ \mathbf{x} & 2P^{-1} \end{bmatrix} \geq 0,\ P > 0,$$

  is equivalent to

$$\frac{1}{2}\mathbf{x}^\top P\mathbf{x} + \mathbf{q}^\top \mathbf{x} + r \leq t.$$

- Now minimising the auxiliary variable t subject to the above constraint and any affine constraints on $\mathbf{x}$ is a QP.
- Specialised software for solving SDPs (e.g. SeDuMi in matlab) can solve QP or LP, although this will rarely be adviseable!

SDP appears at unexpected places $\cdots$

- In complex analysis: consider a set of points $\{x_i\}$ (respectively, $\{y_i\}$), $i = 1, 2, \cdots, n$ in open (resp., closed) unit disk centred at origin in the complex plane.

SDP appears at unexpected places $\cdots$

- In complex analysis: consider a set of points $\{x_i\}$ (respectively, $\{y_i\}$), $i = 1, 2, \cdots, n$ in open (resp., closed) unit disk centred at origin in the complex plane.
- A rational function $f : \mathcal{C} \mapsto \mathcal{C}$ which is analytic in unit disk interpolates these points ($f(x_i) = y_i$) iff

$$P = \left[\frac{1-y_i y_j^*}{1-x_i x_j^*}\right] \geq 0.$$

SDP appears at unexpected places $\cdots$

- In complex analysis: consider a set of points $\{x_i\}$ (respectively, $\{y_i\}$), $i = 1, 2, \cdots, n$ in open (resp., closed) unit disk centred at origin in the complex plane.
- A rational function $f : \mathcal{C} \mapsto \mathcal{C}$ which is analytic in unit disk interpolates these points ($f(x_i) = y_i$) iff

$$P = \left[\frac{1-y_i y_j^*}{1-x_i x_j^*}\right] \geq 0.$$

- P can be expressed as an affine function of y_i, via Schur complement.
- This result is useful in control theory, where analyticity of f relates to stability of the underlying linear system.

SDP appears at unexpected places $\cdots$

- In probability theory: consider a set of real numbers m_i, $i = 1, 2, \cdots, n$. There exists a random variable z with m_i as moments iff

$$H_n := \left[m_{i+j-2} \right] \geq 0, \ i, j \in [0, n], \text{ with } m_0 = 1.$$

SDP appears at unexpected places $\cdots$

- In probability theory: consider a set of real numbers m_i, $i = 1, 2, \cdots, n$. There exists a random variable z with m_i as moments iff

$$H_n := \left[m_{i+j-2} \right] \geq 0, \ i, j \in [0, n], \text{ with } m_0 = 1.$$

- For example:

$$H_3 = \begin{bmatrix} 1 & m_1 & m_2 \\ m_1 & m_2 & m_3 \\ m_2 & m_3 & m_4 \end{bmatrix}.$$

- This is useful in statistical experiment design.

Optimality conditions

- For a convex $f(\mathbf{x})$, we know that $\nabla f(\mathbf{x}^\star) = 0$ defines the point $\mathbf{x}^\star$ which achieves the minimum.

Optimality conditions

- For a convex $f(\mathbf{x})$, we know that $\nabla f(\mathbf{x}^\star) = 0$ defines the point $\mathbf{x}^\star$ which achieves the minimum.
- To seek the conditions for optimality in a constrained problem, the idea of duality introduced.

Duality I

- Recall, we are solving

$$\begin{aligned} &\text{minimize } f_0(\mathbf{x}) \\ &\text{subject to } f_i(\mathbf{x}) \leq 0, \ i = 1, 2, \ldots, m \\ &\text{and } h_i(\mathbf{x}) = 0, \ i = 1, 2, \ldots, p. \end{aligned} \tag{3}$$

- Associated with (3), there is a *dual* convex maximization problem (concave objective, convex constraints).

Duality I

- Recall, we are solving

$$\begin{aligned} &\text{minimize } f_0(\mathbf{x}) \\ &\text{subject to } f_i(\mathbf{x}) \leq 0,\ i = 1, 2, \ldots, m \\ &\text{and } h_i(\mathbf{x}) = 0,\ i = 1, 2, \ldots, p. \end{aligned} \tag{3}$$

- Associated with (3), there is a *dual* convex maximization problem (concave objective, convex constraints).
- We define *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \mapsto \mathbb{R}$ as

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}), \tag{4}$$

where $\text{dom}\,(L) = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$.

Duality II

- We refer to λ_i (respectively, ν_i) as the *Lagrange multiplier* associated with the i^{th} inequality constraint $f_i(\mathbf{x}) \leq 0$ (respectively, i^{th} equality constraint $h_i(\mathbf{x}) = 0$).
- The *Lagrange dual function* (or simply *dual function*) $g : \mathbb{R}^m \times \mathbb{R}^p \mapsto \mathbb{R}$ is defined as the minimum value of Lagrangian over $\mathbf{x}$:

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) \right). \end{aligned} \tag{5}$$

If the Lagrangian is unbounded from below in $\mathbf{x}$, we set $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = -\infty$.

Duality II

- We refer to λ_i (respectively, ν_i) as the *Lagrange multiplier* associated with the i^{th} inequality constraint $f_i(\mathbf{x}) \leq 0$ (respectively, i^{th} equality constraint $h_i(\mathbf{x}) = 0$).
- The *Lagrange dual function* (or simply *dual function*) $g : \mathbb{R}^m \times \mathbb{R}^p \mapsto \mathbb{R}$ is defined as the minimum value of Lagrangian over $\mathbf{x}$:

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) \right). \end{aligned} \tag{5}$$

  If the Lagrangian is unbounded from below in $\mathbf{x}$, we set $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = -\infty$.
- The infimum in defining g is taken over $\mathcal{D}$, not over $\mathcal{F}$. If $\mathbf{x} \in \mathcal{F}$ and $\boldsymbol{\lambda} \geq 0$, note that $f_0(\mathbf{x}) \geq L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$.

Duality III

- Whenever $\boldsymbol{\lambda} \geq 0$, $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$. A natural question to ask is whether the largest possible value of $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is a good approximation to our optimal value $p^\star$.

Duality III

- Whenever $\boldsymbol{\lambda} \geq 0$, $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$. A natural question to ask is whether the largest possible value of $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is a good approximation to our optimal value $p^\star$.
- Assume that $f_i(\cdot)$ are convex and $h_i(\cdot)$ are affine functions of $\mathbf{x}$, $f_0(\cdot)$ is convex and assume that there exists $\mathbf{x} \in \mathcal{D}$ such that $f_i(\mathbf{x}) < 0$, i.e. the problem is strictly feasible. Then we have

$$\max_{\boldsymbol{\lambda} \geq 0, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) := d^\star = p^\star.$$

Duality IV

- g is a pointwise infimum of affine functions of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ and is hence concave. Hence maximizing $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ subject to $\boldsymbol{\lambda} \geq 0$ is a convex problem and is usually referred to as the *Lagrange dual problem* to (3). The multipliers $(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ which achieve the optimal value of g are called the optimal Lagrange multipliers.

Duality IV

- g is a pointwise infimum of affine functions of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ and is hence concave. Hence maximizing $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ subject to $\boldsymbol{\lambda} \geq 0$ is a convex problem and is usually referred to as the *Lagrange dual problem* to (3). The multipliers $(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ which achieve the optimal value of g are called the optimal Lagrange multipliers.
- The difference $d^\star - p^\star$ is called the *duality gap*. It *can* be zero in conditions more general than those given above; the above conditions are sufficient but not necessary.

Duality V: LP example

- Primal LP problem is

$$\begin{aligned} &\text{minimize } \mathbf{c}^\top \mathbf{x} \\ &\text{subject to } A\mathbf{x} = \mathbf{b}, \\ &\mathbf{x} \geq 0. \end{aligned}$$

Duality V: LP example

- Primal LP problem is

$$\begin{aligned} &\text{minimize } \mathbf{c}^\top \mathbf{x} \\ &\text{subject to } A\mathbf{x} = \mathbf{b}, \\ &\mathbf{x} \geq 0. \end{aligned}$$

- Lagrange dual is

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= -\mathbf{b}^\top \boldsymbol{\nu}, \text{ if } c + A^\top \boldsymbol{\nu} = \boldsymbol{\lambda}, \\ &= -\infty \text{ otherwise.} \end{aligned}$$

Duality V: LP example

- Primal LP problem is

$$\begin{aligned} &\text{minimize } \mathbf{c}^\top \mathbf{x} \\ &\text{subject to } A\mathbf{x} = \mathbf{b}, \\ &\mathbf{x} \geq 0. \end{aligned}$$

- Lagrange dual is

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= -\mathbf{b}^\top \boldsymbol{\nu}, \text{ if } c + A^\top \boldsymbol{\nu} = \boldsymbol{\lambda}, \\ &= -\infty \text{ otherwise.} \end{aligned}$$

- Dual problem is

$$\begin{aligned} &\text{maximize } -\mathbf{b}^\top \boldsymbol{\nu} \\ &\text{subject to } A^\top \boldsymbol{\nu} + \mathbf{c} \geq 0. \end{aligned}$$

KKT conditions for optimality I

- For differentiable convex f_0, f_i and affine h_i, let $\tilde{\mathbf{x}}$, $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\nu}}$ be the points which satisfy the following (Karush-Kuhn-Tucker or KKT) conditions:

$$f_i(\tilde{\mathbf{x}}) \leq 0,\ i = 1, 2, \ldots, m, \tag{6}$$

$$h_i(\tilde{\mathbf{x}}) = 0,\ i = 1, 2, \ldots, p, \tag{7}$$

$$\tilde{\lambda}_i \geq 0,\ i = 1, 2, \ldots, m, \tag{8}$$

$$\tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = 0,\ i = 1, 2, \ldots, m, \tag{9}$$

$$\nabla f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \tilde{\nu}_i \nabla h_i(\tilde{\mathbf{x}}) = 0, \tag{10}$$

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

KKT conditions for optimality I

- For differentiable convex f_0, f_i and affine h_i, let $\tilde{\mathbf{x}}$, $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\nu}}$ be the points which satisfy the following (Karush-Kuhn-Tucker or KKT) conditions:

$$f_i(\tilde{\mathbf{x}}) \leq 0,\ i = 1, 2, \ldots, m, \tag{6}$$

$$h_i(\tilde{\mathbf{x}}) = 0,\ i = 1, 2, \ldots, p, \tag{7}$$

$$\tilde{\lambda}_i \geq 0,\ i = 1, 2, \ldots, m, \tag{8}$$

$$\tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = 0,\ i = 1, 2, \ldots, m, \tag{9}$$

$$\nabla f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \tilde{\nu}_i \nabla h_i(\tilde{\mathbf{x}}) = 0, \tag{10}$$

KKT conditions for optimality I

- For differentiable convex f_0, f_i and affine h_i, let $\tilde{\mathbf{x}}$, $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\nu}}$ be the points which satisfy the following (Karush-Kuhn-Tucker or KKT) conditions:

$$f_i(\tilde{\mathbf{x}}) \leq 0,\ i = 1, 2, \ldots, m, \tag{6}$$

$$h_i(\tilde{\mathbf{x}}) = 0,\ i = 1, 2, \ldots, p, \tag{7}$$

$$\tilde{\lambda}_i \geq 0,\ i = 1, 2, \ldots, m, \tag{8}$$

$$\tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = 0,\ i = 1, 2, \ldots, m, \tag{9}$$

$$\nabla f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \tilde{\nu}_i \nabla h_i(\tilde{\mathbf{x}}) = 0, \tag{10}$$

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

KKT conditions for optimality II

- *Solving* the optimization problem is equivalent to solving the KKT system.
- Note:(6)-(7) indicate that the point $\tilde{\mathbf{x}}$ is feasible for the primal problem, while the *complementary slackness condition* (9) guarantees that $f_0(\tilde{\mathbf{x}}) = L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$.
- The last equation (10) generalises the first order convexity condition for the constrained case.

KKT conditions for optimality II

- *Solving* the optimization problem is equivalent to solving the KKT system.
- Note:(6)-(7) indicate that the point $\tilde{\mathbf{x}}$ is feasible for the primal problem, while the *complementary slackness condition* (9) guarantees that $f_0(\tilde{\mathbf{x}}) = L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$.
- The last equation (10) generalises the first order convexity condition for the constrained case.
- We will return to Lagrangian and KKT later in this lecture.

Unconstrained minimization I

- We now look at solving unconstrained problems of the form

$$\text{minimize} \quad f(\mathbf{x}), \tag{11}$$

  with a convex and differentiable f.

- $\text{dom}(f)$ is either $\mathbb{R}^n$ or is an open $\mathcal{D}$ such that

$$\lim_{\mathbf{x} \to \text{bd}(\mathcal{D})} f(\mathbf{x}) = \infty.$$

Unconstrained minimization I

- We now look at solving unconstrained problems of the form

$$\text{minimize} \quad f(\mathbf{x}), \tag{11}$$

  with a convex and differentiable f.

- $\text{dom}(f)$ is either $\mathbb{R}^n$ or is an open $\mathcal{D}$ such that

$$\lim_{\mathbf{x} \to \text{bd}(\mathcal{D})} f(\mathbf{x}) = \infty.$$

- A first order necessary and sufficient condition for a point $\mathbf{x}^\star$ to be optimal is

$$\nabla f(\mathbf{x}^\star) = 0. \tag{12}$$

- Sometimes we have a closed-form solution (e.g. unconstrained linear least squares problem). If not, we use iterative methods to solve the problem.

Unconstrained minimization II

- **Given** an initial point $\mathbf{x}^{(0)}$, set $k = 0$.
  Do
  - Choose a search direction vector $\mathbf{d}^{(k)}$, a stepsize $\alpha_k > 0$, and set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$;
  - Set $k = k + 1$;

  until a stopping criterion is satisfied.

Unconstrained minimization II

- **Given** an initial point $\mathbf{x}^{(0)}$, set $k = 0$.
  Do
  - Choose a search direction vector $\mathbf{d}^{(k)}$, a stepsize $\alpha_k > 0$, and set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$;
  - Set $k = k + 1$;

  until a stopping criterion is satisfied.
- We need $\mathbf{d}^{(k)}$ to be such that $f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$, with α_k chosen to achieve sufficient decrease along each line.
- These are called *descent* methods.

Choice of $\mathbf{d}^{(k)}$

- Since f is convex, $\nabla f(\mathbf{x}^{(k)})^\top(\mathbf{y} - \mathbf{x}^{(k)}) \geq 0 \Rightarrow f(\mathbf{y}) \geq f(\mathbf{x}^{(k)})$. Hence we need $\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)} < 0$.
- Different descent methods differ in the choice of $\mathbf{d}^{(k)}$.

Choice of $\mathbf{d}^{(k)}$

- Since f is convex, $\nabla f(\mathbf{x}^{(k)})^\top(\mathbf{y} - \mathbf{x}^{(k)}) \geq 0 \Rightarrow f(\mathbf{y}) \geq f(\mathbf{x}^{(k)})$. Hence we need $\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)} < 0$.
- Different descent methods differ in the choice of $\mathbf{d}^{(k)}$.
- *Gradient descent method*: $\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})$.
- Stepsize chosen by line search to minimize $f(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)})$ over α_k - or by backtracking line search.

Choice of $\mathbf{d}^{(k)}$

- Since f is convex, $\nabla f(\mathbf{x}^{(k)})^\top(\mathbf{y} - \mathbf{x}^{(k)}) \geq 0 \Rightarrow f(\mathbf{y}) \geq f(\mathbf{x}^{(k)})$. Hence we need $\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)} < 0$.
- Different descent methods differ in the choice of $\mathbf{d}^{(k)}$.
- *Gradient descent method*: $\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})$.
- Stepsize chosen by line search to minimize $f(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)})$ over α_k - or by backtracking line search.
- Stopping criterion: $\|\nabla f(\mathbf{x})\|_2 \leq \eta$.

Backtracking line-search I

- **Given**, a descent direction $\mathbf{d}^{(k)}$, $\beta \in (0, 0.5)$, $\gamma \in (0, 1)$, $\hat{\alpha}_{k,j} = 1$, $j = 1$.
  Do

  1. set $\hat{\alpha}_{k,j+1} = \gamma \hat{\alpha}_{k,j}$,
  2. j=j+1,

  until $f(\mathbf{x}^{(k)} + \hat{\alpha}_{k,j}\mathbf{d}^{(k)}) < f(\mathbf{x}^{(k)}) + \beta \hat{\alpha}_{k,j} \nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)}$.

- Idea: start with a unit step size; reduce it by factor γ until the stopping criterion is satisfied.

Backtracking line-search I

- **Given**, a descent direction $\mathbf{d}^{(k)}$, $\beta \in (0, 0.5)$, $\gamma \in (0, 1)$, $\hat{\alpha}_{k,j} = 1$, $j = 1$.
  Do

  1. set $\hat{\alpha}_{k,j+1} = \gamma \hat{\alpha}_{k,j}$,
  2. j=j+1,

  until $f(\mathbf{x}^{(k)} + \hat{\alpha}_{k,j}\mathbf{d}^{(k)}) < f(\mathbf{x}^{(k)}) + \beta \hat{\alpha}_{k,j} \nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)}$.

- Idea: start with a unit step size; reduce it by factor γ until the stopping criterion is satisfied.

- This corresponds to choosing $\hat{\alpha}_{k,j}$ such that $f(\mathbf{x}^{(k)} + \hat{\alpha}_{k,j}\mathbf{d}^{(k)})$ decreases *sufficiently*, starting from $f(\mathbf{x}^{(k)})$, although we may not actually minimize it.

Backtracking line search II

- Since

$$f(\mathbf{x}^{(k)} + \hat{\alpha}_{k,j}\mathbf{d}^{(k)}) \geq f(\mathbf{x}^{(k)}) + \hat{\alpha}_{k,j}\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)}$$

  as f is convex, β must be less than 1.

- For a sufficiently small $\hat{\alpha}_{k,j}$,

$$f(\mathbf{x}^{(k)} + \hat{\alpha}_{k,j}\mathbf{d}^{(k)}) - f(\mathbf{x}^{(k)}) \approx \hat{\alpha}_{k,j}\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)} \leq \beta\hat{\alpha}_{k,j}\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)},$$

  so that this method eventually converges (recall that $\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)}$ is negative).

Backtracking line search II

- Since

$$f(\mathbf{x}^{(k)} + \hat{\alpha}_{k,j}\mathbf{d}^{(k)}) \geq f(\mathbf{x}^{(k)}) + \hat{\alpha}_{k,j}\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)}$$

  as f is convex, β must be less than 1.

- For a sufficiently small $\hat{\alpha}_{k,j}$,

$$f(\mathbf{x}^{(k)} + \hat{\alpha}_{k,j}\mathbf{d}^{(k)}) - f(\mathbf{x}^{(k)}) \approx \hat{\alpha}_{k,j}\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)} \leq \beta\hat{\alpha}_{k,j}\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)},$$

  so that this method eventually converges (recall that $\nabla f(\mathbf{x}^{(k)})^\top \mathbf{d}^{(k)}$ is negative).

- Usually, β is between 0.01 and 0.3 while γ is between 0.1 and 0.8.

Properties of gradient descent

- The gradient descent method has approximately linear convergence. If there exist constants m, M such that $mI \leq \nabla^2 f(\mathbf{x}) \leq MI$, then the error $f(\mathbf{x}^{(k)}) - p^\star$ at iteration k can be bounded by

$$f(\mathbf{x}^{(k)}) - p^\star \leq c^k (f(\mathbf{x}^{(0)}) - p^\star),$$

where $c = 1 - \min(2m\beta, 2\beta\gamma m/M)$ and $p^\star$ is the optimum.

Properties of gradient descent

- The gradient descent method has approximately linear convergence. If there exist constants m, M such that $mI \leq \nabla^2 f(\mathbf{x}) \leq MI$, then the error $f(\mathbf{x}^{(k)}) - p^\star$ at iteration k can be bounded by

$$f(\mathbf{x}^{(k)}) - p^\star \leq c^k (f(\mathbf{x}^{(0)}) - p^\star),$$

  where $c = 1 - \min(2m\beta, 2\beta\gamma m/M)$ and $p^\star$ is the optimum.

- The choice of constants γ, β has a fairly limited impact. The use of exact line search in place of backtracking search improves the convergence, but not significantly.

- The condition number M/m of the Hessian has a significant impact on the rate of convergence. For problems where Hessian is poorly conditioned (say, > 1000), gradient descent method is *not* used as c is very close to one.

Newton's method

- In Newton's method, $\mathbf{d}^{(k)} = -\nabla^2 f(\mathbf{x}^{(k)})^{-1} \nabla f(\mathbf{x}^{(k)})$, where $\nabla^2 f(\mathbf{x}^{(k)})$ is Hessian matrix of function f. As before, backtracking line search or exact line search is used to choose the stepsize α_k.
- The choice of $\mathbf{v} = \mathbf{d}^{(k)}$ minimizes the right hand side of the (convex) quadratic approximation

$$f(\mathbf{x}^{(k)} + \mathbf{v}) \approx f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)})^\top \mathbf{v} + \frac{1}{2} \mathbf{v}^\top \nabla^2 f(\mathbf{x}^{(k)}) \mathbf{v}.$$

Newton's method: properties

- Assume that there exists constant m such that $mI \leq \nabla^2 f(\mathbf{x})$ and that the Hessian is Lipschitz continuous with a constant L, i.e.:

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$$

  holds over domain of f. Then there are numbers $\eta \geq 0$, $\gamma > 0$ such that

Newton's method: properties

- Assume that there exists constant m such that $mI \leq \nabla^2 f(\mathbf{x})$ and that the Hessian is Lipschitz continuous with a constant L, i.e.:

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$$

holds over domain of f. Then there are numbers $\eta \geq 0$, $\gamma > 0$ such that

1. If $\|\nabla f(\mathbf{x})^{(k)}\|_2 \geq \eta$, then $f(\mathbf{x}^{(k+1)}) \leq f(x^{(k)}) - \gamma$.

Newton's method: properties

- Assume that there exists constant m such that $mI \leq \nabla^2 f(\mathbf{x})$ and that the Hessian is Lipschitz continuous with a constant L, i.e.:

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$$

holds over domain of f. Then there are numbers $\eta \geq 0$, $\gamma > 0$ such that

1. If $\|\nabla f(\mathbf{x})^{(k)}\|_2 \geq \eta$, then $f(\mathbf{x}^{(k+1)}) \leq f(x^{(k)}) - \gamma$.
2. If $\|\nabla f(\mathbf{x})^{(k)}\|_2 < \eta$, then, for $l \geq k$,

$$f(\mathbf{x}^{(l)}) - p^\star \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{l-k+1}}.$$

- The method converges rapidly once we are close to the solution.

Newton's method: common heuristics I

- Apart from the numerical difficulties associated with inverting poorly conditioned matrices, the convergence of Newton's method does *not* depend on the conditioning of the Hessian.
- The main numerical difficulty is solving the system of linear equations

$$\nabla^2 f(\mathbf{x}^{(k)})\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)}).$$

  This problem is usually solved using Cholesky factorization.

Newton's method: common heuristics I

- Apart from the numerical difficulties associated with inverting poorly conditioned matrices, the convergence of Newton's method does *not* depend on the conditioning of the Hessian.
- The main numerical difficulty is solving the system of linear equations

$$\nabla^2 f(\mathbf{x}^{(k)})\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)}).$$

  This problem is usually solved using Cholesky factorization.
- If f is approximately quadratic, the Hessian matrix is factorized only once in every few iterations (say, 10,); $\mathbf{d}^{(k)} = H^{-1}\nabla f(\mathbf{x}^{(k)})$, where H is the last Hessian evaluated.
- If the cross terms $\frac{\partial^2 f}{\partial x_i x_j}$, $i \neq j$ are small, the Hessian is replaced by its diagonal.

Newton's method: common heuristics II

- If the Hessian is sparse (i.e. has only a few non-zero entries) and structured (i.e. some of the cross derivatives are identically 0), the system of linear equations can be solved efficiently using sparse matrix techniques.

Newton's method: common heuristics II

- If the Hessian is sparse (i.e. has only a few non-zero entries) and structured (i.e. some of the cross derivatives are identically 0), the system of linear equations can be solved efficiently using sparse matrix techniques.
- For solving

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{m} f_i(\mathbf{x})^2,$$

where $f_i(\mathbf{x})$ are twice differentiable convex functions, *Gauss Newton* method is often employed:

$$\begin{aligned}\nabla^2 f(\mathbf{x}) &= \sum_{i=1}^{m} \left(\nabla f_i(\mathbf{x}) \nabla f_i(\mathbf{x})^\top + f_i(\mathbf{x}) \nabla^2 f_i(\mathbf{x})\right), \\ &\approx \sum_{i=1}^{m} \left(\nabla f_i(\mathbf{x}) \nabla f_i(\mathbf{x})^\top\right).\end{aligned}$$

Some analysis of Gauss Newton method

- Let $g(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{m} f_i(\mathbf{x})^2$, let A be a matrix with $\nabla f_i(\mathbf{x})$ as its i^{th} column and let $\mathbf{f} = \begin{bmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & \cdots & f_m(\mathbf{x}) \end{bmatrix}^\top$.
- Then $\nabla g(\mathbf{x}) = A\mathbf{f}$ and

$$\begin{aligned} \mathbf{d}^{(k)} &= -\left\{\sum_{i=1}^{m} \left(\nabla f_i(\mathbf{x}) \nabla f_i(\mathbf{x})^\top\right)\right\}^{-1} \nabla g(\mathbf{x}) \\ &= -(AA^\top)^{-1} A\mathbf{f}, \end{aligned}$$

so that $\left(\nabla g(\mathbf{x})\right)^\top \mathbf{d}^{(k)} < 0$, as required, *i.e.* we still have a descent method.

Some analysis of Gauss Newton method

- Let $g(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{m} f_i(\mathbf{x})^2$, let A be a matrix with $\nabla f_i(\mathbf{x})$ as its i^{th} column and let $\mathbf{f} = \begin{bmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & \cdots & f_m(\mathbf{x}) \end{bmatrix}^\top$.
- Then $\nabla g(\mathbf{x}) = A\mathbf{f}$ and

$$\begin{aligned} \mathbf{d}^{(k)} &= -\left\{\sum_{i=1}^{m} \left(\nabla f_i(\mathbf{x}) \nabla f_i(\mathbf{x})^\top\right)\right\}^{-1} \nabla g(\mathbf{x}) \\ &= -(AA^\top)^{-1} A\mathbf{f}, \end{aligned}$$

  so that $\left(\nabla g(\mathbf{x})\right)^\top \mathbf{d}^{(k)} < 0$, as required, *i.e.* we still have a descent method.
- This does away entirely with computing the Hessian of individual functions $\nabla^2 f_i(\mathbf{x})$.

Constrained optimization

- Our discussion of general optimization scheme so far has avoided any constraints.

Constrained optimization

- Our discussion of general optimization scheme so far has avoided any constraints.
- We will next look at what happens when we put in constraints, starting with affine equality constraints.

Equality constrained minimization I

- Consider a problem of the form

$$\begin{aligned} &\text{minimize } f_0(\mathbf{x}) \\ &\text{subject to } A\mathbf{x} = \mathbf{b}. \end{aligned}$$

Equality constrained minimization I

- Consider a problem of the form

$$\begin{aligned} &\text{minimize } f_0(\mathbf{x}) \\ &\text{subject to } A\mathbf{x} = \mathbf{b}. \end{aligned}$$

- The simplest way (although not always the most efficient way) to deal with an affine equality constraint is to eliminate one or more variables. As an example, consider

$$\begin{aligned} &\text{minimize} \quad f(x_1) + f(x_2) + f(x_3) \\ &\text{subject to} \quad x_1 + 2x_2 = x_3, \end{aligned}$$

  where $f(x)$ is a scalar and convex.

Equality constrained minimization II

- Equivalent *unconstrained* minimization problem in two variables is:

$$\text{minimize} \quad f(x_1) + f(x_2) + f(x_1 + 2x_2).$$

Equality constrained minimization II

- Equivalent *unconstrained* minimization problem in two variables is:

$$\text{minimize} \quad f(x_1) + f(x_2) + f(x_1 + 2x_2).$$

- If $\hat{x}_1$, $\hat{x}_2$ solve this modified problem, $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3 := \hat{x}_1 + 2\hat{x}_2$ solve the original optimization problem and both yield the same optimal cost.

Inequality constrained minimization I

- Consider again a problem of the form

$$\begin{aligned} &\text{minimize } f_0(\mathbf{x}) \\ &\text{subject to } f_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m. \end{aligned}$$

where f_0, f_i are convex, differentiable.

Inequality constrained minimization I

- Consider again a problem of the form

$$\begin{aligned} &\text{minimize } f_0(\mathbf{x}) \\ &\text{subject to } f_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m. \end{aligned}$$

  where f_0, f_i are convex, differentiable.

- An equivalent unconstrained problem using an *indicator* function is:

$$\text{minimize } f_0(\mathbf{x}) + \sum_{i=1}^{m} I_-(f_i(\mathbf{x})), \tag{13}$$

  where the function $I_- : \mathbb{R} \mapsto \mathbb{R}$ is defined by

$$\begin{aligned} I_-(u) &= 0 \text{ if } u \leq 0, \\ &= \infty \text{ if } u > 0. \end{aligned}$$

Inequality constrained minimization

- A typical approximation to $I_-(u)$ is a *logarithmic barrier* function, defined by

$$\hat{I}_-(u) = -\frac{1}{t}\log(-u),$$

where $t > 0$ is chosen by the user.

Inequality constrained minimization

- A typical approximation to $I_-(u)$ is a *logarithmic barrier* function, defined by

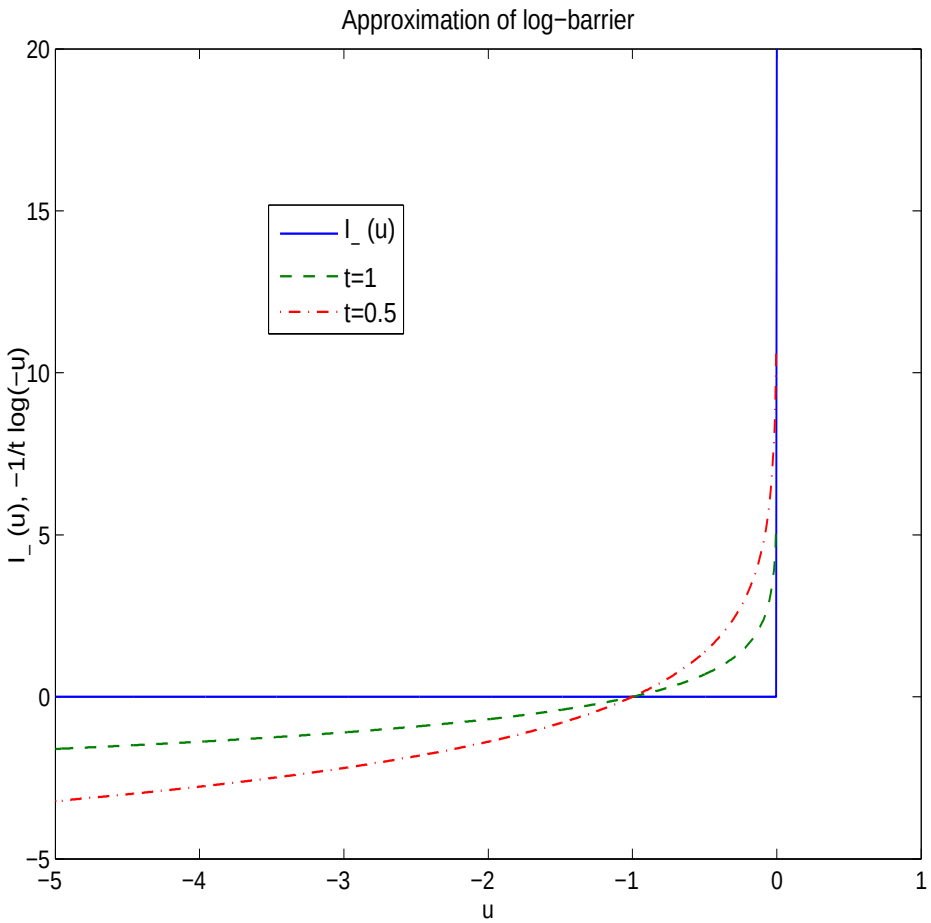
$$\hat{I}_-(u) = -\frac{1}{t}\log(-u),$$

  where $t > 0$ is chosen by the user.

- This is a convex function on non-positive reals and replacing I_- in (13) by $\hat{I}_-$ yields an unconstrained convex minimization problem with a differentiable objective.

- One can solve a sequence of unconstrained minimization problems, increasing t at each minimization step and starting each minimization with the optimal $\mathbf{x}$ obtained at the minimization at the previous step.

Approximation of log-barrier



KKT conditions under log barrier

- For simplicity of notation, we consider $\mathbf{x} \geq 0$ as the only inequality constraint.
- The modified problem is

$$\begin{aligned} \text{minimize} \quad & f_0(\mathbf{x}) - \mu \sum_{i=1}^{n} \log(x_i), \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b}, \end{aligned}$$

KKT conditions for optimality re-visited I

- LP with log barrier:

$$\text{minimize } \mathbf{c}^\top \mathbf{x} - \mu \sum_{i=1}^{n} \log(x_i),$$

$$\text{subject to } A\mathbf{x} = \mathbf{b}.$$

- Lagrangian with log barrier is

$$\mathbf{c}^\top \mathbf{x} - \mu \sum_{i=1}^{n} \log(x_i) + \boldsymbol{\nu}^\top (A\mathbf{x} - \mathbf{b})$$

KKT conditions for optimality re-visited II

- KKT conditions give

$$c + (A^\top \boldsymbol{\nu}) - \mu \operatorname{vec}\left(\frac{1}{\mathbf{x}}\right) = 0,$$

$$A\mathbf{x} - \mathbf{b} = 0,$$

$$\mathbf{x} > 0,$$

  where, for $g : \mathbb{R} \mapsto \mathbb{R}$ and $\mathbf{z} \in \mathbb{R}^N$,

$$\operatorname{vec}(g(\mathbf{z})) = \begin{bmatrix} g(z_1) & g(z_2) & \cdots & g(z_N) \end{bmatrix}^\top.$$

- In LP with log barrier, an *interior point method* finds a sequence of feasible points $\mathbf{x}^{(k)}$ with μ decreasing at eack k; each feasible point found via a Newton step.

KKT conditions for optimality re-visited III

- (Simplified) QP with log barrier:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\mathbf{x}^\top P\mathbf{x} + \mathbf{c}^\top \mathbf{x} - \mu \sum_{i=1}^{n} \log(x_i) \\ & A\mathbf{x} = \mathbf{b}. \end{aligned} \tag{14}$$

- KKT conditions give

$$\begin{aligned} & c + (A^\top \boldsymbol{\nu}) - \mu \,\text{vec}\begin{pmatrix} \frac{1}{\mathbf{x}} \end{pmatrix} + P\mathbf{x} = 0, \\ & A\mathbf{x} - \mathbf{b} = 0, \\ & \mathbf{x} > 0. \end{aligned}$$

De-tour: Quasiconvex optimization problem

- $$\min f(\mathbf{x}), \mathbf{x} \in \mathcal{F}$$

  where f is quasiconvex (has convex sublevel sets), feasible set $\mathcal{F}$ is convex.

- $\mathcal{F}$ can be defined by linear/convex quadratic inequalities and/or linear equalities, for example.

- Quasiconvex problems can be solved using a sequence of *convex feasibility problems* via bisection method.

Convex feasibility problems

- Equivalent to a convex optimization problem with $f_0(\mathbf{x}) =$ constant; find *any* feasible $\mathbf{x}$, if it exists.
- It is often non-trivial to find a feasible point!

Convex feasibility problems

- Equivalent to a convex optimization problem with $f_0(\mathbf{x}) =$ constant; find *any* feasible $\mathbf{x}$, if it exists.
- It is often non-trivial to find a feasible point!
- Typically feasibility problem is solved iteratively, using *sequential projection* type methods.
- Basic idea of iteration: given a point $\mathbf{x}^{(k)}$ in a set $\bigcap_{j=1}^{i} \mathcal{F}_j$, find the next point $\mathbf{x}^{(k+1)}$ in a set $\bigcap_{j=1}^{i+1} \mathcal{F}_j$ by using a projection of $\mathbf{x}^{(k)}$ onto $\mathcal{F}_{i+1}$.

Solving quasiconvex optimization problems

- To solve

$$\min t \text{ such that } f(\mathbf{x}) \leq t,\ \mathbf{x} \in \mathcal{F},$$

  we can use the following method:

- Given upper, lower limits u, l and a tolerance ϵ,

  1. $t_k = (u + l)/2$;
  2. Solve the feasibility problem to find $\hat{\mathbf{x}} \in \mathcal{F}$, $f(\hat{\mathbf{x}}) \leq t_k$;
  3. next $u := t_k$ if the feasible set is non-empty ($f(\mathbf{x}) \leq t_k$ for some $\mathbf{x}$ in $\mathcal{F}$); next $l := t_k$ otherwise.
  4. Iterate till $u - l \leq \epsilon$.

Solving quasiconvex optimization problems

- To solve

$$\min t \text{ such that } f(\mathbf{x}) \leq t,\ \mathbf{x} \in \mathcal{F},$$

  we can use the following method:

- Given upper, lower limits u, l and a tolerance ϵ,

  1. $t_k = (u + l)/2$;
  2. Solve the feasibility problem to find $\hat{\mathbf{x}} \in \mathcal{F}$, $f(\hat{\mathbf{x}}) \leq t_k$;
  3. next $u := t_k$ if the feasible set is non-empty ($f(\mathbf{x}) \leq t_k$ for some $\mathbf{x}$ in $\mathcal{F}$); next $l := t_k$ otherwise.
  4. Iterate till $u - l \leq \epsilon$.

- Eventually: $f(\mathbf{x}) \leq t_k$ is feasible but $f(\mathbf{x}) \leq t_{k-1}$ is not, with $|t_k - t_{k-1}| \leq \epsilon$, which is what we are looking for.

An example of a quasiconvex problem

- $$\begin{aligned} &\text{minimize } \bar{\lambda}\left\{B^{-1}(\mathbf{x})A(\mathbf{x})\right\} \\ &\text{subject to } B(\mathbf{x}) > 0, \mathbf{x} \in \mathcal{X}, \end{aligned}$$

where A, B are affine in $\mathbf{x}$, $\mathcal{X}$ is a convex set.

An example of a quasiconvex problem

- $$\begin{aligned} &\text{minimize } \bar{\lambda}\left\{B^{-1}(\mathbf{x})A(\mathbf{x})\right\} \\ &\text{subject to } B(\mathbf{x}) > 0, \mathbf{x} \in \mathcal{X}, \end{aligned}$$

  where A, B are affine in $\mathbf{x}$, $\mathcal{X}$ is a convex set.

- This reduces to an equivalent quasiconvex problem

$$\begin{aligned} &\text{minimize } t \text{ subject to} \\ &A(\mathbf{x}) \leq tB(\mathbf{x}), \\ &B(\mathbf{x}) > 0, \mathbf{x} \in \mathcal{X}, \end{aligned}$$

An example of a quasiconvex problem

- $$\begin{aligned} &\text{minimize } \bar{\lambda}\left\{B^{-1}(\mathbf{x})A(\mathbf{x})\right\} \\ &\text{subject to } B(\mathbf{x}) > 0, \mathbf{x} \in \mathcal{X}, \end{aligned}$$

  where A, B are affine in $\mathbf{x}$, $\mathcal{X}$ is a convex set.

- This reduces to an equivalent quasiconvex problem

$$\begin{aligned} &\text{minimize } t \text{ subject to} \\ &A(\mathbf{x}) \leq tB(\mathbf{x}), \\ &B(\mathbf{x}) > 0, \mathbf{x} \in \mathcal{X}, \end{aligned}$$

- If A, B are scalar functions, this is a *linear fractional program*.

What happens next?

- A small theoretical case study;

What happens next?

- A small theoretical case study;
- A lot more on interior point methods;
- Some hands-on programming;
- Enjoy the rest of the course!