MS-E2122 - Nonlinear Optimization
Lecture 9

Fabricio Oliveira

Systems Analysis Laboratory
Department of Mathematics and Systems Analysis

Aalto University
School of Science

October 13, 2021

Outline of this lecture

Constrained methods: Penalty methods

- Penalty functions
- Exterior penalty function methods
- Augmented Lagrangian method of multipliers
- Alternating direction method of multipliers

Penalty functions

We want to penalise constraint violations, turning the problem unconstrained.

Let $P = \text{min. } \{f(x) : g(x) \leq 0, h(x) = 0, x \in X\}$. Then a penalised version of P is:

$$P_\mu = \text{min. } \{f(x) + \mu\alpha(x) : x \in X\},$$

where $\mu > 0$ is a penalty term and $\alpha(x) : \mathbb{R}^n \mapsto \mathbb{R}$ is a penalty function of the form

$$\alpha(x) = \sum_{i=1}^{m} \phi(g_i(x)) + \sum_{i=1}^{l} \psi(h_i(x))$$

and ϕ and ψ are continuous and satisfy:

$$\phi(y) = 0 \text{ if } y \leq 0 \text{ and } \phi(y) > 0 \text{ if } y > 0$$
$$\psi(y) = 0 \text{ if } y = 0 \text{ and } \psi(y) > 0 \text{ if } y \neq 0.$$

Suitable penalty functions

Typical options are $\phi(y) = ([y]^+)^p$ with $p \in \mathbb{Z}_+$ and $\psi(y) = |y|^p$.

Example: (P) : min. $\left\{x_1^2 + x_2^2 : x_1 + x_2 = 1, x \in \mathbb{R}^2\right\}$. Notice that the optimal solution is $(1/2, 1/2)$ with objective $1/2$.

Given a large enough $\mu > 0$, the (penalised) auxiliary problem is:

$$(P_\mu) : \text{min. } \left\{f_\mu(x) = x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1)^2 : x \in \mathbb{R}^2\right\}$$

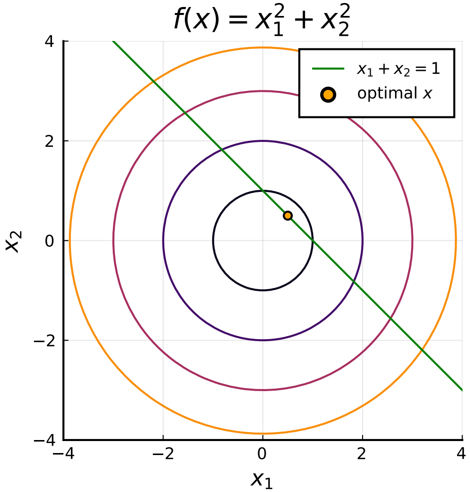
Since f_μ is convex and differentiable, necessary and sufficient optimality conditions $\nabla f_\mu(x) = 0$ imply:

$$\begin{aligned} x_1 + 2\mu(x_1 + x_2 - 1) = 0 \\ x_2 + 2\mu(x_1 + x_2 - 1) = 0, \end{aligned}$$

which gives $x_1 = x_2 = \frac{\mu}{2\mu+1}$.

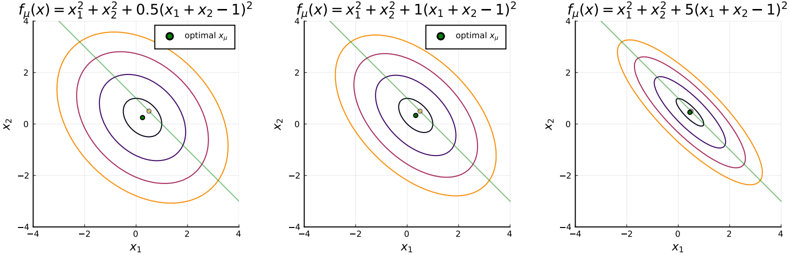
Suitable penalty functions

$$(P): \text{min. } \{x_1^2 + x_2^2 : x_1 + x_2 = 1, x \in \mathbb{R}^2\}$$



Suitable penalty functions

Solving (P_μ) : min. $\left\{x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1)^2 : x \in \mathbb{R}^2\right\}$ with $\mu = 0.5, 1,$ and 5 (from left to right).

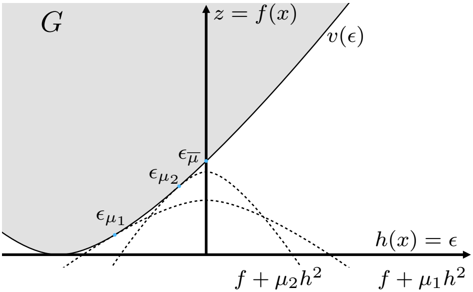


The line represents the original constraint $x_1 + x_2 = 1$ and the orange dot is the optimal $(1/2, 1/2)$ to P.

As μ increases, the optimal of P_μ converges to the optimal of P.

Geometric interpretation

Let $G : \mathbb{R}^2 \to \mathbb{R}^2$ be a mapping $\left\{[h(x), f(x)] : x \in \mathbb{R}^2\right\}$, and let $v(\epsilon) = \text{min. } \left\{x_1^2 + x_2^2 : x_1 + x_2 - 1 = \epsilon,\ x \in \mathbb{R}^2\right\}$. The optimal solution is $x_1 = x_2 = \frac{1+\epsilon}{2}$ with $v(\epsilon) = \frac{(1+\epsilon)^2}{2}$.



Geometric representation of penalised problems in the mapping $G = [h(x), f(x)]$

Minimising $f(x) + \mu(h(x)^2)$ consists of moving the curve downwards until a single contact point ϵ_μ remains.

As $\mu \to \infty$, $f + \mu h$ becomes "sharper" ($\mu_2 > \mu_1$), and ϵ_μ converges to the optimum $\epsilon_{\overline{\mu}}$.

Geometric interpretation

The shape of the penalised problem curve is due to the following:

$$
\begin{aligned}
&\min_{x} \; \left\{ f(x) + \mu \sum_{i=1}^{l} (h_i(x))^2 \right\} \\
=&\min_{x,\epsilon} \; \left\{ f(x) + \mu||\epsilon||^2 : h_i(x) = \epsilon, i = 1, \dots, l \right\} \\
=&\min_{\epsilon} \; \left\{ \mu||\epsilon||^2 + \min_{x} \; \{ f(x) : h_i(x) = \epsilon, i = 1, \dots, l \} \right\} \\
=&\min_{\epsilon} \; \left\{ \mu||\epsilon||^2 + v(\epsilon) \right\}.
\end{aligned}
$$

Consider $l = 1$, and let $x_\mu = \arg\min_\epsilon \left\{ \mu||\epsilon||^2 + v(\epsilon) \right\}$ with $h(x_\mu) = \epsilon_\mu$.

1. $f(x_\mu) + \mu(h(x_\mu))^2 = \mu\epsilon_\mu^2 + v(\epsilon_\mu) \Rightarrow f(x_\mu) = v(\epsilon_\mu)$
2. $v'(\epsilon_\mu) = \frac{\partial}{\partial\epsilon}(f(x_\mu) + \mu(h(x_\mu))^2 - \mu\epsilon_\mu^2) = -2\mu\epsilon_\mu$

Therefore, $(h(x_\mu), f(x_\mu)) = (\epsilon_\mu, v(\epsilon_\mu))$. Letting $f(x_\mu) + \mu h(x_\mu)^2 = k_\mu$, we see the parabolic function $f = k_\mu - \mu\epsilon^2$ matching $v(\epsilon_\mu)$ for $\epsilon = \epsilon_\mu$.

Penalty-based methods

Consider the problem:

$$(P) : \text{min. } \{f(x) : g_i(x) \leq 0, \ i = 1, \dots, m, \\ h_i(x) = 0, \ i = 1, \dots, l, \ \ x \in X\}.$$

We seek to solve P by solving $\sup_\mu \{\theta(\mu)\}$ for $\mu > 0$, where

$$\theta(\mu) = \inf \{f(x) + \mu\alpha(x) : x \in X\}$$

and $\alpha(x)$ is a penalty function. We need a result guaranteeing that

$$\inf \{f(x) : g(x) \leq 0, h(x) = 0, x \in X\} = \sup_{\mu \geq 0} \theta(\mu) = \lim_{\mu \to \infty} \theta(\mu).$$

Remark: in practice, we will calculate $\theta(\mu_k)$ repeatedly increasing μ_k to approximate $\mu \to \infty$.

Penalty-based methods

Theorem 1 (Convergence of penalty-based methods)

Consider the (primal) problem

$$\begin{aligned}(P) : \textit{min.} \ \{f(x) : g_i(x) \le 0, \ i = 1, \dots, m, \\ h_i(x) = 0, \ i = 1, \dots, l, \ x \in X\},\end{aligned}$$

with continuous f, g_i for $i = 1, \dots, m$, and h_i for $i = 1, \dots, l$, and $X \subset \mathbb{R}^n$ a compact set. Suppose that, for each μ, there exists $x_\mu = \arg\min \{f(x) + \mu\alpha(x) : x \in X\}$, where α is a suitable penalty function and $\{x_\mu\}$ is contained within X. Then

$$\inf \{f(x) : g(x) \le 0, h(x) = 0, x \in X\} = \sup_{\mu \ge 0} \{\theta(\mu)\} = \lim_{\mu \to \infty} \theta(\mu),$$

where $\theta(\mu) = \inf \{f(x) + \mu\alpha(x) : x \in X\} = f(x_\mu) + \mu\alpha(x_\mu)$. Also, the limit of any convergent subsequence of $\{x_\mu\}$ is optimal to the original problem and $\mu\alpha(x_\mu) \to 0$ as $\mu \to \infty$.

Penalty-based methods

One important corollary from Theorem 1 is the following.

Corollary 2

If $\alpha(x_\mu) = 0$ for some μ, then x_μ is optimal for P.

Proof.

If $\alpha(x_\mu) = 0$, then x_μ is feasible. Moreover, x_μ is optimal, since

$$\begin{aligned}\theta(\mu) &= f(x_\mu) + \mu\alpha(x_\mu) \\ &= f(x_\mu) \leq \inf\left\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\right\}. \quad \square\end{aligned}$$

Remarks:

- Notice that X needs to be compact (e.g. bounded variables), or optimal primal and penalty function values may not match.
- Making μ arbitrarily large, x_μ can be made arbitrarily close to the feasible region and $f(x_\mu) + \mu\alpha(x_\mu)$ can be made arbitrary close to the optimal value.

Computational issues with penalty methods

One might wonder why not start with a very large μ to reduce the number of iterations. The answer for this is ill-conditioning.

Some of the eigenvalues of the Hessians of penalty functions are proportional to the penalty terms, thus affecting conditioning.

Recall that conditioning is measured by $\kappa = \frac{\max_{i=1,\dots,n} \lambda_i}{\min_{i=1,\dots,n} \lambda_i}$, where $\{\lambda_i\}_{i=1,\dots,n}$ are the eigenvalues of the Hessian.

Example: $f_\mu(x) = x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1)^2$.

The Hessian of $f_\mu(x)$ at x is

$$\nabla^2 f_\mu(x) = \begin{bmatrix} 2(1+\mu) & 2\mu \\ 2\mu & 2(1+\mu) \end{bmatrix}.$$

Solving $\mathbf{det}(\nabla^2 f_\mu(x) - \lambda I) = 0$, we get $\lambda_1 = 2$, $\lambda_2 = 2(1+2\mu)$, with eigenvectors $(1,-1)$ and $(1,1)$, which gives $\kappa = (1+2\mu)$.

Augmented Lagrangian methods

We will develop a penalty method that works with finite penalties by shifting the curve implied by the penalty term.

For simplicity, consider the (primal) problem P as

$$(P) : \text{min. } \{f(x) : h_i(x) = 0,\ i = 1, \dots, l\}.$$

The shifted penalty defines an augmented Lagrangian of P:

$$\begin{aligned} f_\mu(x) &= f(x) + \mu \sum_{i=1}^{l} (h_i(x) - \theta_i)^2 \\ &= f(x) + \mu \sum_{i=1}^{l} h_i(x)^2 - \sum_{i=1}^{l} 2\mu\theta_i h_i(x) + \mu \sum_{i=1}^{l} \theta_i^2 \\ &= f(x) + \sum_{i=1}^{l} v_i h_i(x) + \mu \sum_{i=1}^{l} h_i(x)^2, \end{aligned}$$

with $v_i = -2\mu\theta_i$. The last term is a constant and can be dropped.

Augmented Lagrangian methods

The name refers to the fact that

$$f_\mu(x) = f(x) + \sum_{i=1}^{l} v_i h_i(x) + \mu \sum_{i=1}^{l} h_i(x)^2$$

is equivalent to the Lagrangian function of problem P, augmented with the penalty term.

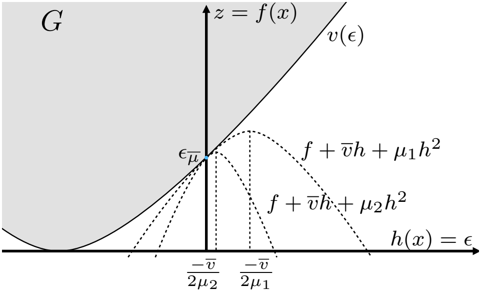
Moreover, assuming that $(\overline{x}, \overline{v})$ is a KKT solution to P, we have

$$\nabla_x f_\mu(x) = \nabla f(x) + \sum_{i=1}^{l} \overline{v}_i \nabla h_i(x) + 2\mu \sum_{i=1}^{l} h_i(x) \nabla h_i(x) = 0,$$

which implies that the optimal solution $\overline{x}$ can be recovered using a finite penalty, unlike with the previous penalty-based methods.

Augmented Lagrangian - geometric interpretation

Let $v(\epsilon) = \text{min. } \{f(x) : h(x) = \epsilon\}$ be the perturbation function.
We will minimise $f(x) + \overline{v}h(x) + \mu h(x)^2$ for a given $\mu > 0$.



Geometric representation of augmented Lagrangians in the mapping $G = [h(x), f(x)]$

The minimum is attained for $f + \overline{v}h + \mu h^2 = k$, or equivalently $f = -\mu\left[h + (\overline{v}/2\mu)\right]^2 + \left[k + (\overline{v}^2/4\mu)\right]$, with k touching $v(\epsilon)$.
Notice that f is a parabola shifted by $h = -\overline{v}/2\mu$.

(Augmented Lagrangian) method of multipliers (MM)

Define the augmented Lagrangian function

$$L_\mu(x, v) = f(x) + \sum_{i=1}^{l} v_i h_i(x) + \mu \sum_{i=1}^{l} h_i(x)^2$$

The strategy is to search for KKT points (or primal-dual pairs) $(\overline{x}, \overline{v})$ by iteratively operating in both primal (x) and dual (v) spaces.

1. **Primal space:** optimise $L_\mu(x, v^k)$ using an unconstrained optimisation method
2. **Dual space:** perform a dual variable update step retaining $\nabla_x L_\mu(x^{k+1}, v^k) = \nabla_x L_\mu(x^{k+1}, v^{k+1}) = 0$

(Augmented Lagrangian) method of multipliers (MM)

The dual variable update step is $\overline{v}^{k+1} = \overline{v}^k + 2\mu h(\overline{x}^{k+1})$, which is justified as follows:

1. $h(\overline{x}^k)$ is a subgradient of $L_\mu(x, v)$ at $\overline{x}^k$ for any v.
2. The step size is devised such that the optimality condition of the Lagrangian is retained, i.e., $\nabla_x L(\overline{x}^k, \overline{v}^{k+1}) = 0$.

Part 2. refers to the following:

$$\begin{aligned}
\nabla_x L(\overline{x}^k, \overline{v}^{k+1}) &= \nabla f(\overline{x}^k) + \sum_{i=1}^{l} \overline{v}_i^{k+1} \nabla h_i(\overline{x}^k) = 0 \\
&= \nabla f(\overline{x}^k) + \sum_{i=1}^{l} (\overline{v}_i^k + 2\mu h_i(\overline{x}^k)) \nabla h_i(\overline{x}^k) = 0 \\
&= \nabla f(\overline{x}^k) + \sum_{i=1}^{l} \overline{v}_i^k \nabla h_i(\overline{x}^k) + \sum_{i=1}^{l} 2\mu h_i(\overline{x}^k) \nabla h_i(\overline{x}^k) = 0.
\end{aligned}$$

(Augmented Lagrangian) method of multipliers (ALMM)

Algorithm (Augmented Lagrangian) method of multipliers

1: **initialise.** tolerance $\epsilon > 0$, initial dual solution v^0, iteration count $k = 0$
2: **while** $|h(\overline{x}^k)| > \epsilon$ **do**
3: $\quad \overline{x}^{k+1} = \arg\min L_\mu(x, \overline{v}^k)$
4: $\quad \overline{v}^{k+1} = \overline{v}^k + 2\mu h(\overline{x}^{k+1})$
5: $\quad k = k + 1$
6: **end while**
7: **return** x^k.

Remarks:

- μ can be individualised for each constraint: $\sum_{i=1}^{l} \mu_i h_i(x)^2$.
- Increasing μ_i for most violated constraints $\max_{i=1,\dots,l} h_i(x)$ is often used. Provides convergence guarantees as $\mu \to \infty$.
- Due to the gradient-like step in the dual space, we can expect linear convergence from the ALMM.

Alternating direction method of multipliers - ADMM

ADMM is a distributed version of the method of multipliers.

Best suited for large problems with decomposable structure, so computations can be performed in a distributed manner.

Consider a problem P of the form:

$$\begin{aligned} (P) : \text{min. } & f(x) + g(y) \\ \text{subject to: } & Ax + By = c \end{aligned}$$

Problems of this form appear in several important applications in stochastic programming and regularisation for example.

We aim to solve problems of this form in a distributed manner in terms of x and y.

Alternating direction method of multipliers - ADMM

We start by formulating the augmented Lagrangian function

$$\phi(x, y, v) = f(x) + g(y) + v^\top(c - Ax - By) + \mu(c - Ax - By)^2$$

The penalty term $\mu(c - Ax - By)^2$ prevents separation, which is recovered by optimising x and y in a coordinate descent fashion.

Algorithm ADMM

1: **initialise.** tolerance $\epsilon > 0$, initial dual and primal solutions v^0 and y^0, $k = 0$
2: **while** $|c - A\overline{x}^k - B\overline{y}^k|$ and $||y^{k+1} - y^k|| > \epsilon$ **do**
3: $\quad\overline{x}^{k+1} = \arg\min \phi_\mu(x, \overline{y}^k, \overline{v}^k)$
4: $\quad\overline{y}^{k+1} = \arg\min \phi_\mu(\overline{x}^{k+1}, y, \overline{v}^k)$
5: $\quad\overline{v}^{k+1} = \overline{v}^k + 2\mu(c - A\overline{x}^{k+1} - B\overline{y}^{k+1})$
6: $\quad k = k + 1$
7: **end while**
8: **return** (x^k, y^k).

Remark: the stopping criteria in Line 2 consider primal and dual (indirectly) residuals that can take different values.

Alternating direction method of multipliers - ADMM

Remarks

1. Optimising with respect to (x, y) requires additional steps in Lines 3 and 4. However, this is not needed for convergence.
2. Variants consider more than one (x, y) step. No clear benefit has been observed in practice.
3. For ADMM, no generally good update rule for μ is known.
4. Convergence of ADMM is worse compared to the method of multipliers. The benefit of ADMM comes from the ability to separate x and y.
5. Notice that, if we can further separate x (or y), Lines 3 (or 4) can be calculated in a distributed fashion.