Chapter 23 Algorithms for Constrained Optimization

1. Projections
2. Projected Gradient Methods with Linear Constraints
3. Lagrangian Algorithms
4. Penalty Methods

Projections

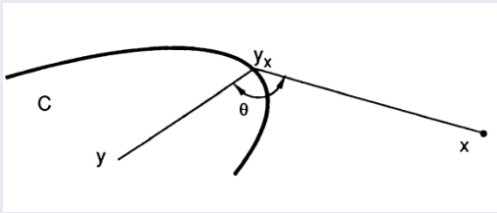
constrained optimization: $\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \Omega\}$

- $f : \mathbb{R}^n \to \mathbb{R}$ is multi-dimensional objective function.
- $\boldsymbol{x} \in \Omega \subset \mathbb{R}^n$ is n-dimensional decision variables.

★ generic linear search $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}$ for unconstrained optimization may not satisfy the constraints.

Definition (projection)

Let $\Omega \subset \mathbb{R}^n$ be nonempty closed convex set.
The projection onto Ω is defined as
$\Pi_\Omega[\boldsymbol{x}] := \arg\min\{\|\boldsymbol{z} - \boldsymbol{x}\| \mid \boldsymbol{z} \in \Omega\},\ \forall \boldsymbol{x} \in \mathbb{R}^n.$



properties of projection

- If Ω is nonempty closed convex, then the projection exists and is unique.
- $\boldsymbol{x}^*$ is the projection of $\boldsymbol{x}$ onto $\Omega \Leftrightarrow (\boldsymbol{x} - \boldsymbol{x}^*)^\top (\boldsymbol{z} - \boldsymbol{x}^*) \le 0,\ \forall \boldsymbol{z} \in \Omega$. Moreover, "=" holds $\iff$ Ω is affine.
- projection mapping is nonexpansive, $\|\Pi_\Omega[\boldsymbol{x}] - \Pi_\Omega[\boldsymbol{y}]\| \le \|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

Projected Methods

projected method for $\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \Omega\}$

- (revisit) unconstrained line search: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}$.
- projected methods: $\boldsymbol{x}^{(k+1)} = \Pi[\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}]$.
- projected gradient methods: $\boldsymbol{x}^{(k+1)} = \Pi[\boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})]$.

projection of peculiar set with explicit solution

1. If $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a} \leq \boldsymbol{x} \leq \boldsymbol{b}\}$ is a box with $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^n$, then $[\Pi_\Omega(\boldsymbol{x})]_i$ =median$\{a_i, x_i, b_i\}$ for $i = 1, 2, ..., n$.
2. If $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_2 \leq \boldsymbol{a}\}$ is a ball with $\boldsymbol{a} > 0$, then $\Pi_\Omega(\boldsymbol{x}) = \min(1, \frac{\boldsymbol{a}}{\|\boldsymbol{x}\|_2})\boldsymbol{x}$.
3. If $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ is an affine set with $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$, then $\Pi_\Omega(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{A}^\dagger(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$, where $\boldsymbol{A}^\dagger$ is the Moore-Penrose pseudo-inverse of A. Particularly, if $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}^\top \boldsymbol{x} = \boldsymbol{b}, \boldsymbol{a} \neq 0 \in \mathbb{R}^n\}$ is a nonvertical hypeplane, then $\Pi_\Omega(\boldsymbol{x}) = \boldsymbol{x} - \frac{\boldsymbol{a}^\top \boldsymbol{x} - \boldsymbol{b}}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a}$.

More examples: Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, 2006.

Projected Gradient Methods (PGM)

Example ($\min\ \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x},\ \text{s.t. } \|\boldsymbol{x}\|^2 = 1, \text{ where } \boldsymbol{Q} = \boldsymbol{Q}^\top \succ 0$)

1. Derive a formula for the update equation for the algorithm.
2. Whether the algorithm converges to an optimum or not?
3. Show that for $0 < \alpha < \frac{1}{\lambda_{\max}}$ (where $\lambda_{\max}$ is the largest eigenvalue of $\boldsymbol{Q}$), the algorithm converges to an optimum, provided that $\boldsymbol{x}^{(0)}$ is not orthogonal to the eigenvectors of $\boldsymbol{Q}$ corresponding to the smallest eigenvalue.

1. $\boldsymbol{x}^{(k+1)} = \beta_k(\boldsymbol{x}^{(k)} - \alpha\boldsymbol{Q}\boldsymbol{x}^{(k)}) = \beta_k(\boldsymbol{I} - \alpha\boldsymbol{Q})\boldsymbol{x}^{(k)}$, where $\beta_k = \frac{1}{\|(\boldsymbol{I}-\alpha\boldsymbol{Q})\boldsymbol{x}^{(k)}\|}$.
2. If we start with $\boldsymbol{x}^{(0)}$ being a unit eigenvector of $\boldsymbol{Q}$, it is very easy to check that $\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(0)}$ for all k.
3. Let $\{\boldsymbol{v}_i\}_{i=1}^n$ be the linear independent eigenvector corresponding to the eigenvalue $\{\lambda_i(\boldsymbol{Q})\}_{i=1}^n$. We have $\{\boldsymbol{v}_i\}_{i=1}^n$ is basis of $\mathbb{R}^n$. Thus,

$$\boldsymbol{x}^{(k)} = y_1^{(k)}\boldsymbol{v}_1 + \cdots + y_n^{(k)}\boldsymbol{v}_n, \quad (1)$$

$$\boldsymbol{x}^{(k+1)} = y_1^{(k+1)}\boldsymbol{v}_1 + \cdots + y_n^{(k+1)}\boldsymbol{v}_n. \quad (2)$$

Projected Gradient Methods (PGM)

It follows from the iterative scheme in item 1 that

$$\begin{aligned}\boldsymbol{x}^{(k+1)} &= \beta_k(\boldsymbol{I}-\alpha\boldsymbol{Q})\boldsymbol{x}^{(k)} = \beta_k(\boldsymbol{I}-\alpha\boldsymbol{Q})\left(y_1^{(k)}\boldsymbol{v}_1+\cdots+y_n^{(k)}\boldsymbol{v}_n\right)\\ &= \beta_k\left(y_1^{(k)}(\boldsymbol{I}-\alpha\boldsymbol{Q})\boldsymbol{v}_1+\cdots+y_n^{(k)}(\boldsymbol{I}-\alpha\boldsymbol{Q})\boldsymbol{v}_n\right)\\ &= \beta_k\left(y_1^{(k)}\left(1-\alpha\lambda_1\right)\boldsymbol{v}_1+\cdots+y_n^{(k)}\left(1-\alpha\lambda_n\right)\boldsymbol{v}_n\right).\end{aligned}$$

Compared to (2), it yields $y_i^{(k+1)} = \beta_k y_i^{(k)}\left(1-\alpha\lambda_i\right)$.

$\therefore\ y_i^{(k)} = (\Pi_{i=0}^{k-1}\beta_k)y_i^{(0)}\left(1-\alpha\lambda_i\right)^k$.

It follows from (1) that $\boldsymbol{x}^{(k)} = \sum_{i=1}^{n} y_i^{(k)}\boldsymbol{v}_i = y_1^{(k)}\left(\boldsymbol{v}_1+\sum_{i=2}^{n}\frac{y_i^{(k)}}{y_1^{(k)}}\boldsymbol{v}_i\right)$.

Assume that $y_1^{(0)}\neq 0$, we obtain $\frac{y_i^{(k)}}{y_1^{(k)}} = \frac{y_i^{(0)}(1-\alpha\lambda_i)^k}{y_1^{(0)}(1-\alpha\lambda_1)^k} = \frac{y_i^{(0)}}{y_1^{(0)}}\left(\frac{1-\alpha\lambda_i}{1-\alpha\lambda_1}\right)^k$.

Using the fact that $\frac{1-\alpha\lambda_i}{1-\alpha\lambda_1}<1$, we deduce that $\frac{y_i^{(k)}}{y_1^{(k)}}\to 0$,

which implies that $\boldsymbol{x}^{(k)}\to\boldsymbol{v}_1$.

PGM with Linear Constraints

constrained optimization: $\min f(\boldsymbol{x})$, s.t. $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

- $f : \mathbb{R}^n \to \mathbb{R}$ is multi-dimensional objective function.
- $\boldsymbol{x} \in \mathbb{R}^n$ is n-dimensional decision variables.
- $\boldsymbol{A} \in \mathbb{R}^{m \times n} (m < n)$, $\text{rank}(\boldsymbol{A}) = m$, $\boldsymbol{b} \in \mathbb{R}^m$.

Lemma

Let $\Omega = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$. Then the projection onto Ω can be defined by the orthogonal projector matrix $\boldsymbol{P}$, i.e.,

$$\Pi_\Omega = \boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{A}^\top)^{-1} \boldsymbol{A}. \tag{3}$$

key point of proof: By the KKT condition of constrained optimization.

Lemma

For the projection matrix in (3)*, it holds that* $\mathcal{N}(\boldsymbol{P}) = \mathcal{R}\left(\boldsymbol{A}^\top\right)$, $\mathcal{N}(\boldsymbol{A}) = \mathcal{R}(\boldsymbol{P})$.

proof. $\Rightarrow$) For any $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{P})$, we have

$$0 = \boldsymbol{P}\boldsymbol{v} = \left(\boldsymbol{I}_n - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\right)\boldsymbol{v} = \boldsymbol{v} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\boldsymbol{v}.$$

$\therefore \boldsymbol{v} = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\boldsymbol{v}$, and hence $\boldsymbol{v} \in \mathcal{R}(\boldsymbol{A}^\top)$.

$\Leftarrow$) For any $\boldsymbol{v} \in \mathcal{R}(\boldsymbol{A}^\top)$, then there exists $\boldsymbol{u} \in \mathbb{R}^m$ such that $\boldsymbol{v} = \boldsymbol{A}^\top\boldsymbol{u}$.

$$\therefore \boldsymbol{P}\boldsymbol{v} = \left(\boldsymbol{I}_n - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\right)\boldsymbol{A}^\top\boldsymbol{u}$$

$$= \boldsymbol{A}^\top\boldsymbol{u} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u} = 0.$$

Hence, we have proved that $\mathcal{N}(\boldsymbol{P}) = \mathcal{R}\left(\boldsymbol{A}^\top\right)$.

By arguments similar to the above, we can show $\mathcal{N}(\boldsymbol{A}) = \mathcal{R}\left(\boldsymbol{P}\right)$.

PGM with Linear Constraints

revisit: for the unconstrained optimization $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$, the first-order necessary condition (FONC) for a point $\boldsymbol{x}^*$ to be a local minimizer is $\nabla f(\boldsymbol{x}^*) = 0$.

Theorem

For constrained optimization $\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \Omega\}$ with $\Omega = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$, the FONC for a point $\boldsymbol{x}^*$ to be a local minimizer is $\boldsymbol{P}\nabla f(\boldsymbol{x}^*) = 0$.

proof. $\boldsymbol{P}\nabla f(\boldsymbol{x}^*) = 0$,

$\Longleftrightarrow \nabla f(\boldsymbol{x}^*) \in \mathcal{N}(\boldsymbol{P})$, [by the definition of $\mathcal{N}(\boldsymbol{P})$]

$\Longleftrightarrow \nabla f(\boldsymbol{x}^*) \in \mathcal{R}(\boldsymbol{A}^\top)$, [by the above lemma]

$\Longleftrightarrow \exists \boldsymbol{\lambda}^* \in \mathbb{R}^m$, such that $\nabla f(\boldsymbol{x}^*) + \boldsymbol{A}^\top \boldsymbol{\lambda}^* = 0$,

$\Longleftrightarrow$ by combining $\boldsymbol{x}^* \in \Omega = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$,

we can obtain the KKT condition

$$\begin{cases} \nabla f(\boldsymbol{x}^*) + \boldsymbol{A}^\top \boldsymbol{\lambda}^* = 0, \\ \boldsymbol{A}^\top \boldsymbol{x}^* = \boldsymbol{b}. \end{cases}$$

PGM with Linear Constraints

iterative scheme of projection gradient method (PGM)

$\min\{f(\boldsymbol{x}) \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$, the iterative scheme of PGM is

$$\boldsymbol{x}^{(k+1)} = \Pi[\boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})] \overset{\text{why?}}{=\!=\!=} \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}).$$

Theorem

In PGM, if $\boldsymbol{x}^{(0)}$ is feasible, then each $\boldsymbol{x}^{(k)}$ is feasible, i.e., $\boldsymbol{A}\boldsymbol{x}^{(k)} = \boldsymbol{b}$, $\forall k \geq 0$.

proof. (by induction).

- The result holds for $k = 0$ by assumption.
- Suppose that $\boldsymbol{A}\boldsymbol{x}^{(k)} = \boldsymbol{b}$. We now show that $\boldsymbol{A}\boldsymbol{x}^{(k+1)} = \boldsymbol{b}$.

$$\because \boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}) \in \mathcal{R}(\boldsymbol{P}) \overset{\text{lemma}}{\Longrightarrow} \boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}) \in \mathcal{N}(\boldsymbol{A}).$$

$$\begin{aligned} \therefore \boldsymbol{A}\boldsymbol{x}^{(k+1)} &= \boldsymbol{A}(\boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)})) \\ &= \boldsymbol{A}\boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{A}\boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}) \\ &= \boldsymbol{b}. \quad (\text{why? Ans: } \boldsymbol{A}\boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}) = 0) \end{aligned}$$

Properties of PGM with Linear Constraints

Theorem

If $\{\boldsymbol{x}^{(k)}\}$ is a sequence generated by the projected steepest descent method and if $\boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}) \neq \boldsymbol{0}$, then $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$.

proof. $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)})$ with $\alpha_k \geq 0$ is the exact step size.

Let $\phi_k(\alpha) := f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}))$.

Then, $\phi_k(\alpha_k) \leq \phi_k(\alpha)$ for all $\alpha \geq 0$.

By the chain rule,

$$\begin{aligned}\phi_k'(0) &= -\nabla f(\boldsymbol{x}^{(k)} - 0\boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}))^\top \boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}) \\ &= -\nabla f(\boldsymbol{x}^{(k)})^\top \boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)}) \\ &= -\|\boldsymbol{P}\nabla f(\boldsymbol{x}^{(k)})\|^2 < 0. \quad [\because \boldsymbol{P} = \boldsymbol{P}^2 = \boldsymbol{P}^\top \boldsymbol{P}]\end{aligned}$$

$\therefore \exists \bar{\alpha} > 0$ such that $\phi_k(0) > \phi_k(\alpha)$, $\forall \alpha \in (0, \bar{\alpha}]$.

$\therefore f(\boldsymbol{x}^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(\bar{\alpha}) < \phi_k(0) = f(\boldsymbol{x}^{(k)})$.

PGM for Linear Program

(LP) $\min \ \boldsymbol{c}^\top \boldsymbol{x}, \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{c} \in \mathbb{R}^n$.

Lemma

Let the solution set of (LP) be nonempty, then for any $\boldsymbol{x}_0 \in \mathbb{R}^n$, there exists $\varepsilon_0 > 0$, such that for all $\varepsilon \in (0, \varepsilon_0]$,

$$\text{(LP)} \iff \boxed{\min \ \frac{\varepsilon}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \boldsymbol{c}^\top \boldsymbol{x}, \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}.}$$

Proof: By using the KKT condition.

PGM for Linear Program

discussion

$$\boxed{\min\ \boldsymbol{c}^\top \boldsymbol{x},\quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}.}$$

$$\Updownarrow$$

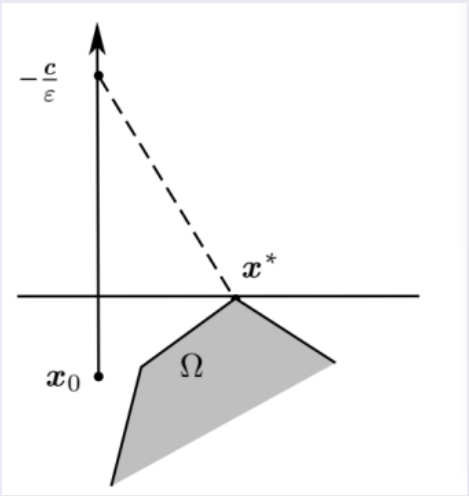
$$\boxed{\min\ \frac{\varepsilon}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|^2 + \boldsymbol{c}^\top \boldsymbol{x},\quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}.}$$

$$\Updownarrow$$

$$\boxed{\min\ \frac{\varepsilon}{2}\left\|\boldsymbol{x}-(\boldsymbol{x}_0-\frac{\boldsymbol{c}}{\varepsilon})\right\|^2,\quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}.}$$

$$\Updownarrow$$

$$\boxed{\text{Projection of } (\boldsymbol{x}_0-\frac{\boldsymbol{c}}{\varepsilon}) \text{ onto the feasible set } \Omega=\{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}\}.}$$



★ For any initial point x_0, as long as the step size is large enough for projected gradient algorithm, the optimum is obtained in one step (Mangasarian,1979).

Lagrangian Algorithms for Equality Constraints

motivation for Lagrangian algorithms

By using gradient algorithms to update simultaneously the decision variable and Lagrange multiplier vector.

equality constrained optimization

$$\boxed{\min\ f(\boldsymbol{x}),\ \ \text{s.t.}\ \ \boldsymbol{h}(\boldsymbol{x})=0,\ \ \text{where}\ \ \boldsymbol{h}:\mathbb{R}^n\to\mathbb{R}^m.} \tag{4}$$

Lagrangian function of (4) is $l(\boldsymbol{x},\boldsymbol{\lambda})=f(\boldsymbol{x})+\boldsymbol{\lambda}^\top\boldsymbol{h}(\boldsymbol{x})$.

Iterative scheme of Lagrangian method for (4) is

$$\begin{cases}\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\alpha_k\nabla_{\boldsymbol{x}}l(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)})=\boldsymbol{x}^{(k)}-\alpha_k(\nabla f(\boldsymbol{x}^{(k)})+\boldsymbol{Dh}(\boldsymbol{x}^{(k)})^\top\boldsymbol{\lambda}^{(k)}),\\ \boldsymbol{\lambda}^{(k+1)}=\boldsymbol{\lambda}^{(k)}+\beta_k\nabla_{\boldsymbol{\lambda}}l(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)})=\boldsymbol{\lambda}^{(k)}+\beta_k\boldsymbol{h}(\boldsymbol{x}^{(k)}).\end{cases} \tag{5}$$

★ $\boldsymbol{x}^{(k)}$ is a gradient method for minimizing Lagrangian w.r.t $\boldsymbol{x}$ variable;
$\boldsymbol{\lambda}^{(k)}$ is a gradient method for maximizing Lagrangian w.r.t $\boldsymbol{\lambda}$ variable.

Lagrangian Algorithms for Equality Constraints

Lemma

When Lagrangian method (5) *updates* $\boldsymbol{x}^{(k)}$ *and* $\boldsymbol{\lambda}^{(k)}$, *the pair* $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ *is a fixed point if and only if it satisfies the Lagrange condition.*

Theorem

When Lagrangian method (5) *updates* $\boldsymbol{x}^{(k)}$ *and* $\boldsymbol{\lambda}^{(k)}$, *provided that* α_k *and* β_k *are sufficiently small, there is a neighborhood of* $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ *such that: if* $(\boldsymbol{x}^{(0)}, \boldsymbol{\lambda}^{(0)})$ *is in this neighborhood, then the method converges to* $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ *with at least a linear order of convergence.*

(Proof: By using the contraction mapping fixed point theorem.)

Lagrangian Algorithms for Inequality Constraints

inequality constrained optimization

$$\boxed{\min\ f(\boldsymbol{x}),\quad \text{s.t. } \boldsymbol{g}(\boldsymbol{x}) \le 0,\ \text{ where } \boldsymbol{g}:\mathbb{R}^n \to \mathbb{R}^p.} \tag{6}$$

Lagrangian function of (6): $l(\boldsymbol{x}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \boldsymbol{\mu}^\top \boldsymbol{g}(\boldsymbol{x}),\ \boldsymbol{\mu} \ge 0.$

Lagrangian method for (6):

$$\begin{cases} \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k(\nabla f(\boldsymbol{x}^{(k)}) + \boldsymbol{D}\boldsymbol{g}(\boldsymbol{x}^{(k)})^\top \boldsymbol{\mu}^{(k)}), \\ \boldsymbol{\mu}^{(k+1)} = [\boldsymbol{\mu}^{(k)} + \beta_k \boldsymbol{g}(\boldsymbol{x}^{(k)})]_+, \quad \text{where } [\cdot]_+ = \max\{\cdot, 0\}. \end{cases} \tag{7}$$

★ $\boldsymbol{x}^{(k)}$ is a gradient method for minimizing Lagrangian w.r.t its $\boldsymbol{x}$ variable;
$\boldsymbol{\mu}^{(k)}$ is a projected gradient method for maximizing Lagrangian w.r.t $\boldsymbol{\mu}$.

Lagrangian Algorithms for Inequality Constraints

Lemma

When Lagrangian method (7) updates $\boldsymbol{x}^{(k)}$ and $\boldsymbol{\mu}^{(k)}$, the pair $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a fixed point if and only if it satisfies the KKT condition.

Theorem

When Lagrangian method (7) updates $\boldsymbol{x}^{(k)}$ and $\boldsymbol{\mu}^{(k)}$, provided that α_k and β_k are sufficiently small, there is a neighborhood of $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ such that if the pair $(\boldsymbol{x}^{(0)}, \boldsymbol{\mu}^{(0)})$ is in this neighborhood, then

1. *the nonactive multipliers reduce to zero in finite time and remain at zero thereafter;*
2. *the algorithm converges to $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ with at least a linear order of convergence.*

proof: by using contraction mapping fixed point theorem and nonexpansive mapping theorem.

Penalty Method

motivation for penalty method

The constrained optimization is approximated and solved by a sequence of unconstrained optimization.

Definition

indicator function Let $\Omega \subset \mathbb{R}^n$ be nonempty. The indicator function is defined by

$$\iota_\Omega(\boldsymbol{x}) := \begin{cases} 0, & \text{if } \boldsymbol{x} \in \Omega. \\ +\infty, & \text{otherwise .} \end{cases}$$

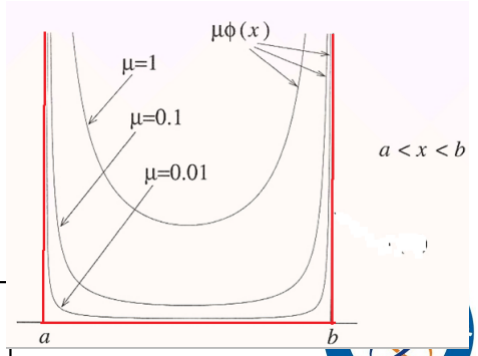
constrained optimization: $\min \quad f(\boldsymbol{x}), \quad \text{s.t.} \quad \boldsymbol{x} \in \Omega.$

$\Updownarrow$ by definition of ι_Ω

unconstrained optimization: $\min \quad f(\boldsymbol{x}) + \iota_\Omega(\boldsymbol{x}).$

$\Downarrow$ approximated by following problem

unconstrained optimization: $\min \quad f(\boldsymbol{x}) + \gamma \boldsymbol{P}(\boldsymbol{x}),$
where $\gamma > 0$ is a constant; $\boldsymbol{P} : \mathbb{R}^n \to \mathbb{R}$ is a given function.



Penalty Method

$$\text{(I)} \quad \begin{array}{ll} \min & f(\boldsymbol{x}), \\ \text{s.t.} & \boldsymbol{x} \in \Omega. \end{array} \quad \approx \quad \text{(II)} \quad \begin{array}{l} \min \ f(\boldsymbol{x}) + \gamma \boldsymbol{P}(\boldsymbol{x}), \\ \text{where } \gamma > 0 \text{ is a constant;} \\ \boldsymbol{P} : \mathbb{R}^n \to \mathbb{R} \text{ is a given function.} \end{array}$$

★ γ is called the penalty parameter, $\boldsymbol{P}$ is called the penalty function.

Definition (penalty function)

A function $\boldsymbol{P} : \mathbb{R}^n \mapsto \mathbb{R}$ is called a penalty function for the constrained optimization if it satisfies the following three conditions:

1. $\boldsymbol{P}$ is continuous.
2. $\boldsymbol{P}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.
3. $\boldsymbol{P}(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x}$ is feasible (i.e., $\boldsymbol{x} \in \Omega$).

★ The accuracy of (II) for approximating (I) depends on γ and $\boldsymbol{P}$.

★ The larger the γ, the closer the approximation. Ideally, as $\gamma \to +\infty$, the (II) should yield the true solution to (I).

Penalty Methods

Given $\boldsymbol{h}: \mathbb{R}^n \to \mathbb{R}^p$, $\boldsymbol{g}: \mathbb{R}^n \to \mathbb{R}^m$.

constrained optimization: $\min\ f(\boldsymbol{x})$, s.t. $\boldsymbol{h}(\boldsymbol{x}) = 0,\ \boldsymbol{g}(\boldsymbol{x}) \le 0.$

- exact penalty function:

$$\min f(\boldsymbol{x}) + \gamma \left(\sum_{i=1}^{p} |\boldsymbol{h}_i(\boldsymbol{x})| + \sum_{i=1}^{m} |\boldsymbol{g}_i^+(\boldsymbol{x})| \right),$$

where $\boldsymbol{g}_i^+(\boldsymbol{x}) = \max\{0, \boldsymbol{g}_i(\boldsymbol{x})\} = \begin{cases} 0, & \boldsymbol{g}_i(\boldsymbol{x}) \le 0. \\ \boldsymbol{g}_i(\boldsymbol{x}), & \boldsymbol{g}_i(\boldsymbol{x}) > 0. \end{cases}$

- quadratic penalty function: $\min f(\boldsymbol{x}) + \gamma \left(\sum_{i=1}^{p} |\boldsymbol{h}_i(\boldsymbol{x})|^2 + \sum_{i=1}^{m} |\boldsymbol{g}_i^+(\boldsymbol{x})|^2 \right).$
- Generalized penalty function:
$\min f(\boldsymbol{x}) + \boldsymbol{P}(\boldsymbol{x}) = f(\boldsymbol{x}) + \gamma \left(\|\boldsymbol{h}(\boldsymbol{x})\|_b^a + \|[\boldsymbol{g}(\boldsymbol{x})]^+\|_b^a \right),\ a > 0, b > 0.$

★ exact penalty function, nondifferentiable; quadratic penalty, generally differentiable.

Penalty Methods

constraint $\Omega = \{x \in \mathbb{R} \mid g_1(x) = x - 2 \le 0, g_2(x) = -(x+1)^3 \le 0\}$

Ans: Indeed, $\Omega = [-1, 2]$.

$$g_1^+(x) = \max\{0, g_1(x)\} = \begin{cases} 0, & \text{if } x \le 2, \\ x - 2, & \text{otherwise.} \end{cases}$$

$$g_2^+(x) = \max\{0, g_2(x)\} = \begin{cases} 0, & \text{if } x \ge -1, \\ -(x+1)^3, & \text{otherwise.} \end{cases}$$

Thus, the exact penalty function is

$$P(x) = g_1^+(x) + g_2^+(x) = \begin{cases} x - 2, & \text{if } x > 2, \\ 0, & \text{if } -1 \le x \le 2, \\ -(x+1)^3, & \text{if } x < -1. \end{cases}$$

Penalty Methods for Equality Constraints

Example (solve by quadratic penalty function method)

$\min\ f(\boldsymbol{x}) = -x_1 x_2,\ \text{s.t. } g(\boldsymbol{x}) = x_1 + 2x_2 - 4 = 0.$

Ans: quadratic penalty function is

$$\min_{\boldsymbol{x}}\ q(\boldsymbol{x}) = -x_1 x_2 + \frac{1}{2}\gamma\left(x_1 + 2x_2 - 4\right)^2. \tag{8}$$

By solving unconstrained optimization (8) with parameter $\gamma > 0$, we have

$$\begin{cases} -x_2 + \gamma\left(x_1 + 2x_2 - 4\right) = 0 \\ -x_1 + 2\gamma\left(x_1 + 2x_2 - 4\right) = 0 \end{cases} \Rightarrow \begin{cases} x_1(\gamma) = \frac{8\gamma}{4\gamma - 1} \\ x_2(\gamma) = \frac{4\gamma}{4\gamma - 1} \end{cases} \xrightarrow{\gamma \to +\infty} \begin{cases} x_1 = 2 \\ x_2 = 1 \end{cases}$$

By comparing the optimality condition of (8) and KKT of original constrained optimization, it yields a dual solution

$$\lambda(\gamma) \overset{\text{why?}}{=\!=} -\gamma g(\boldsymbol{x}(\gamma)) = \frac{-4\gamma}{4\gamma - 1} \xrightarrow{\gamma \to +\infty} \lambda = -1.$$

Penalty Methods for Equality Constraints

$\min \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$, s.t. $\|\boldsymbol{x}\|^2 = 1$, where $\boldsymbol{Q} = \boldsymbol{Q}^\top$.

1. write the quadratic penalty function with penalty parameter γ.
2. show that $\boldsymbol{x}_\gamma$ is an eigenvector of $\boldsymbol{Q}$ for any γ.
3. show that $\|\boldsymbol{x}_\gamma\|^2 - 1 = O(1/\gamma)$, as $\gamma \to +\infty$.

Ans: quadratic penalty function: $\min\ q(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \gamma \left(\|\boldsymbol{x}\|^2 - 1\right)^2$.

As $\boldsymbol{x}_\gamma$ is a minimizer of $q(\boldsymbol{x})$, it follows from optimality condition for unconstrained optimization that

$$\nabla q(\boldsymbol{x}_\gamma) = 0 \Longrightarrow 2\boldsymbol{Q}\boldsymbol{x}_\gamma + 4\gamma \left(\|\boldsymbol{x}_\gamma\|^2 - 1\right) \boldsymbol{x}_\gamma = 0,$$

i.e., $\boldsymbol{Q}\boldsymbol{x}_\gamma = \underbrace{2\gamma \left(1 - \|\boldsymbol{x}_\gamma\|^2\right)}_{=:\lambda_\gamma} \boldsymbol{x}_\gamma =: \lambda_\gamma \boldsymbol{x}_\gamma$, $\therefore \boldsymbol{x}_\gamma$ is an eigenvector of $\boldsymbol{Q}$.

$$\lambda_\gamma = 2\gamma \left(1 - \|\boldsymbol{x}_\gamma\|^2\right) \le \lambda_{\max},$$

$$\therefore \|\boldsymbol{x}_\gamma\|^2 - 1 = -\frac{\lambda_{\max}}{2\gamma} = O(\tfrac{1}{\gamma}) \text{ as } \gamma \to \infty.$$

Penalty Methods for Inequality Constraints

solve constrained optimization by quadratic penalty function method

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) = x_1 + x_2, \\ \text{s.t} \quad & g_1(\boldsymbol{x}) = -x_1^2 + 2x_2 \geq 0,\ g_2(\boldsymbol{x}) = x_1 \geq 0. \end{aligned}$$

Ans: quadratic penalty function is

$$\min\ q(\boldsymbol{x}) = f(\boldsymbol{x}) + \gamma \left[\min\left(0, -x_1^2 + x_2\right)\right]^2 + \gamma \left[\min\left(0, x_1\right)\right]^2.$$

By solving the above unconstrained optimization problem with parameter γ,

$$\begin{cases} 1 + 2\gamma[\min(0, -x_1^2 + x_2)](-2x_1) + 2\gamma[\min(0, x_1)] = 0, \\ 1 + 2\gamma[\min(0, -x_1^2 + x_2)] = 0, \end{cases}$$

$$\Rightarrow \begin{cases} 1 + 4\gamma x_1\left(x_1^2 - x_2\right) + 2\gamma x_1 = 0 \\ 1 - 2\gamma\left(x_1^2 - x_2\right) = 0 \end{cases} \Rightarrow \begin{cases} x_1 = -\frac{1}{2+2\gamma} \\ x_2 = \frac{1}{(2+2\gamma)^2} - \frac{1}{2\gamma} \end{cases}$$

$$\xrightarrow{\gamma \to +\infty} \begin{cases} x_1 = 0. \\ x_2 = 0. \end{cases}$$ (check the solution by KKT)

Penalty Methods

constrained optimization with only inequality constraints

$\min\ f(\boldsymbol{x}),\ \ \text{s.t. } \boldsymbol{g}(\boldsymbol{x}) \geq 0$, where $\boldsymbol{g}: \mathbb{R}^n \to \mathbb{R}^m$.

1. logarithmic penalty function: $\min\ q(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{1}{\gamma}\left(\sum_{i=1}^{m} \log(\boldsymbol{g}_i(\boldsymbol{x}))\right)$.
2. reciprocal penalty function: $\min\ q(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{1}{\gamma}\left(\sum_{i=1}^{m} \frac{1}{\boldsymbol{g}_i(\boldsymbol{x})}\right)$.

solve constrained optimization by logarithmic penalty function method

$\min f(\boldsymbol{x}) = x_1 - 2x_2, \quad \text{s.t. } 1 + x_1 - x_2^2 \geq 0,\ x_2 \geq 0.$

Ans: logarithmic penalty function is

$\min\ q(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{1}{\gamma}[\log(1 + x_1 - x_2^2) + \log(x_2)]$.

By solving the above unconstrained optimization

$$\begin{cases} 1 - \frac{1}{\gamma(1+x_1-x_2^2)} = 0 \\ -2 + \frac{2x_2}{\gamma(1+x_1-x_2^2)} - \frac{1}{\gamma x_2} = 0 \end{cases} \Rightarrow \begin{cases} x_1(\gamma) = \frac{\sqrt{\gamma^2+2\gamma}+3-\gamma}{2\gamma} \\ x_2(\gamma) = \frac{\gamma+\sqrt{\gamma^2+2\gamma}}{2\gamma} \end{cases} \xrightarrow{\gamma\uparrow+\infty} \begin{cases} x_1 = 0 \\ x_2 = 1 \end{cases}$$

Penalty Methods

Definition (associated function)

Let $\{\gamma_k\}_{k=0}^{\infty}$ be a positive sequence. The associated function $q(\gamma_k, \cdot) : \mathbb{R}^n \to \mathbb{R}$, is defined by

$$q(\gamma_k, \boldsymbol{x}) = f(\boldsymbol{x}) + \gamma_k P(\boldsymbol{x}). \tag{9}$$

★ γ_k are dynamic penalty factors. Let $\boldsymbol{x}^{(k)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} q(\gamma_k, \boldsymbol{x})$.

Lemma

Suppose the sequence $\{\gamma_k\}$ is positive and nondecreasing. Then,

1. $q\left(\gamma_{k+1}, \boldsymbol{x}^{(k+1)}\right) \geq q\left(\gamma_k, \boldsymbol{x}^{(k)}\right)$.
2. $P\left(\boldsymbol{x}^{(k+1)}\right) \leq P\left(\boldsymbol{x}^{(k)}\right)$.
3. $f\left(\boldsymbol{x}^{(k+1)}\right) \geq f\left(\boldsymbol{x}^{(k)}\right)$.
4. $f\left(\boldsymbol{x}^{*}\right) \geq q\left(\gamma_k, \boldsymbol{x}^{(k)}\right) \geq f\left(\boldsymbol{x}^{(k)}\right)$.

1. By the definition of q in (9), the monotonicity of $\{\gamma_k\}$ and $\boldsymbol{x}^{(k)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} q(\gamma_k, \boldsymbol{x})$, we have

$$\begin{aligned} q(\gamma_{k+1}, \boldsymbol{x}^{(k+1)}) &= f(\boldsymbol{x}^{(k+1)}) + \gamma_{k+1} P(\boldsymbol{x}^{(k+1)}) \geq f(\boldsymbol{x}^{(k+1)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}) \\ &\geq f(\boldsymbol{x}^{(k+1)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}) = q(\gamma_k, \boldsymbol{x}^{(k)}). \end{aligned}$$

Penalty Methods

2. $\because \boldsymbol{x}^{(k)} = \arg\min q(\gamma_k, \boldsymbol{x})$ and $\boldsymbol{x}^{(k+1)} = \arg\min q(\gamma_{k+1}, \boldsymbol{x})$. Thus,
$$\begin{cases} q(\gamma_k, \boldsymbol{x}^{(k)}) = f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}) \leq f(\boldsymbol{x}^{(k+1)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}) \\ q(\gamma_{k+1}, \boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k+1)}) + \gamma_{k+1} P(\boldsymbol{x}^{(k+1)}) \leq f\boldsymbol{x}^{(k)}) + \gamma_{k+1} P(\boldsymbol{x}^{(k)}) \end{cases}$$
By adding the above inequalities, it yields
$$\gamma_k P(\boldsymbol{x}^{(k)}) + \gamma_{k+1} P(\boldsymbol{x}^{(k+1)}) \leq \gamma_{k+1} P(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}).$$
By rearranging terms, we get $P(\boldsymbol{x}^{(k+1)}) \leq P(\boldsymbol{x}^{(k)})$.

3. $\because \boldsymbol{x}^{(k)} = \arg\min q(\gamma_k, \boldsymbol{x})$
$\therefore\ q(\gamma_k, \boldsymbol{x}^{(k)}) = f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}) \leq f(\boldsymbol{x}^{(k+1)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}).$
$\therefore\ f(\boldsymbol{x}^{(k+1)}) \geq f(\boldsymbol{x}^{(k)}) + \gamma_k (P(\boldsymbol{x}^{(k)}) - P(\boldsymbol{x}^{(k+1)})).$
By item 2, we get $f(\boldsymbol{x}^{(k+1)}) \geq f(\boldsymbol{x}^{(k)})$.

4. $\because \boldsymbol{x}^{(k)} = \arg\min q(\gamma_k, \boldsymbol{x})$
$\therefore\ f(\boldsymbol{x}^*) + \gamma_k P(\boldsymbol{x}^*) \geq q(\gamma_k, \boldsymbol{x}^{(k)}) = f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}).$
$\because \boldsymbol{x}^*$ is a minimizer of constrained optimization problem,
$\therefore\ P(x^*) = 0$ and $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}).$
$\because\ P(\boldsymbol{x}^{(k)}) \geq 0$ and $\gamma_k \geq 0$,
$\therefore\ f(\boldsymbol{x}^*) \geq q(\gamma_k, \boldsymbol{x}^{(k)}) \geq f(\boldsymbol{x}^{(k)}).$

Theorem

Suppose that f is continuous and $\gamma_k \to \infty$ as $k \to \infty$. Then, any limit point of $\left\{\boldsymbol{x}^{(k)}\right\}$ is a solution to the constrained optimization problem.

proof: suppose $\left\{\boldsymbol{x}^{(m_k)}\right\}$ is a subsequence of $\left\{\boldsymbol{x}^{(k)}\right\}$ converging to $\hat{\boldsymbol{x}}$.

By above Lemma, $\left\{q\left(\gamma_k, \boldsymbol{x}^{(k)}\right)\right\}$ is nondecreasing and bounded above by $f\left(\boldsymbol{x}^*\right)$.

$\therefore$ $\left\{q\left(\gamma_k, \boldsymbol{x}^{(k)}\right)\right\}$ has a limit, i.e., $q^* = \lim_{k\to\infty} q\left(\gamma_k, \boldsymbol{x}^{(k)}\right)$ such that $q^* \le f\left(\boldsymbol{x}^*\right)$.

$\because$ f is continuous and $f\left(\boldsymbol{x}^{(m_k)}\right) \le f\left(\boldsymbol{x}^*\right)$ by the above Lemma,

$$\lim_{k\to\infty} f(\boldsymbol{x}^{(m_k)}) = f(\lim_{k\to\infty} \boldsymbol{x}^{(m_k)}) = f(\hat{\boldsymbol{x}}) \le f(\boldsymbol{x}^*).$$

$\because$ both $\left\{f(\boldsymbol{x}^{(m_k)})\right\}$ and $\left\{q(\gamma_{m_k}, \boldsymbol{x}^{(m_k)})\right\}$ converge,

Penalty Methods

$\therefore \ \left\{\gamma_{m_k} \boldsymbol{P}\left(\boldsymbol{x}^{(m_k)}\right)\right\} = \left\{q\left(\gamma_{m_k}, \boldsymbol{x}^{(m_k)}\right) - f\left(\boldsymbol{x}^{(m_k)}\right)\right\}$ also converges, with

$$\lim_{k\to\infty} \gamma_{m_k} \boldsymbol{P}\left(\boldsymbol{x}^{(m_k)}\right) = q^* - f(\hat{\boldsymbol{x}}).$$

By the above Lemma, $\left\{\boldsymbol{P}\left(\boldsymbol{x}^{(k)}\right)\right\}$ is nonincreasing and bounded from below by 0.
$\therefore \ \left\{\boldsymbol{P}\left(\boldsymbol{x}^{(k)}\right)\right\}$ converges, and hence so does $\left\{\boldsymbol{P}\left(\boldsymbol{x}^{(m_k)}\right)\right\}$.
$\because \ \gamma_{m_k} \to \infty, \therefore \ \lim_{k\to\infty} \boldsymbol{P}\left(\boldsymbol{x}^{(m_k)}\right) = 0$. By continuity of $\boldsymbol{P}(x)$, we have

$$0 = \lim_{k\to\infty} \boldsymbol{P}(\boldsymbol{x}^{(m_k)}) = \boldsymbol{P}(\lim_{k\to\infty} \boldsymbol{x}^{(m_k)}) = \boldsymbol{P}(\hat{\boldsymbol{x}}),$$

$\therefore \ \hat{\boldsymbol{x}}$ is a feasible point.
$\because \ f\left(\boldsymbol{x}^*\right) \geq f(\hat{\boldsymbol{x}})$ from above,
$\therefore \ \hat{\boldsymbol{x}}$ must be a solution to the constrained optimization problem.
Conclusion: If $\gamma \to \infty$, the limit of any convergent subsequence is a minimizer to the original constrained optimization problem.

Penalty Methods

★ (exact penalty functions): To find a minimizer to the original constrained optimization by a single (not a sequence) unconstrained optimization approximating the original problem. It requires an exact solution to the original constrained optimization by solving the unconstrained problem with a finite $\gamma > 0$.

Drawback of exact penalty functions: nondifferentiable.

$\min f(x) = 5 - 3x$, s.t. $0 \le x \le 1$

The minimizer is $x^* = 1$.

Suppose that the penalty function P is differentiable at $x^* = 1$.

$\because$ $P(x) = 0$ for all $x \in [0, 1]$ (why?), $P'(x^*) = 0$.

If $g := f + \gamma P$, then

$$g'(x^*) = f'(x^*) + \gamma P'(x^*) \neq 0, \ \forall \gamma > 0.$$

$\therefore$ $x^* = 1$ does not satisfy the optimality condition to be a local minimizer of g.

$\therefore$ $\boldsymbol{P}$ is not an exact penalty function.

Penalty Methods

Theorem ($\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \Omega\}$ with $\Omega \subset \mathbb{R}^n$ as a convex set)

Suppose that the minimizer $\boldsymbol{x}^*$ lies on bd(Ω) and there exists a feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}^*$ such that $\boldsymbol{d}^\top \nabla f(\boldsymbol{x}^*) > 0$. If $\boldsymbol{P}$ is an exact penalty function, then $\boldsymbol{P}$ is nondifferentiable at $\boldsymbol{x}^*$.

proof. (contradiction) Suppose that $\boldsymbol{P}$ is differentiable at $\boldsymbol{x}^*$.

$\because$ $P(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Omega$,

$\therefore \nabla P(\boldsymbol{x}^*) = 0$.

Then, $\boldsymbol{d}^\top \nabla P(\boldsymbol{x}^*) = 0$ for all $\boldsymbol{d}$.

If $\boldsymbol{g} := f + \gamma \boldsymbol{P}$, then $\boldsymbol{d}^\top \nabla \boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{d}^\top \nabla f(\boldsymbol{x}^*) > 0$ for all $\gamma > 0$, which implies that $\nabla \boldsymbol{g}(\boldsymbol{x}^*) \neq \boldsymbol{0}$.

$\therefore$ $\boldsymbol{x}^*$ is not a local minimizer of $\boldsymbol{g}$,

$\therefore$ $\boldsymbol{P}$ is not an exact penalty function.

Augmented Lagrangian Methods

motivation of augmented Lagrangian method (ALM)

By combining Lagrange function with penalty function to avoid large penalty parameters, i.e., use a constant penalty parameter.

Given $\boldsymbol{h}: \mathbb{R}^n \to \mathbb{R}^p$, $\boldsymbol{g}: \mathbb{R}^n \to \mathbb{R}^m$.

constrained optimization: $\min\ f(\boldsymbol{x})$, s.t. $\boldsymbol{h}(\boldsymbol{x}) = 0$, $\boldsymbol{g}(\boldsymbol{x}) \leq 0$.

Augmented Lagrangian function $L: \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m_+ \to \mathbb{R}$,

$$
\begin{aligned}
L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) =& f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{\mu}^\top \boldsymbol{g}(\boldsymbol{x}) + \gamma(\|\boldsymbol{h}(\boldsymbol{x})\|^2 + \|\boldsymbol{g}^+(\boldsymbol{x})\|^2) \\
=& f(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i h_i(\boldsymbol{x}) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}) \\
&+ \gamma \left(\sum_{i=1}^{p} |h_i(\boldsymbol{x})|^2 + \sum_{i=1}^{m} |g_i^+(\boldsymbol{x})|^2 \right).
\end{aligned}
$$

Augmented Lagrangian Methods

iterative scheme of ALM

$$\begin{cases} \boldsymbol{x}^{k+1} = \arg\min L\left(\boldsymbol{x}, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k\right), \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \gamma \sum_{i=1}^{p} h_i\left(\boldsymbol{x}^{k+1}\right), \\ \boldsymbol{\mu}^{k+1} = \left(\boldsymbol{\mu}^k + \gamma \sum_{i=1}^{m} \left|g_i^+(\boldsymbol{x})\right|^2\right)^+. \end{cases}$$