A general constraint optimization problem: $\min_{x \in \mathbb{R}^n} f(x)$ s.t. $c_i(x) = 0, \ i \in \mathcal{E}$
$c_i(x) \geq 0, \ i \in \mathcal{I}$.

$f, c_i : \mathbb{R}^n \to \mathbb{R}^1$ are smooth (C^1 or C^2.)

Definition 12.4 (LICQ).

Given the point x and the active set $\mathcal{A}(x)$ defined in Definition 12.1, we say that the linear independence constraint qualification (LICQ) *holds if the set of active constraint gradients $\{\nabla c_i(x), i \in \mathcal{A}(x)\}$ is linearly independent.*

Theorem 12.1 (First-Order Necessary Conditions).

Suppose that x^* is a local solution of (12.1), that the functions f and c_i in (12.1) are continuously differentiable, and that the LICQ holds at x^*. Then there is a Lagrange multiplier vector λ^*, with components λ_i^*, $i \in \mathcal{E} \cup \mathcal{I}$, such that the following conditions are satisfied at (x^*, λ^*)

$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x)$

KKT conditions

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0, \tag{12.34a}$$
$$c_i(x^*) = 0, \quad \text{for all } i \in \mathcal{E}, \tag{12.34b}$$
$$c_i(x^*) \geq 0, \quad \text{for all } i \in \mathcal{I}, \tag{12.34c}$$
$$\lambda_i^* \geq 0, \quad \text{for all } i \in \mathcal{I}, \tag{12.34d}$$
$$\lambda_i^* c_i(x^*) = 0, \quad \text{for all } i \in \mathcal{E} \cup \mathcal{I}. \tag{12.34e}$$

Without inequality constraint, this is the standard result of Lagrange multiplier.

Linear Program (LP) : f is linear and c_i are affine functions

Note : A LP is pretty boring if $\mathcal{I} = \emptyset$, as $\min c^T x$ s.t. $Ax = b$ has only three boring possibilities :

(i) infeasible (when the equations $Ax = b$ are inconsistent), or
(ii) unbounded (when $c \neq 0$ and $Ax = b$ are consistent) or
(iii) bounded in a lame way, namely when $c = 0$.

Quadratic Program (QP) : f is quadratic, c_i are affine functions

Without inequality constraints, a QP is already "not boring", but it is relatively easy as it is about solving a linear system.

$$\min \; q(x) = \frac{1}{2} x^T Q x + x^T c \qquad \text{s.t. } Ax = b$$

First order necessary condition for optimality :

$$\begin{aligned} Qx + c - A^T \lambda &= 0 \\ Ax &= b \end{aligned} \quad \text{or} \quad \begin{bmatrix} Q & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix}.$$

In general, dealing with inequality constraints is challenging.
To state the obvious: there are up to $2^{|\mathcal{I}|}$ possible subsets of $\mathcal{I}$. Any practical algorithm cannot be based on considering all possible subsets of active subsets of $\mathcal{I}$.
(Recall what are done in the simplex and the interior methods for LPs).

There are a number of specific types of convex constraint optimization problems not discussed here, such as SDP, SOCP etc.

QP is important by itself, its structure can be exploited by efficient algorithms, and in the (successful) sequential quadratic programming (SQP) methods for solving general constraint optimization problems, QP subproblems need to be solved.

Let's first see an well-known application of QP.

Application of QP in classification problems

Suppose we have labelled training data:

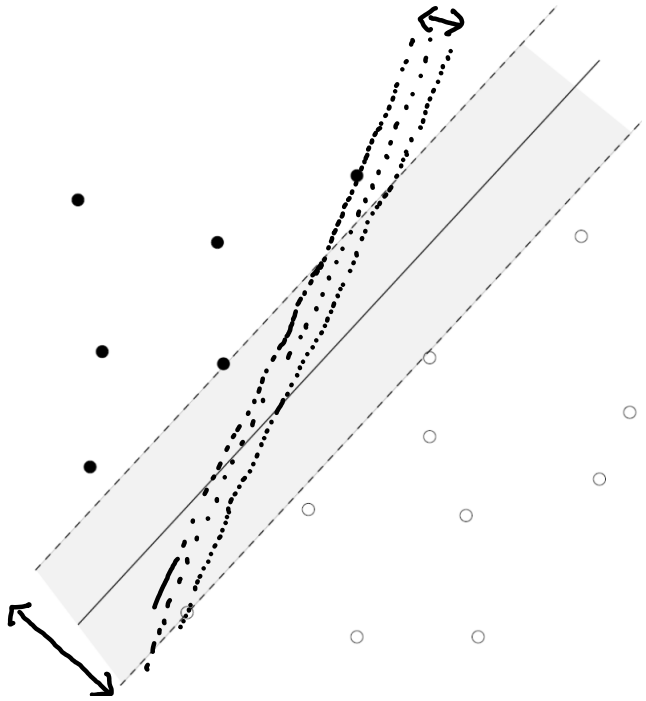
Type A	$x_1, \cdots, x_m,$	$\in \mathbb{R}^n$
Type B	$x_{m+1}, \cdots, x_{m+p}$	

First, assume the two groups of points are linearly separable, i.e.

$$\exists\ w \in \mathbb{R}^n \setminus \{0\},\ \gamma \in \mathbb{R} \text{ s.t.}$$
$$w^T x_i - \gamma > 0 \quad i = 1, \cdots, m$$
$$w^T x_i - \gamma < 0 \quad i = m+1, \cdots, m+p.$$

Problem:

Find the linear classifier with the largest margin.

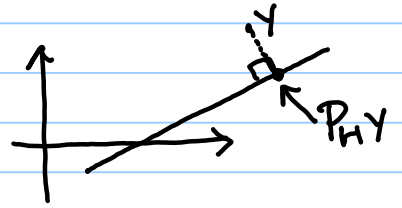


Lemma: Let $H(w,\gamma) = \{x \in \mathbb{R}^n : w^T x = \gamma\}$, $w \in \mathbb{R}^n \setminus \{0\}$, $\gamma \in \mathbb{R}$,

Let $y \in \mathbb{R}^n$.

Then $d_2(y, H(w,\gamma)) = |w^T y - \gamma| / \|w\|_2$.

Proof:



The orthogonal projection of y on $H(w,\gamma)$ is the unique solution of the strictly convex problem:

$$\min \frac{1}{2}\|x-y\|_2^2 \quad \text{s.t.} \quad w^T x = \gamma, \qquad \leftarrow \text{a QP with only (one) equality constraint.}$$

which is the unique solution of the linear system: (the "easy kind")

$$w^T x = \gamma \;, \quad \mathcal{L}(x,\lambda) = \frac{1}{2}(x-y)^T(x-y) - \lambda(w^T x - \gamma) \;, \quad \nabla_x \mathcal{L}(x,\lambda) = x - y - \lambda w = 0 \Rightarrow x = y + \lambda w$$

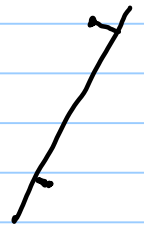
Substitute $x = y + \lambda w$ into the first equation gives $w^T y + \lambda w^T w = 0$

$$\Rightarrow \lambda = -\frac{w^T y - \gamma}{w^T w}$$

$$\text{dist}(y, H(w,\gamma)) = \|\lambda w\| = |w^T y - \gamma| / \|w\|_2 . \qquad \text{Q.E.D.}$$

The problem is then to solve.

$$\max_{w,\gamma} \min_i \frac{|w^T x_i - \gamma|}{\|w\|_2} \quad \text{s.t.} \quad \begin{matrix} w^T x_i - \gamma > 0 & i \le m \\ w^T x_i - \gamma < 0 & i > m \end{matrix} \qquad \text{(I)}$$



This does not look like a tractable problem, and is certainly not a QP.

Note if (w,γ) is a solution, then $\alpha(w,\gamma)$ is also a solution for any $\alpha>0$. So we may 'normalize' (w,γ) by requiring $\min_i |w^T x_i - \gamma| = 1$, which would then imply

(i) $w^T x_i - \gamma \geq 1$, $i \leq m$, $w^T x_i - \gamma \leq -1$, $i > m$

and

(ii) $\min_i |w^T x_i - \gamma| / \|w\|_2 = 1/\|w\|_2$

So (I) is equivalent to:

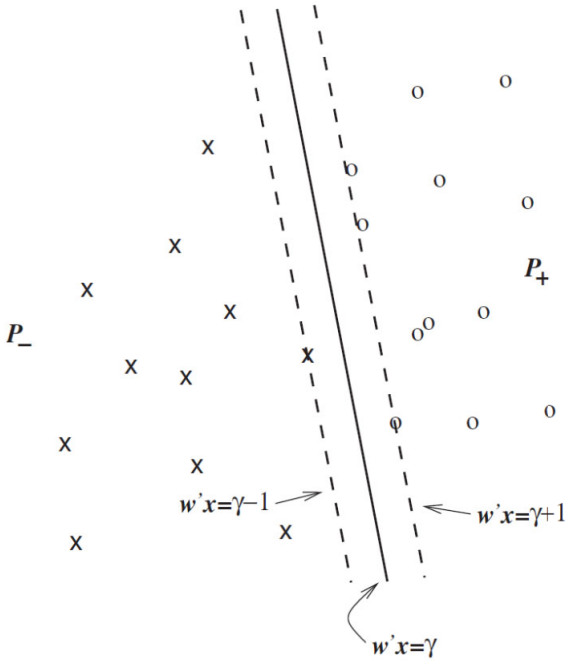
$$\max_{w,\gamma} 1/\|w\|_2 \quad \text{s.t.} \quad \begin{array}{l} \min_i |w^T x_i - \gamma| = 1, \\ w^T x_i - \gamma \geq 1, \quad i \leq m, \\ w^T x_i - \gamma \leq -1, \quad i > m, \end{array} \qquad \text{(II)}$$

which, in turn, is equivalent to

$$\min_{w,\gamma} \tfrac{1}{2} w^T w \quad \text{s.t.} \quad \begin{array}{l} w^T x_i - \gamma \geq 1, \quad i \leq m, \\ w^T x_i - \gamma \leq -1, \quad i > m. \end{array} \qquad \text{(III)}$$

This is a QP (with many linear inequality constraints).

Ex: Explain why (II) and (III) are equivalent.



Nonseparable case:

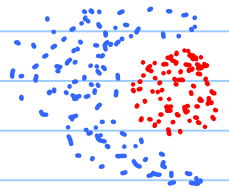
When the data points are not linearly separable (due to, say, noise in the data), the QP above is infeasible. But we may still aim to find a classifier that simultaneously maximize 'margin' and minimize misclassification by solving:

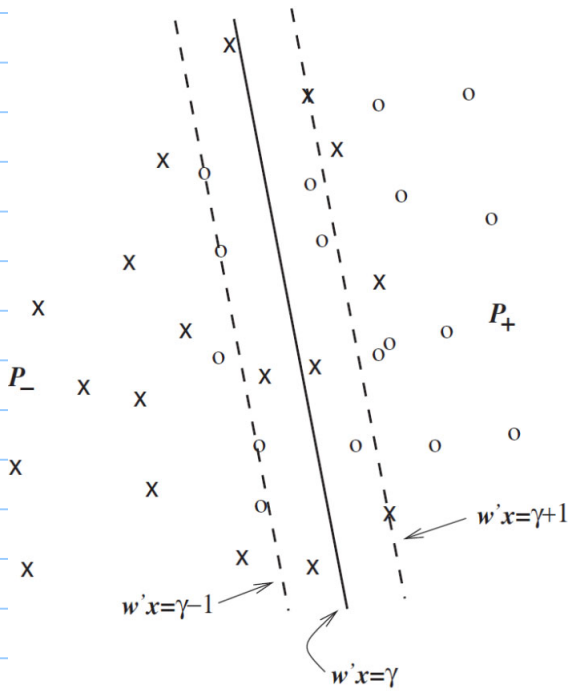
$$\min \tfrac{1}{2} w^T w + \nu \sum \gamma_i \quad \text{s.t.} \quad \begin{aligned} w^T x_i - \gamma + \gamma_i &\geq 1, && i \leq m \\ w^T x_i - \gamma - \gamma_i &\leq -1, && i > m \\ \gamma_i &\geq 0, && i = 1, \dots, m+p \end{aligned}$$

ν: a penalization parameter that needs to be tuned

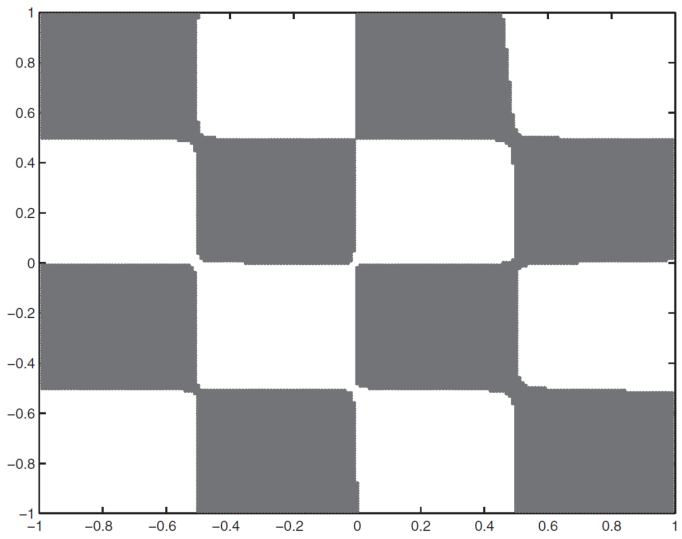
This is still a QP, referred to as a (linear) support vector machine (SVM).

What if the linear non-separability is not due to just noise?





Not discussed : the kernel trick for nonlinear SVM.



← separating hyperplane seems useless for data like this, or is it?

Methods for solving general constrained optimization problem

I. Quadratic Penalty Method

For a equality-constrained problem $\min_x f(x)$ s.t. $c_i(x) = 0 \quad i \in \mathcal{E}$, (17.1) its quadratic penalty function is

$$Q(x;\mu) := f(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i(x)^2 \quad , \quad \mu > 0 \text{ is the penalty parameter.} \tag{17.2}$$

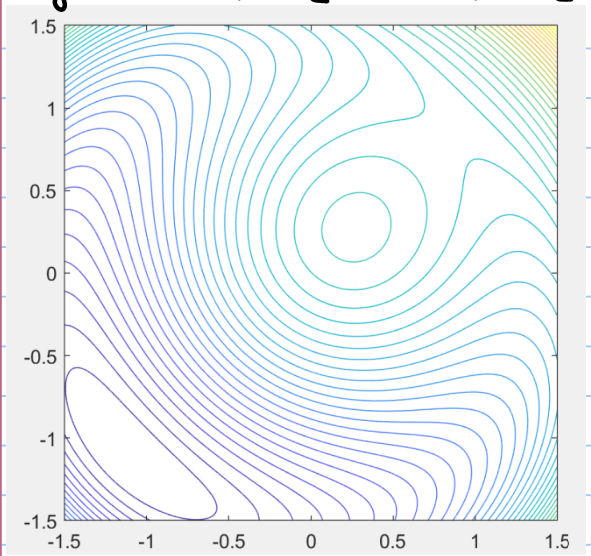
For the general constrained optimization problem $\min f(x)$ s.t. $c_i(x) = 0 \quad i \in \mathcal{E}$
$c_i(x) \geq 0 \quad i \in \mathcal{I}$,
we can define the quadratic penalty function as

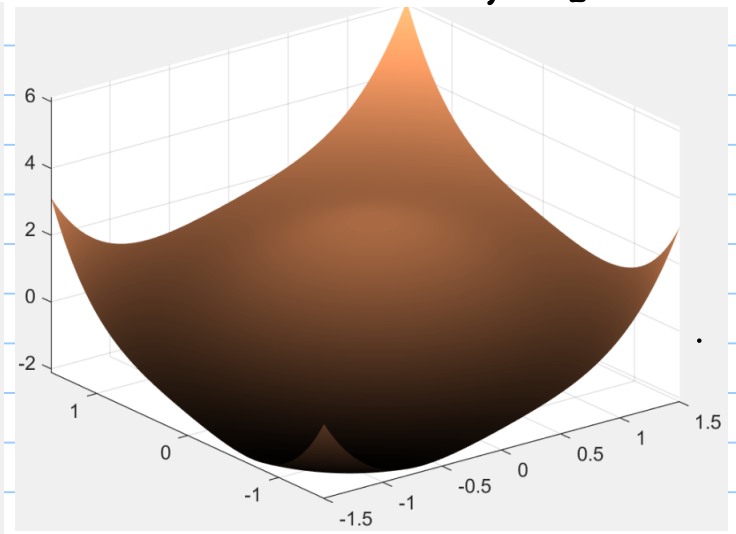
$$Q(x;\mu) := f(x) + \frac{\mu}{2} \left\{ \sum_{i \in \mathcal{E}} c_i(x)^2 + \sum_{i \in \mathcal{I}} ([c_i(x)]^-)^2 \right\},$$

where $[y]^- := \max(-y, 0)$.

We expect that if x_k solves (or approximately solves) $\min_x Q(x;\mu_k)$,
$\mu_k \to \infty$ and x_k (or some subsequence of it) $\to x^*$,
then x^* is a solution of the original problem.

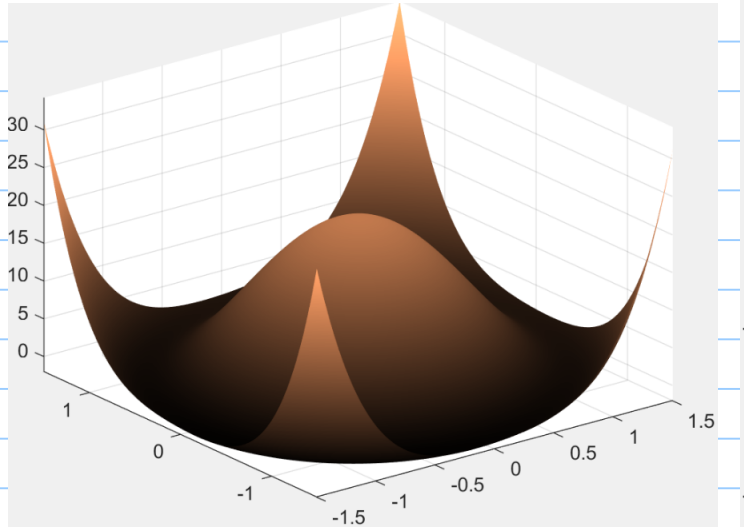
E.g. $\min x_1 + x_2$ s.t. $x_1^2 + x_2^2 - 2 = 0$ Solution: $[-1, -1]^T$

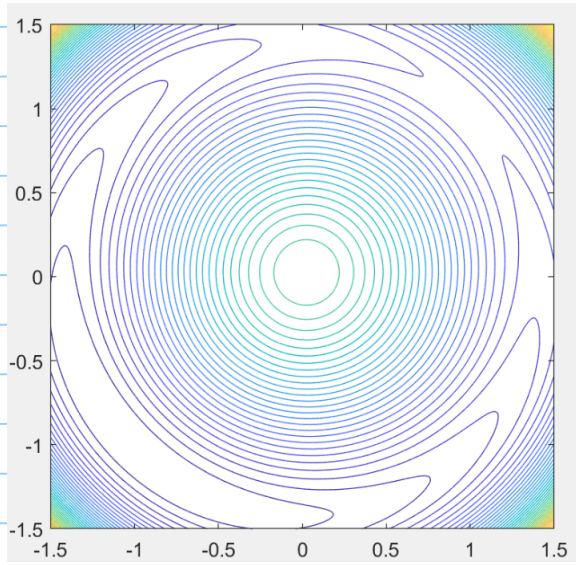




← $\mu = 1$

$\mu = 10$ →





Issues with the method:

- when μ is small, $Q(x;\mu)$ may be unbounded below even when the constrained problem is bounded.

  E.g $\min -5x_1^2+x_2^2$ st $x_1=1$ (solution at $[1,0]^T$.)

  $Q(x;\mu) = -5x_1^2+x_2^2+\mu(x_1-1)^2$ is unbounded below for $\mu<10$.

- The penalty function is non-smooth when $\mathcal{I}\neq\emptyset$

  E.g. If $x_1\geq 0$ is an inequality constraint, $([x_1]^-)^2 = \max(-x_1,0)^2$
  $= \min(0,x_1)^2$
  ↑
  C^1 but not twice differentiable

  (Again, inequality constraints cause difficulties.)

- (In the case with only equality constraints), the condition number of $\nabla^2 Q(x^*;\mu) \uparrow \infty$ as $\mu\uparrow\infty$.

Framework 17.1 (Quadratic Penalty Method).

Given $\mu_0 > 0$, a nonnegative sequence $\{\tau_k\}$ with $\tau_k \to 0$, and a starting point x_0^s;
for $k = 0, 1, 2, \ldots$
    Find an approximate minimizer x_k of $Q(\cdot; \mu_k)$, starting at x_k^s,
        and terminating when $\|\nabla_x Q(x; \mu_k)\| \leq \tau_k$;
    if final convergence test satisfied
        stop with approximate solution x_k;
    end (if)
    Choose new penalty parameter $\mu_{k+1} > \mu_k$;
    Choose new starting point x_{k+1}^s;
end (for)

Thm 17.1 in NSW: Assume the equality constrained problem $\min f(x)$ s.t. $c_i(x)=0$, $i \in \mathcal{E}$ has a global solution.
Assume also that $Q(\cdot; \mu_k)$ has a global minimizer, and $\mu_k \uparrow \infty$.
Then every limit point x^* of the sequence $\{x_k\}$ is a global sol. of the constrained problem.

Proof: Let $\bar{x}$ be a solution of $\min f(x)$ s.t. $c_i(x)=0$, $i\in\mathcal{E}$, so

$$f(\bar{x}) \le f(x) \quad \forall x \text{ with } c_i(x)=0,\ i\in\mathcal{E}.$$

Since x_k minimizes $Q(\cdot;\mu_k)$ for each k, $Q(x_k,\mu_k) \le Q(\bar{x},\mu_k)$, which leads to:

$$f(x_k) + \frac{\mu_k}{2}\sum_{i\in\mathcal{E}} c_i^2(x_k) \le f(\bar{x}) + \frac{\mu_k}{2}\sum_{i\in\mathcal{E}} c_i^2(\bar{x}) = f(\bar{x}),$$

which implies:

$$\sum_{i\in\mathcal{E}} c_i^2(x_k) \le \frac{2}{\mu_k}[f(\bar{x}) - f(x_k)] \to 0,$$

So if (a subsequence of) $x_k \to x^*$, then $\sum_{i\in\mathcal{E}} c_i^2(x^*) = \lim_{k\to\infty} \sum_{i\in\mathcal{E}} c_i^2(x_k) = 0$,

which implies $c_i(x^*)=0 \quad \forall i\in\mathcal{E}$.

moreover, $f(x^*) = \lim_k f(x_k) + \frac{\mu_k}{2}\sum_{i\in\mathcal{E}} c_i^2(x_k) \le f(\bar{x})$, meaning that x^* is a global solution. Q.E.D.

Among other issues, this result requires us to find a global minimizer of each subproblem, which isn't practical in general.

The next result is perhaps more interesting. It allows for approximate solution of

$\min_x Q(x;\mu_k)$, it allows the problem (17.1) to be infeasible, and it illustrates a connection of the square penalty method and the method of Lagrange multipliers. The latter forms the basis of the augmented Lagrangian method also.

Theorem 17.2.

Suppose that the tolerances and penalty parameters in Framework 17.1 satisfy $\tau_k \to 0$ *and* $\mu_k \uparrow \infty$. *Then if a limit point* x^* *of the sequence* $\{x_k\}$ *is infeasible, it is a stationary point of the function* $\|c(x)\|^2$. *On the other hand, if a limit point* x^* *is feasible and the constraint gradients* $\nabla c_i(x^*)$ *are linearly independent, then* x^* *is a KKT point for the problem (17.1). For such points, we have for any infinite subsequence* $\mathcal{K}$ *such that* $\lim_{k\in\mathcal{K}} x_k = x^*$ *that*

$$\lim_{k\in\mathcal{K}} -\mu_k c_i(x_k) = \lambda_i^*, \qquad \text{for all } i \in \mathcal{E}, \tag{17.10}$$

where λ^* *is the multiplier vector that satisfies the KKT conditions (12.34) for the equality-constrained problem (17.1).*

See N&W for the complete proof.

The proof begins with differentiating $Q(x;\mu_k)$: w.r.t. x

$$\nabla_x Q(x_k;\mu_k) = \nabla f(x_k) + \sum_{i\in\mathcal{E}} \underbrace{\mu_k c_i(x_k)}_{\approx -\lambda_i^* \ (k \text{ large})} \nabla c_i(x_k) \approx 0$$

Ill conditioning and reformulations

$$\nabla^2 Q(x;\mu) = \nabla^2 f(x) + \mu \sum_{i\in\mathcal{E}} \{c_i(x)\nabla^2 c_i(x) + \nabla c_i(x)\nabla c_i(x)^T\}$$

$$\nabla^2 Q(x_k;\mu_k) = \nabla^2 f(x_k) + \sum_{i\in\mathcal{E}} \underbrace{\mu_k c_i(x_k)}_{\approx -\lambda_i^*} \nabla^2 c_i(x_k) + \mu_k \sum_{i\in\mathcal{E}} \nabla c_i(x_k)\nabla c_i(x_k)^T$$

$$\approx \nabla_{xx}^2 \mathcal{L}(x_k, \lambda^*) + \mu_k \sum_{i\in\mathcal{E}} \nabla c_i(x_k)\nabla c_i(x_k)^T$$

So when k is large and $x \approx x_k \approx x^*$,

$$\nabla^2 Q(x;\mu_k) \approx \underbrace{\nabla_{xx}^2 \mathcal{L}(x^*;\lambda^*)}_{\text{independent of } \mu_k} + \mu_k \underbrace{\sum_{i\in\mathcal{E}} \nabla c_i(x^*)\nabla c_i(x^*)^T}_{\text{rank } |\mathcal{E}|, \text{ whose non-zero eigenvalues are of order } \mu_k}$$

When $|\mathcal{E}| < n$ (which better be the case, otherwise the optimization problem is over-constrained), perturbation results of eigenvalues tell us that the overall matrix has some of its eigenvalues approaching values independent of μ_k, while others are of order μ_k, meaning that

$$\frac{\lambda_{\max}(\nabla^2 Q(x;\mu_k)}{\lambda_{\min}(\nabla^2 Q(x;\mu_k)} \to \infty \quad \text{as} \quad x \to x_k \text{ and } k \uparrow \infty.$$

This ill-conditioning makes many unconstrained minimization algorithms such as quasi-Newton and conjugate gradient perform poorly. Newton's method, on the other hand, is not sensitive to ill-conditioning of the Hessian, but it, too, may encounter difficulties for large μ_k for two other reasons:

1. The ill-conditioning might be expected to cause numerical problems when we solve the linear equations $-\nabla^2 Q(x;\mu_k)\,p = \nabla Q(x;\mu_k)$ for the Newton step.

2. The quadratic approximation is a reasonable approximation of the true function only in a small neighborhood of x.

Remedy for 2: use warm start, i.e. choose the starting point for step $k+1$ to be the (approximate) solution of step k, and choosing μ_{k+1} to be only modestly larger than μ_k.

Remedy for 1: use the following reformulation of $\nabla^2 Q(x;\mu_k)\,p = -\nabla Q(x;\mu_k)$:

$$\nabla^2 Q(x;\mu) = \nabla^2 f(x) + \mu \sum_{i\in\mathcal{E}} \{c_i(x)\,\nabla^2 c_i(x) + \nabla c_i(x)\,\nabla c_i(x)^T\}$$

$$= \nabla^2 f(x) + \mu \sum_{i\in\mathcal{E}} c_i(x)\,\nabla^2 c_i(x) + \mu\, A(x)^T A(x), \quad A(x)^T = \left[\nabla c_1(x) \cdots \nabla c_{|\mathcal{E}|}(x)\right]$$

Write $\mu_k A(x) p = \zeta$, then the Newton step p is the solution of

$$\begin{bmatrix} \nabla^2 f(x) + \mu_k \sum_{i \in \mathcal{E}} c_i(x) \nabla^2 c_i(x) & A(x)^T \\ A(x) & -\frac{1}{\mu_k} I \end{bmatrix} \begin{bmatrix} p \\ \zeta \end{bmatrix} = \begin{bmatrix} -\nabla Q(x; \mu_k) \\ 0 \end{bmatrix}.$$

↑ larger than $\nabla^2 Q(x; \mu_k)$ ($|\mathcal{E}|$ more rows and columns), but well-conditioned.

Despite these ideas, the quadratic penalty method is usually not used directly in practice. However, it inspires two very interesting methods:

II. Nonsmooth Exact Penalty Methods

III. Augmented Lagrangian Methods.

Both avoid the need to increase the penalty parameter to arbitrarily large values.

II Nonsmooth Exact Penalty Methods

Let's see a little "miracle":

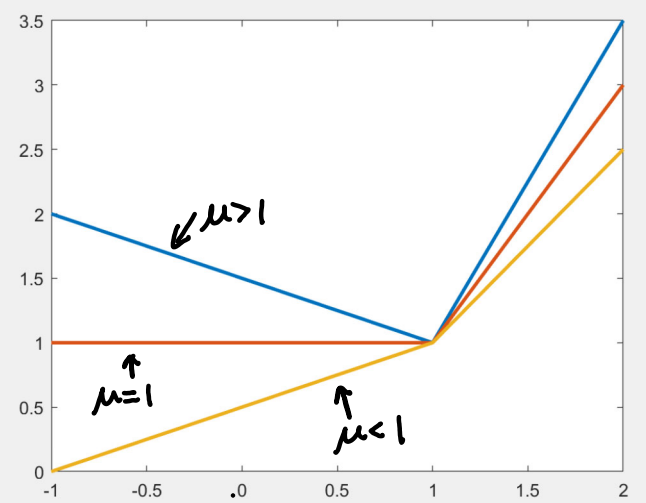
Eg. $\min x$ s.t. $x=1$ (*)

$Q(x;\mu) = x + \frac{\mu}{2}(x-1)^2$. $Q' = 1+\mu(x-1)$, which vanishes at $x_\mu = 1-\frac{1}{\mu} \to 1$ as $\mu \to \infty$.

solution of (*)

Now, consider $\phi_1(x;\mu) = x + \mu|x-1|$, which is not C^1.

But notice a spectacular property: $\operatorname{argmin}_x \phi_1(x;\mu) \equiv 1 \quad \forall \mu > 1$.



There is no need to make μ arbitrarily large, and no need to approximate!

If the problem is $\min x$ s.t. $x \geq 1$, (**)

set

$$\phi_1(x;\mu) = x + \mu[x-1]^- = \begin{cases} x, & x \geq 1 \\ (1-\mu)x + \mu, & x < 1 \end{cases}$$

Again, $\operatorname{argmin}_x \phi_1(x;\mu) \equiv 1$, $\forall \mu > 1$.

↑ Sol. of (**)

$[y]^- = \max(0, -y)$.

It is not an accident. For a general constrained problem

$$\min_x f(x) \quad \text{st} \quad \begin{array}{ll} c_i(x) = 0 & i \in \mathcal{E} \\ c_i(x) \geq 0 & i \in \mathcal{I} \end{array}, \qquad (17.6)$$

its ℓ_1 penalty function is $\phi_1(x;\mu) := f(x) + \mu \sum_{i \in \mathcal{E}} |c_i(x)| + \mu \sum_{i \in \mathcal{I}} [c_i(x)]^-$

ϕ_1 is not C^1, and it is the non-smoothness that gives it the following exactness property:

Theorem 17.3. [Han & Magarsanian 1979]

Suppose that x^* is a strict local solution of the nonlinear programming problem (17.6) at which the first-order necessary conditions of Theorem 12.1 are satisfied, with Lagrange multipliers λ_i^*, $i \in \mathcal{E} \cup \mathcal{I}$. Then x^* is a local minimizer of $\phi_1(x;\mu)$ for all $\mu > \mu^*$, where

$$\mu^* = \|\lambda^*\|_\infty = \max_{i \in \mathcal{E} \cup \mathcal{I}} |\lambda_i^*|. \qquad (17.23)$$

If, in addition, the second-order sufficient conditions of Theorem 12.6 hold and $\mu > \mu^*$, then x^* is a strict local minimizer of $\phi_1(x;\mu)$.

One needs to venture into techniques of non-smooth optimization in order to solve $\min_x \phi(x;\mu)$. Here we give a taste of this subject.

Even though ϕ_1 is not differentiable, it has a directional derivative $D(\phi_1(x;\mu);p)$ along any direction p.

E.g. $$D(\|x\|_1;p) = \lim_{\varepsilon\to 0} \frac{\|x+\varepsilon p\|_1 - \|x\|_1}{\varepsilon}$$

$$= \lim_{\varepsilon\to 0} \frac{\sum_{i=1}^n |x_i+\varepsilon p_i| - \sum_{i=1}^n |x_i|}{\varepsilon}.$$

$$= \sum_{i=1}^n \underbrace{\lim_{\varepsilon\to 0} \frac{|x_i+\varepsilon p_i| - |x_i|}{\varepsilon}}_{\begin{array}{ll} -p_i & \text{if } x_i<0 \\ p_i & \text{if } x_i>0 \\ |p_i| & \text{if } x_i=0 \end{array}} = \sum_{i|x_i<0} -p_i + \sum_{i|x_i>0} p_i + \sum_{i|x_i=0} |p_i|.$$

A slightly more general derivation would show that $D(\phi_1(x;\mu);p)$ exists for any x and p.

Def: A point $\hat{x} \in \mathbb{R}^n$ is a stationary point for ϕ_1 if $D(\phi_1(\hat{x};\mu);p) \geq 0 \quad \forall p \in \mathbb{R}^n$.

Similarly, $\hat{x} \in \mathbb{R}^n$ is a stationary point for the infeasibility measure

$$h(x) = \sum_{i \in \mathcal{E}} |c_i(x)| + \sum_{i \in \mathcal{I}} [c_i(x)]^- \quad \text{if} \quad D(h(\hat{x});p) \geq 0 \quad \forall p \in \mathbb{R}^n.$$

If $\hat{x}$ is infeasible (i.e. $h(\hat{x}) > 0$) and is stationary for h, we say that it is an infeasible stationary point.

For $\min x$ s.t. $x \geq 1$, the solution is $x^* = 1$,

$$\phi_1(x;\mu) = x + \mu[x-1]^- = \begin{cases} x, & x \geq 1 \\ (1-\mu)x + \mu, & x < 1 \end{cases}$$

$$D(\phi_1(x^*;\mu);p) = \begin{cases} p & \text{if } p \geq 0 \\ (1-\mu)p & \text{if } p < 0. \end{cases}$$

So when $\mu \geq 1$, $D(\phi_1(x^*;\mu);p) \geq 0$ for all $p \in \mathbb{R}$.

The following is a partial converse of the previous result:

Theorem 17.4.

Suppose that $\hat{x}$ is a stationary point of the penalty function $\phi_1(x;\mu)$ for all μ greater than a certain threshold $\hat{\mu} > 0$. Then, if $\hat{x}$ is feasible for the nonlinear program (17.6), it satisfies the KKT conditions (12.34) for (17.6). Is $\hat{x}$ is not feasible for (17.6), it is an infeasible stationary point.

Proof: Suppose first that $\hat{x}$ is feasible. It can be shown that

$$D(\phi_1(\hat{x};\mu);p) = \nabla f(\hat{x})^T p + \mu \sum_{i\in\mathcal{E}} |\nabla c_i(\hat{x})^T p| + \mu \sum_{i\in\mathcal{I}\cap\mathcal{A}(\hat{x})} [\nabla c_i(\hat{x})^T p]^- \quad -①$$

↑ the active set $\mathcal{A}(\hat{x}) = \{i : c_i(\hat{x}) = 0\}$.

(Ex: check it.)

Consider any $p \in \mathcal{F}(\hat{x})$ = the linearized feasible direction set

$$= \{ d \in \mathbb{R}^n : d^T c_i(\hat{x}) = 0,\ i\in\mathcal{E},\ d^T c_i(x) \geq 0,\ i \in \mathcal{I}\cap\mathcal{A}(\hat{x})\},$$

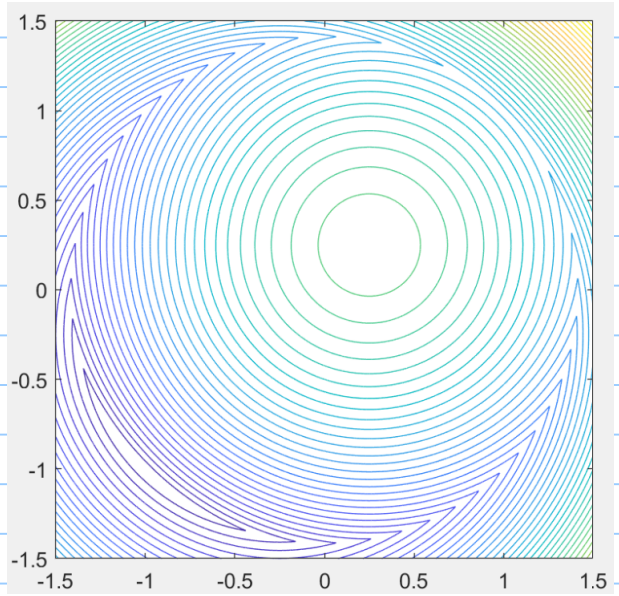
then $$\sum_{i\in\mathcal{E}} |\nabla c_i(\hat{x})^T p| + \sum_{i\in\mathcal{I}\cap\mathcal{A}(\hat{x})} [\nabla c_i(\hat{x})^T p]^- = 0. \quad -②$$

By assumption, $D(\phi_1(\hat{x};\mu);p) \geq 0$. $\quad -③$

①+②+③ $\Rightarrow$ $\nabla f(\hat{x})^T p \geq 0$, $\forall p \in \mathcal{F}(\hat{x})$ $\quad -(*)$

And if you review the proof of KKT, $(*) \Rightarrow$ the KKT conditions
(a key step involves Farkas' lemma.)

The proof of the infeasible case is similar. Q.E.D.



Contours of $\phi_1(x;\mu) = x_1 + x_2 + \frac{2}{\mu}|x_1^2+x_2^2-2|$.

(Observe the kinks on the circle $x_1^2+x_2^2-2=0$.)

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

$\nabla_x\mathcal{L}(x,\lambda) = \begin{bmatrix} 1-2\lambda x_1 \\ 1-2\lambda x_2 \end{bmatrix}$ $\nabla\mathcal{L}(x,\lambda) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ when $x_1=x_2=\frac{1}{2\lambda}$

$x_1^2+x_2^2-2=0 \Rightarrow 2\frac{1}{4\lambda^2}-2=0 \Rightarrow \lambda^2=\frac{1}{4} \Rightarrow$

KKT points: $\lambda^* = \pm\frac{1}{2}$ and $x^* = \pm\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

minimizer: $x^* = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\lambda^* = -\frac{1}{2}$

Thm 17.3 tells us that when $\mu > |\lambda^*| = \frac{1}{2}$, $\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ is also a local minimizer of $\phi_1(x;\mu)$.

Framework 17.2 (Classical ℓ_1 Penalty Method).

Given $\mu_0 > 0$, tolerance $\tau > 0$, starting point x_0^s;

for $k = 0, 1, 2, \ldots$

Find an approximate minimizer x_k of $\phi_1(x; \mu_k)$, starting at x_k^s; ← How?

if $h(x_k) \leq \tau$

stop with approximate solution x_k;

end (if)

Choose new penalty parameter $\mu_{k+1} > \mu_k$;

Choose new starting point x_{k+1}^s;

end (for)

There are techniques for such non-smooth problems, such as the "bundle methods".

But the following approach based on QP is found to work well in practice.

Think : $\phi_1(x+p;\mu) = f(x+p) + \mu \sum_{i\in\mathcal{E}} |c_i(x+p)| + \mu \sum_{i\in\mathcal{I}} [c_i(x+p)]^-$

$\approx f(x) + \nabla f(x)^T p + \frac{1}{2} p^T W p + \mu \sum_{i\in\mathcal{E}} |c_i(x) + \nabla c_i(x)^T p| + \mu \sum_{i\in\mathcal{I}} [c_i(x) + \nabla c_i(x)^T p]^-$

↑ some symmetric matrix which contains 2nd derivative information about f and c_i, $i \in \mathcal{E} \cup \mathcal{I}$, as in Newton or quasi-Newton methods.

Then, for any fixed μ $(= \mu_k)$ in the "outer loop" of Framework 17.2, solve the subproblem $\min \phi(x;\mu)$ iteratively:

"inner loop"

For $l = 0, 1, \cdots$

Solve $p^l = \arg\min_p f(x^l) + \nabla f(x^l)p + \frac{1}{2}p^T W p + \mu \sum_{i \in \mathcal{E}} |c_i(x^l) + \nabla c_i(x^l)^T p| + \mu \sum_{i \in \mathcal{I}} [c_i(x^l) + \nabla c_i(x^l)^T p]^-$ — (✱)

$x^{l+1} \leftarrow x^l + p^l$

We will justify this inner loop later, for now let's see that the subproblem (✱) can be reformulated as a QP.

For any $x \in \mathbb{R}$,

$$x = \max(x, 0) - \max(-x, 0)$$
$$|x| = \max(x, 0) + \max(-x, 0)$$
$$[x]^- = \max(-x, 0).$$

So the non-smooth, unconstrained problem (✱) can be reformulated as:

(drop 'l', write x^l as x)

$$\min_{p,r,s,t} \; f(x) + \nabla f(x)^T p + \tfrac{1}{2} p^T W p + \mu \sum_{i \in \mathcal{E}} (r_i + s_i) + \mu \sum_{i \in \mathcal{I}} t_i$$

s.t.

$$\begin{aligned} c_i(x) + \nabla c_i(x)^T p &= r_i - s_i \;, && i \in \mathcal{E} \\ c_i(x) + \nabla c_i(x)^T p &\geq -t_i \;, && i \in \mathcal{I} \end{aligned} \qquad , \quad r, s, t \geq 0.$$

We may also add a "box-shaped" trust region constraint of the form $\|p\|_\infty \leq \Delta$ ($\iff -\Delta \leq p_i \leq \Delta$). It does not change the fact that the subproblem above is a QP.

All these techniques (exact penalty, quasi-Newton, QP, trust region, etc.) used together in a way similar to the above would accumulate into a successful method for constrained optimization called

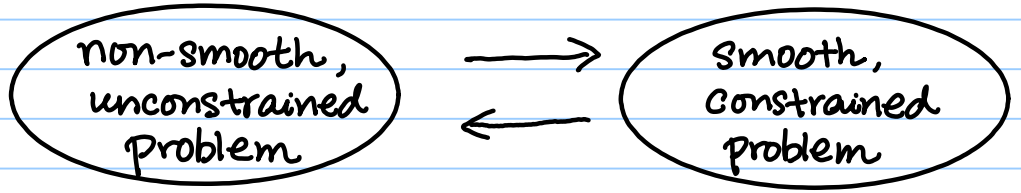
sequential quadratic programming.

(The penultimate chapter of N&W (ed 2) is devoted to SQP.)

Final remarks about exact penalty methods

1. We actually saw the idea used in an opposite way before.

   In CO_1, I showed how to convert an unconstrained, non-smooth problem involving $|\cdot|$ (notably the L^1-regression problem) into a constrained problem with smooth — in fact linear in the case of L^1-regression — objective and constraint functions.



2. One may wonder if it is truely necessary to give up smoothess for exactness. The answer is affirmative.

   For simplicity, consider: $\min f(x)$ s.t. $c_1(x) = 0$.

   Assume, instead of $|\cdot|$, it is possible to find a C^1 function $h: \mathbb{R} \to \mathbb{R}$ s.t. $h(y) \geq 0$, $h(0) = 0$, and the h-penalized problem $\min_x \phi(x; \mu)$

$$\phi(x;\mu) := f(x) + \mu h(c_1(x)),$$

is exact for the constrained problem for some $\mu > 0$, i.e.

$$x^* \text{ solves } \min_x \phi(x;\mu) \Rightarrow x^* \text{ solves } \min f(x) \text{ s.t. } c_1(x) = 0.$$

Since $h(y) \geq 0$ and $h(0) = 0$, then 0 is a minimizer of h, which implies $h'(0) = 0$.

If x^* solves $\min f(x)$ s.t. $c_1(x) = 0$, then $c_1(x^*) = 0$, so $h'(c_1(x^*)) = 0$.

Since we assume x^* also solves $\min_x \phi(x;\mu)$,

$$0 = \nabla\phi(x^*;\mu) = \nabla f(x^*) + \mu h'(c_1(x^*)) \nabla c_1(x^*) = \nabla f(x^*).$$

However, almost any constrained problem you write down would not have a solution at which ∇f vanishes.

So the assumption that h is C^1 must be incorrect.

III. Augmented Lagrangian Method

When x_k is an exact or approximate minimizer of $Q(\cdot;\mu_k) = f(\cdot) + \frac{\mu_k}{2}\sum_{i\in\mathcal{E}} c_i^2(\cdot)$,

- x_k does not quite satisfy the feasibility conditions $c_i(x) = 0$, $i \in \mathcal{E}$
- instead, when k is large, we expect $c_i(x_k) \approx \lambda_i^*/\mu_k$, $i \in \mathcal{E}$.

[$Q(x;\mu_k) = f(x) + \frac{\mu_k}{2}\sum_{i\in\mathcal{E}} c_i^2(x)$. If x_k minimizes $Q(\cdot;\mu_k)$, then

$$\nabla Q(x_k;\mu_k) = \nabla f(\underbrace{x_k}_{\approx x^*}) + \sum_{i\in\mathcal{E}} \underbrace{\mu_k c_i(x_k)}_{\approx -\lambda_i^*} \nabla c_i(\underbrace{x_k}_{\approx x^*}) \approx 0 \qquad (\mu_k \text{ large}) \ ]$$

Certainly $c_i(x_k) \approx \lambda_i^*/\mu_k \to 0$ as $k \to \infty$. (Thm 17.2 makes this rigorous.)

The idea is to alter the function $Q(x;\mu_k)$ to make the approximate minimizers more nearly satisfy the constraints $c_i(x) = 0$, even for moderate values of μ_k.

The trick is to consider

$$\mathcal{L}_A(x,\lambda;\mu) := f(x) - \sum_{i\in\mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(x).$$

a vector (of length $|\mathcal{E}|$), so we put the iteration count in the superscript

If x_k approximately minimizes $\mathcal{L}_A(\cdot, \lambda^k; \mu_k)$, then

$$0 \approx \nabla_x \mathcal{L}_A(x_k, \lambda^k; \mu_k) = \nabla f(x_k) - \sum_{i \in \mathcal{E}} [\lambda_i^k - \mu_k c_i(x_k)] \nabla c_i(x_k)$$

With some luck,

$$x_k \approx x^*, \quad \lambda_i^k - \mu_k c_i(x_k) \approx \lambda_i^* \quad \text{as } k \uparrow.$$

$$\Downarrow$$

$$c_i(x_k) \approx -\frac{1}{\mu_k}(\lambda_i^* - \lambda_i^k) \quad i \in \mathcal{E}, \qquad (*)$$

meaning: 1. $c_i(x_k) \approx 0$ if μ_k is large and/or $\lambda_i^k \approx \lambda_i^*$

meaning that the method does not need to rely only on μ_k being large. ← source of ill-conditioning

2. with λ^k, μ_k, x_k in place, we may update λ^k by

$$\lambda_i^{k+1} \leftarrow \lambda_i^k - \mu_k c_i(x_k), \quad \text{as suggested by } (*).$$

(17.39)

This discussion motivates the following algorithmic framework.

Framework 17.3 (Augmented Lagrangian Method-Equality Constraints).

Given $\mu_0 > 0$, tolerance $\tau_0 > 0$, starting points x_0^s and λ^0;

for $k = 0, 1, 2, \ldots$

Find an approximate minimizer x_k of $\mathcal{L}_A(\cdot, \lambda^k; \mu_k)$, starting at x_k^s,
and terminating when $\|\nabla_x \mathcal{L}_A(x_k, \lambda^k; \mu_k)\| \leq \tau_k$;

if a convergence test for (17.1) is satisfied

stop with approximate solution x_k;

end (**if**)

Update Lagrange multipliers using (17.39) to obtain λ^{k+1};

$\lambda_i^{k+1} \leftarrow \lambda_i^k - \mu_k c_i(x_k)$

Choose new penalty parameter $\mu_{k+1} \geq \mu_k$;

Set starting point for the next iteration to $x_{k+1}^s = x_k$;

Select tolerance τ_{k+1};

end (**for**)

Consider again $\min x_1 + x_2$ s.t. $x_1^2 + x_2^2 - 2 = 0$. Sol at $x^* = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\lambda^* = -0.5$

minimizer of $Q(x; \mu) = x_1 + x_2 + \frac{\mu}{2}(x_1^2 + x_2^2 - 2)^2$

$\approx \begin{bmatrix} -1.1 \\ -1.1 \end{bmatrix}$ when $\mu = 1$.

Compare:

minimizer of $\mathcal{L}_A(x, \lambda; \mu) = x_1 + x_2 - \lambda(x_1^2 + x_2^2 - 2) + \frac{\mu}{2}(x_1^2 + x_2^2 - 2)^2$

$\approx \begin{bmatrix} -1.02 \\ -1.02 \end{bmatrix}$ when $\lambda = -0.4$, $\mu = 1$ ← closer to $\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ compared to $\begin{bmatrix} -1.1 \\ -1.1 \end{bmatrix}$.

Theoretical Results:

Theorem 17.5.

Let x^* be a local solution of (17.1) at which the LICQ is satisfied (that is, the gradients $\nabla c_i(x^*)$, $i \in \mathcal{E}$, are linearly independent vectors), and the second-order sufficient conditions specified in Theorem 12.6 are satisfied for $\lambda = \lambda^*$. Then there is a threshold value $\bar{\mu}$ such that for all $\mu \geq \bar{\mu}$, x^* is a strict local minimizer of $\mathcal{L}_A(x, \lambda^*; \mu)$.

i.e. no need to increase μ to arbitrarily large values, well, if we know λ^*.

The next result describes the more realistic situation of $\lambda \neq \lambda^*$.

Theorem 17.6.

Suppose that the assumptions of Theorem 17.5 are satisfied at x^* and λ^* and let $\bar{\mu}$ be chosen as in that theorem. Then there exist positive scalars δ, ϵ, and M such that the following claims hold:

ind. of k

(a) *For all λ^k and μ_k satisfying*

$$\|\lambda^k - \lambda^*\| \leq \mu_k \delta, \qquad \mu_k \geq \bar{\mu}, \tag{17.44}$$

the problem

$$\min_x \mathcal{L}_A(x, \lambda^k; \mu_k) \quad \text{subject to } \|x - x^*\| \leq \epsilon$$

independent of k

has a unique solution x_k. Moreover, we have

$$\|x_k - x^*\| \leq M\|\lambda^k - \lambda^*\|/\mu_k. \tag{17.45}$$

(b) *For all λ^k and μ_k that satisfy (17.44), we have*

$$\|\lambda^{k+1} - \lambda^*\| \leq M\|\lambda^k - \lambda^*\|/\mu_k, \tag{17.46}$$

where λ^{k+1} is given by the formula (17.39).

(c) *For all λ^k and μ_k that satisfy (17.44), the matrix $\nabla^2_{xx}\mathcal{L}_A(x_k, \lambda^k; \mu_k)$ is positive definite and the constraint gradients $\nabla c_i(x_k)$, $i \in \mathcal{E}$, are linearly independent.*

Practical Augmented Lagrangian Methods

Bound-Constrained Formulation (used in the successful code LANCELOT)

We discussed only problems with equality constraints.

We may express an inequality constraint $c_i(x) \geq 0$ as

$$c_i(x) - s_i = 0 \,, \quad s_i \geq 0 \,.$$

So a general constrained optimization problem can always be written as

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad c_i(x) = 0 \,, \; i = 1, \cdots, m \,, \quad \underbrace{\begin{bmatrix} l_1 \\ \vdots \\ l_n \end{bmatrix}}_{l} \leq \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{x} \leq \underbrace{\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}}_{u} \,. \qquad (B)$$

Note: The slacks s_i have been incorporated into the vector x and the objective f and the constraint functions c_i have been redefined accordingly.

Some of components of l may be set to $-\infty$, signifying that there is no lower bound on the component of x in question, similarly for u.

The bound-constrained Lagrangian (BCL) approach incorporates only the equality constraints from (B) into the augmented Lagrangian, i.e.

$$\mathcal{L}_A(x, \lambda; \mu) = f(x) - \sum_{i=1}^{m} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i=1}^{m} c_i^2(x).$$

The bound constraints are enforced explicitly in the subproblem, which has the form

$$\min_x \ \mathcal{L}_A(x, \lambda; \mu) \quad \text{s.t.} \quad l \leq x \leq u. \qquad \text{(BL)}$$

After this problem has been solved approximately, the multipliers λ and the penalty parameter μ are updated and the process is iterated.

So what's the point of all these? Ans: The equality constraints and the general nonlinear constraints are all 'absorbed' into the augmented Lagrangian. The only constraints left in (BL) are simple (linear) bound constraints on the variables.

There is an efficient (SQP-based) technique for solving the bound constrained problem (BL) (for fixed λ and μ), to be presented in the SQP chapter later.

For now, let's see how the KKT conditions of any bound constraints problem

$$\min_x g(x) \quad \text{st.} \quad l \le x \le u \qquad (G)$$

can be expressed by the simple projection operator $P(\cdot; l, u) = P : \mathbb{R}^n \to [l, u]$ (where $[l, u] = [l_1, u_1] \times \cdots \times [l_n, u_n]$) defined as follows

$$P(y; l, u)_i := \begin{cases} l_i & \text{if} \quad l_i \le y_i \\ y_i & \text{if} \quad l_i < y_i < u_i \\ u_i & \text{if} \quad u_i < y_i \end{cases}.$$

The KKT conditions of (G) are:

$$\nabla g(x) - \lambda_l^T \nabla_x (x - l) - \lambda_u^T \nabla_x (u - x) = 0 \quad , \quad \lambda_l, \lambda_u \ge 0 \quad , \quad l \le x \le u$$

$$\Downarrow$$

$$\nabla g(x) - \lambda_l + \lambda_u = 0$$

Ex: Show that these conditions are equivalent to $x - P(x - \nabla g(x); l, u) = 0$.

Algorithm 17.4 (Bound-Constrained Lagrangian Method). implemented in the LANCELOT package

Choose an initial point x_0 and initial multipliers λ^0;
Choose convergence tolerances η_* and ω_*;
Set $\mu_0 = 10$, $\omega_0 = 1/\mu_0$, and $\eta_0 = 1/\mu_0^{0.1}$;
for $k = 0, 1, 2, \ldots$
    Find an approximate solution x_k of the subproblem (17.50) such that (BL) ← done using a SQP algorithm (CHI18)

$$\left\| x_k - P\left(x_k - \nabla_x \mathcal{L}_A(x_k, \lambda^k; \mu_k), l, u\right)\right\| \le \omega_k;$$

    if $\|c(x_k)\| \le \eta_k$
        (* test for convergence *)
        if $\|c(x_k)\| \le \eta_*$ and $\left\| x_k - P\left(x_k - \nabla_x \mathcal{L}_A(x_k, \lambda^k; \mu_k), l, u\right)\right\| \le \omega_*$
            stop with approximate solution x_k;
        end (if)
        (* update multipliers, tighten tolerances *)
        $\lambda^{k+1} = \lambda^k - \mu_k c(x_k)$;
        $\mu_{k+1} = \mu_k$;
        $\eta_{k+1} = \eta_k / \mu_{k+1}^{0.9}$;
        $\omega_{k+1} = \omega_k / \mu_{k+1}$;
    else
        (* increase penalty parameter, tighten tolerances *)
        $\lambda^{k+1} = \lambda^k$;
        $\mu_{k+1} = 100\mu_k$;
        $\eta_{k+1} = 1/\mu_{k+1}^{0.1}$;
        $\omega_{k+1} = 1/\mu_{k+1}$;
    end (if)
end (for)

There are other related methods for dealing with inequality constraints; see Sec 17.4.
Another successful package called MINOS uses such an augmented Lagrangian based algorithm.