Numerical Optimization

Lecture 16: Penalty Methods

王浩

信息科学与技术学院

Email: wanghao1@shanghaitech.edu.cn

Outline

Outline

Challenges of nonlinear optimization

Theory and algorithms for linear/quadratic optimization are (mostly) clear-cut.

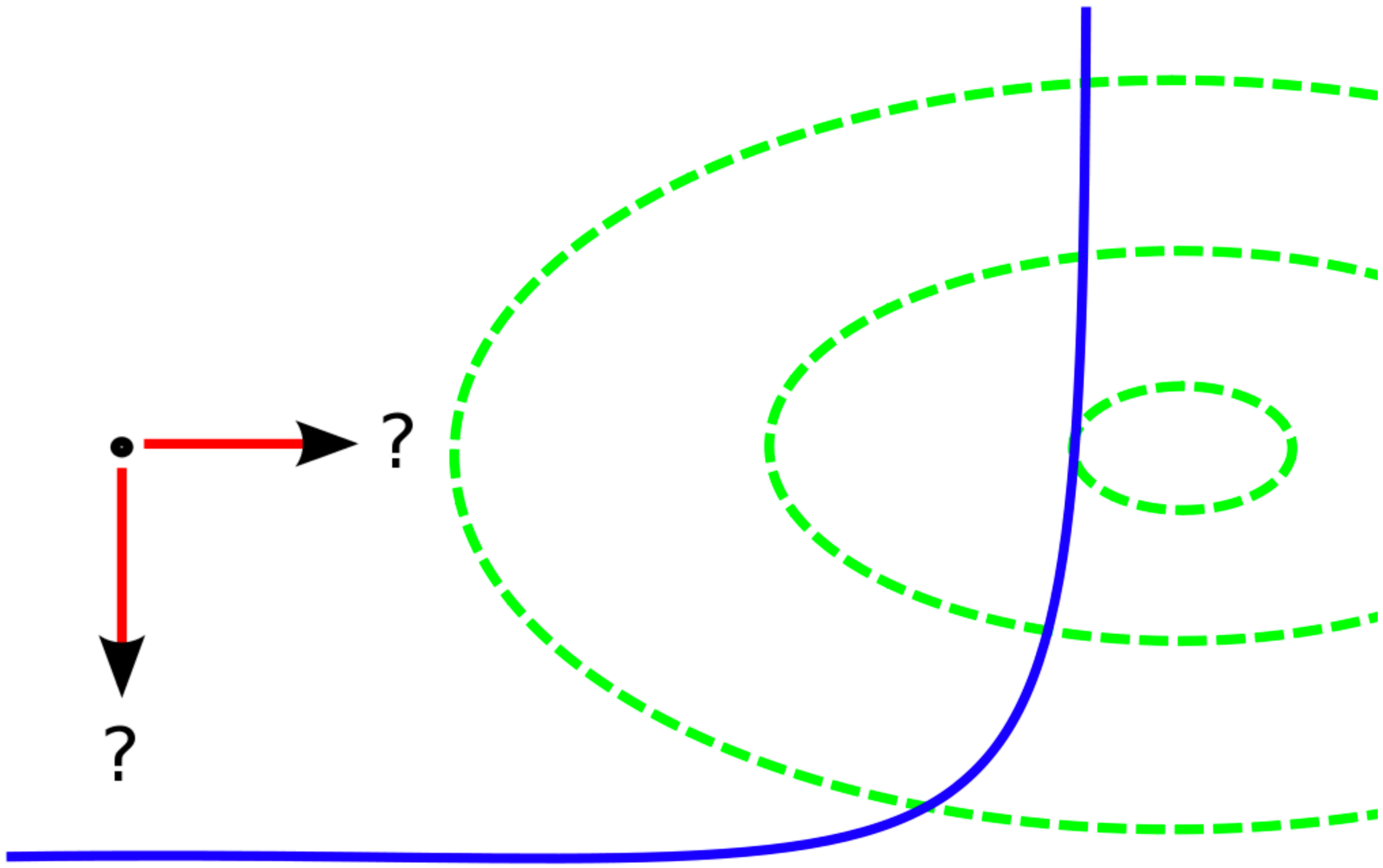
- Constraints are affine, so optimal solutions are always KKT points.
- Feasible region is a polytope.
- It generally makes sense to work to find a feasible point first, then optimize.[1]

The same is not always true for general nonlinear problems.

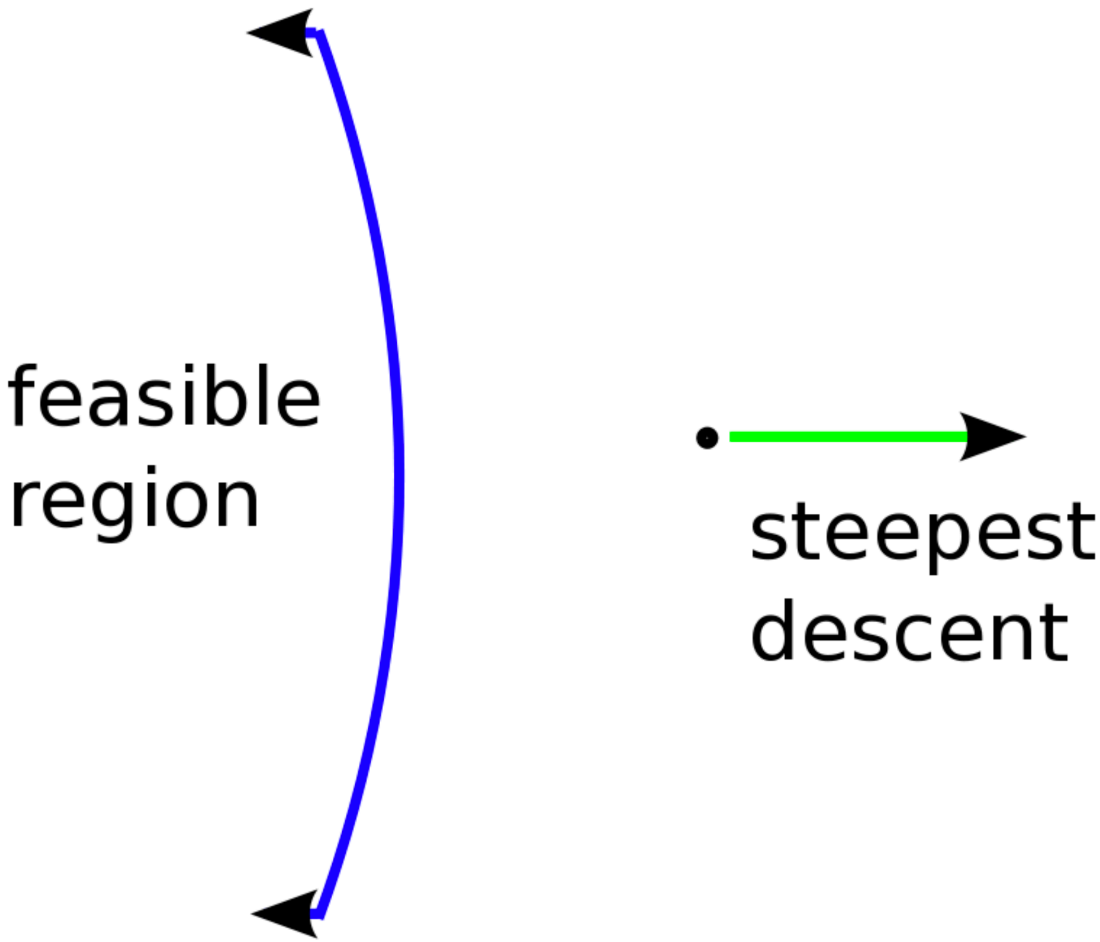
- Lagrange multipliers may not exist. (Should your algorithm be primal-dual?)
- Feasible region may be complicated. (Constraint elimination may fail.)
- Finding a feasible point first may not be the most efficient approach.

[1] But not always; e.g., infeasible interior-point methods are very efficient.

Reduce constraint violation or objective first?



Reduce constraint violation or objective first?



Penalty methods

Early attempts for solving constrained problems were penalty methods.

- Try to avoid difficulties involved in having constraints.
- Try to employ wealth of algorithms for unconstrained optimization.
- Try to avoid dealing directly with Lagrange/dual variables (for the most part).

It will be sufficient to focus primarily on equality constrained problems:

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & c(x) = 0. \end{aligned}$$

We will find useful ideas/tools in our study of penalty methods, but algorithm are generally better if they "handle constraints as constraints".

Outline

Problem reformulation

Reformulate the constrained problem

$$\min_{x} \ f(x)$$
$$\text{s.t. } c(x) = 0$$

as the unconstrained quadratic penalty subproblem

$$\min_{x} \ \phi(x;\nu) := f(x) + \frac{\nu}{2}\|c(x)\|_2^2,$$

where $\nu \geq 0$ is a penalty parameter.

- Why a squared norm? If c is differentiable everywhere, then so is ϕ.
- Penalty parameter balances weight on objective and constraint violation.

Example of quadratic penalty

Recall the equality constrained problem

$$\min_x \; x_1 + x_2$$
$$\text{s.t. } x_1^2 + x_2^2 - 2 = 0$$

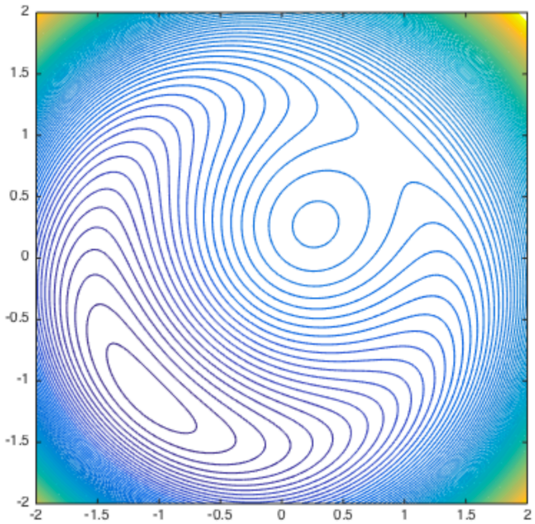
for which we have the corresponding quadratic penalty subproblem

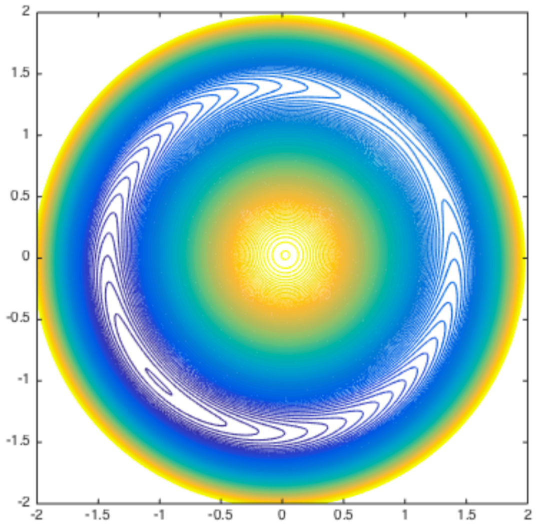
$$\min_x \;\; x_1 + x_2 + \tfrac{\nu}{2}(x_1^2 + x_2^2 - 2)^2.$$

- The solution to the constrained problem is $x_* = (-1, -1)$.
- The solution to the quadratic penalty subproblem depends on ν...

Illustration of quadratic penalty

Contours of quadratic penalty for $\nu = 1$ (left) and $\nu = 10$ (right).





As $\nu \to \infty$, minimizer of quadratic penalty approaches x_*.

Quadratic penalty algorithm

Algorithm 1 Quadratic Penalty Algorithm (Sketch)

1: Choose $\nu_0 > 0$ and $\{\tau_k\} \to 0$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Find an approximate solution x_k to

$$\min_x \ \phi(x; \nu_k) = f(x) + \tfrac{\nu_k}{2} \|c(x)\|_2^2$$

satisfying

$$\|\nabla \phi(x_k; \nu_k)\| \le \tau_k.$$

4: If an optimality test for the constrained problem is satisfied, then stop.
5: Choose $\nu_{k+1} > \nu_k$.
6: **end for**

Convergence result

Theorem

Suppose $\nu_k \to \infty$ and each x_k is the exact global minimizer of $\phi(\cdot;\nu_k)$. Then, every limit point of $\{x_k\}$ is a global solution of the constrained problem.

- Nice! But...
- Unfortunately, finding exact global minimizers for each ν_k is expensive.
- There are other concerns, too, but first a more practical result...

Practical convergence result

Theorem

Suppose $\nu_k \to \infty$ and $\tau_k \to 0$. Let x_* be any limit point of $\{x_k\}$.

- *If x_* is infeasible, then it is a stationary point of $\|c(x)\|^2$.*
- *If x_* is feasible and the constraint gradients are linearly independent, then x_* is a KKT point for the constrained problem. Moreover, for such points, we have that for any infinite subsequence $\mathcal{K}$ such that*

$$\lim_{k \in \mathcal{K}} x_k = x_*$$

the following limit holds

$$\lim_{k \in \mathcal{K}} \nu_k c^i(x_k) = \lambda_*^i,$$

where λ_* is a multiplier vector satisfying the KKT condition

$$\nabla f(x_*) + \nabla c(x_*)\lambda_* = 0.$$

Comments on quadratic penalization theorem

First some comments, then a quick proof.

- We cannot guarantee that we find a feasible point.
- We only guarantee that we find a stationary point of $\|c(x)\|^2$, i.e., a point with

  $$\nabla c(x)c(x) = \sum c^i(x)\nabla c^i(x) = 0.$$

  This may mean $c^i(x) = 0$ for $i \in \mathcal{E}$, but perhaps not if the constraint gradients are linearly dependent for some x.
- If we converge to a feasible point, then we need the LICQ to prove it's a KKT point. The multipliers are then revealed by the penalty parameter and constraint values.

Proof of practical convergence result

Proof, part 1.

- By design of the algorithm, x_k satisfies

$$\|\nabla\phi(x_k;\nu_k)\| = \left\|\nabla f(x_k) + \sum \nu_k c^i(x_k)\nabla c^i(x_k)\right\| \le \tau_k.$$

- Rearranging, and using the inequality $\|a\| - \|b\| \le \|a+b\|$, we find

$$\left\|\sum c^i(x_k)\nabla c^i(x_k)\right\| \le \frac{1}{\nu_k}(\tau_k + \|\nabla f(x_k)\|).$$

- Since $\nu_k \to \infty$ and $\nabla f(x_k) \to \nabla f(x_*)$ for $k \in \mathcal{K}$, this implies

$$\left\|\sum c^i(x_*)\nabla c^i(x_*)\right\| = 0.$$

- This proves the infeasible case, and proves that we converge to a feasible point if the constraint gradients are linearly independent.

Proof of practical convergence result

Proof, part 2.

It remains to show that under LICQ we obtain a KKT point with the given λ_*.

- Define $\lambda_k = \nu_k c(x_k)$, which by definition of ϕ means that for all k

$$\nabla c(x_k)\lambda_k = -(\nabla f(x_k) - \nabla \phi(x_k; \nu_k)).$$

- For $k \in \mathcal{K}$ sufficiently large, $\nabla c(x_k)$ has full rank, so we can write

$$\lambda_k = -[\nabla c(x_k)^T \nabla c(x_k)]^{-1} \nabla c(x_k)^T (\nabla f(x_k) - \nabla \phi(x_k; \nu_k)).$$

- Taking limits on both sides, we obtain

$$\lim_{k \in \mathcal{K}} \lambda_k = -[\nabla c(x_*)^T \nabla c(x_*)]^{-1} \nabla c(x_*)^T \nabla f(x_*) =: \lambda_*$$

which implies that

$$\nabla f(x_*) + \nabla c(x_*)\lambda_* = 0.$$

Inequality constraints

Corresponding to the generally constrained problem

$$\begin{aligned} \min_{x} \ & f(x) \\ \text{s.t. } & c^{\mathcal{E}}(x) = 0 \\ & c^{\mathcal{I}}(x) \leq 0, \end{aligned}$$

we have the following quadratic penalty subproblem

$$\min_{x} \ \phi(x;\nu) := f(x) + \frac{\nu}{2}\|c^{\mathcal{E}}(x)\|_2^2 + \frac{\nu}{2}\|\max\{c^{\mathcal{I}}(x), 0\}\|_2^2.$$

Results similar to those discussed hold in the generally constrained case as well.

Drawbacks of quadratic penalization

Despite some nice features, there are major drawbacks to quadratic penalization.

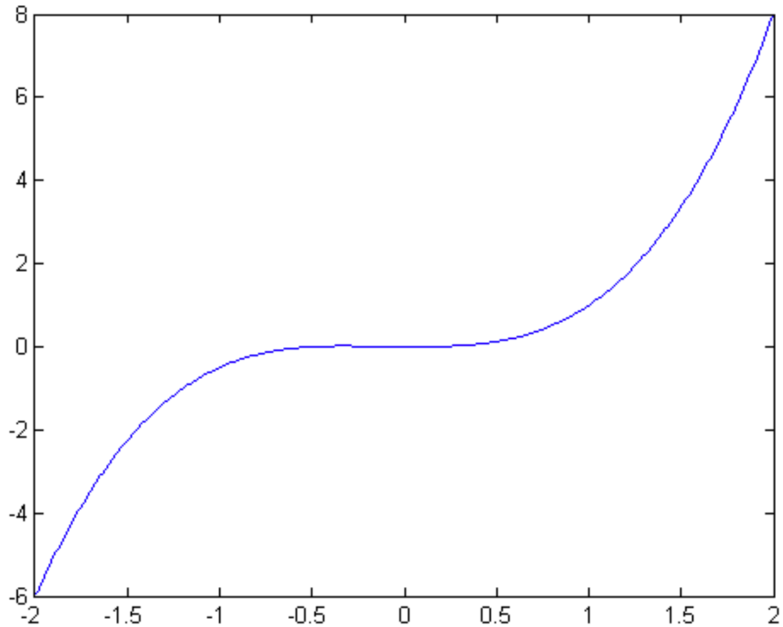
- How do we define $\{\nu_k\}$ so that we converge to a feasible point?
- We may have a hard time finding x_k satisfying

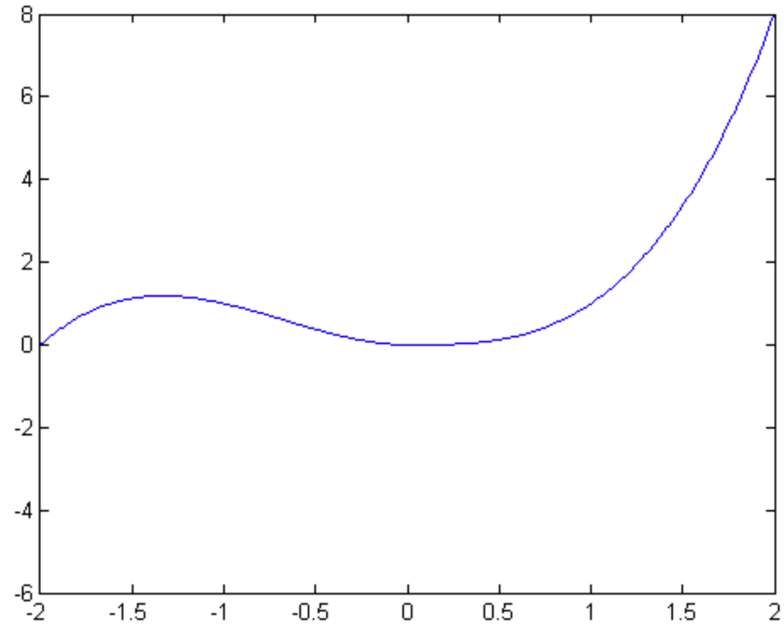
$$\|\nabla\phi(x_k;\nu_k)\| \leq \tau_k.$$

- Quadratic penalization leads to terrible ill-conditioning.

Poorly-posed quadratic penalty problems

$\min_x \ x^3$ s.t. $-x \leq 0$ yields $\phi(x;\nu) = x^3 + \frac{\nu}{2}\max\{-x,0\}^2$. (Here, $\nu = 1, 4$):



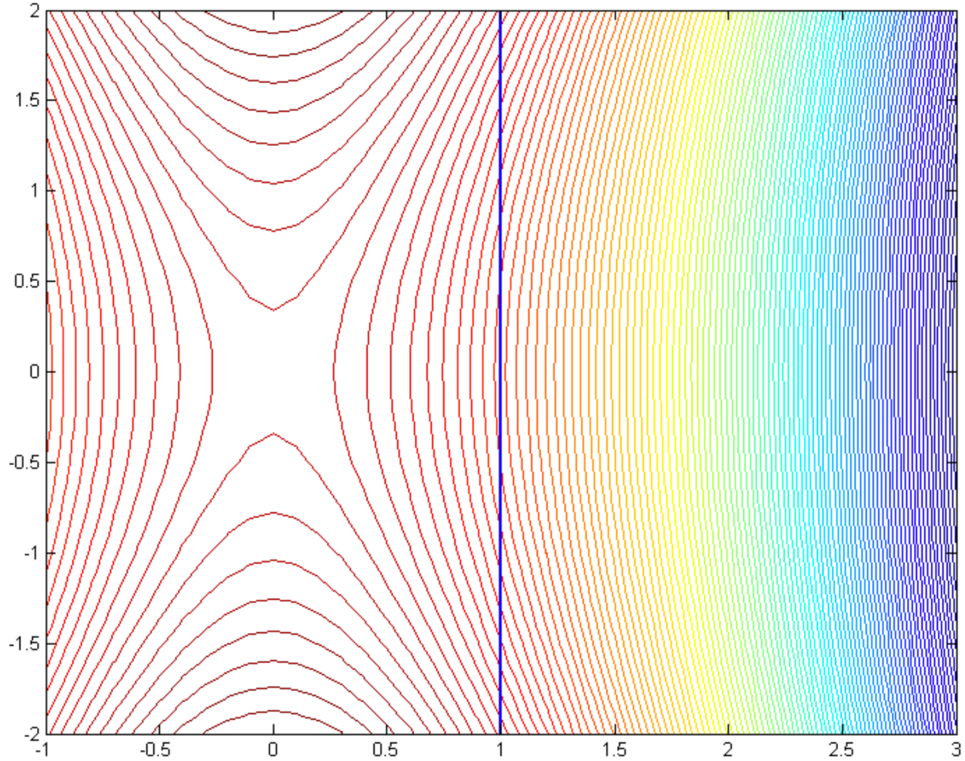


f is convex on the feasible set, but ϕ is unbounded below for all ν!

Poorly-posed quadratic penalty problems

The following constrained problem (convex on the feasible set!) has $x_* = (1, 0)$:

$$\min_x \; f(x) = -5x_1^2 + x_2^2$$
$$\text{s.t. } c(x) = x_1 - 1 = 0$$

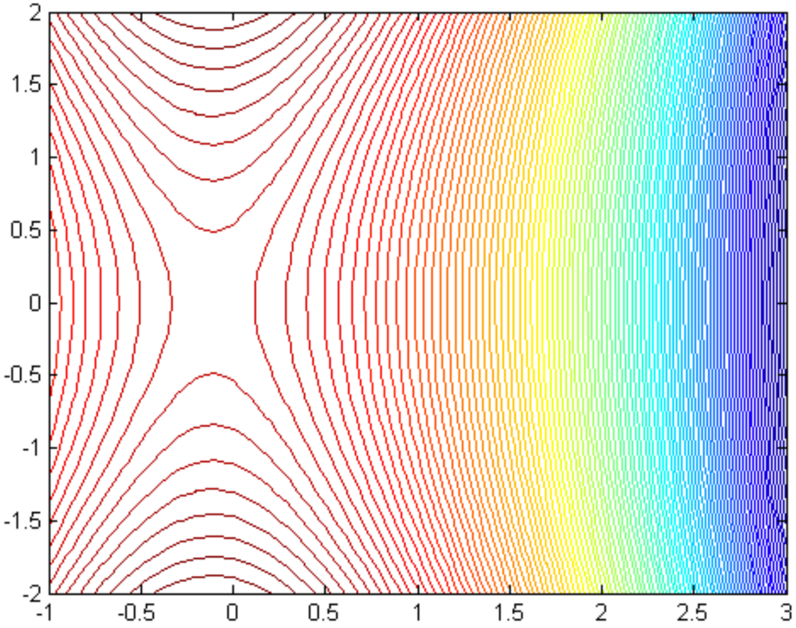


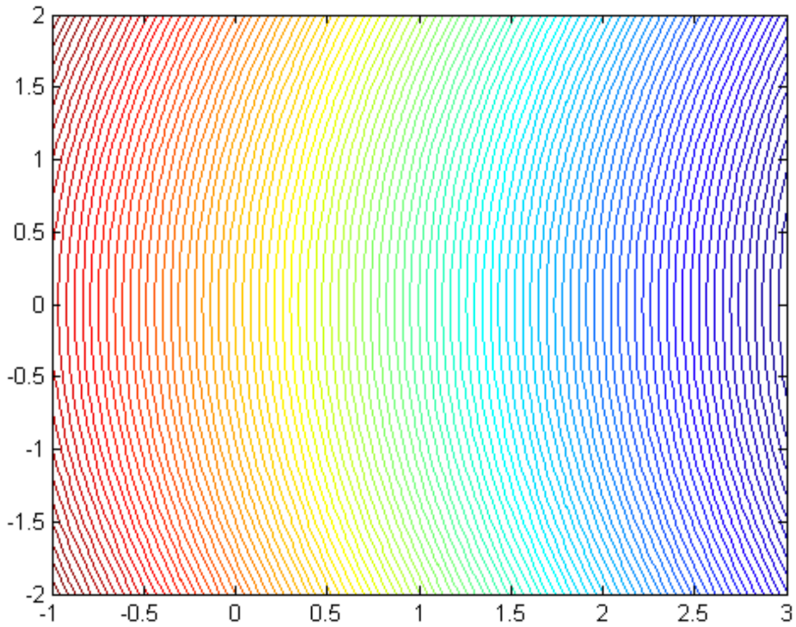
Poorly-posed quadratic penalty problems

However, the corresponding quadratic penalty subproblem

$$\min_{x} \ \phi(x;\nu) = -5x_1^2 + x_2^2 + \frac{\nu}{2}(x_1 - 1)^2$$

is unbounded below for all $\nu < 10$. (Below, $\nu = 1, 10.$)



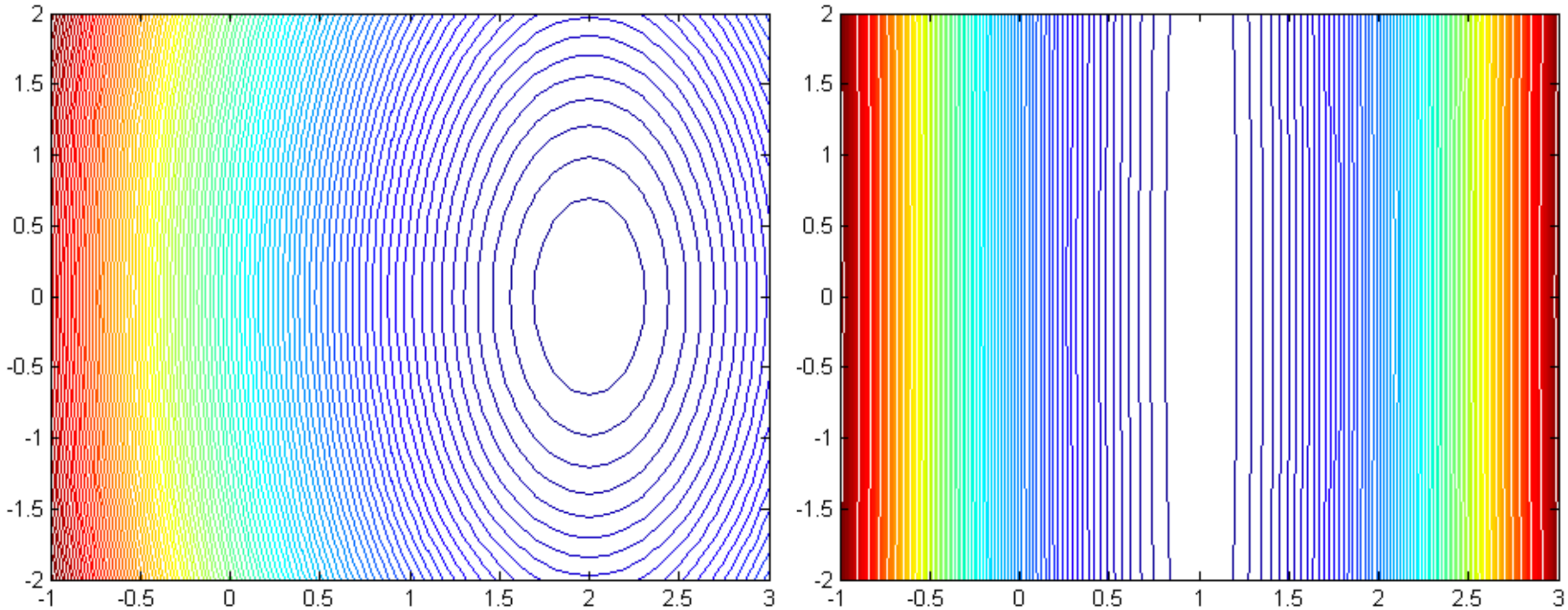


Poorly-posed quadratic penalty problems

However, the corresponding quadratic penalty subproblem

$$\min_x \ \phi(x;\nu) = -5x_1^2 + x_2^2 + \frac{\nu}{2}(x_1 - 1)^2$$

is unbounded below for all $\nu < 10$. (Below, $\nu = 20, 1000$.)



Ill-conditioning

It is apparent in this last example that the Hessian of the quadratic penalty function can become extremely ill-conditioned as $\nu \to \infty$:

$$\begin{aligned}\nabla^2 \phi(x;\nu) = & \ \nabla^2 f(x) + \sum \nu c^i(x) \nabla c^i(x) + \nu \nabla c(x) \nabla c(x)^T \\ = & \ \nabla^2_{xx} L(x,\lambda) + \nu \nabla c(x) \nabla c(x)^T.\end{aligned}$$

- Some eigenvalues of $\nabla^2 \phi(x;\nu)$ increase with ν, even if f and c are "nice".
- The multiplier estimate $\lambda_k = \nu_k c(x_k)$ may be a poor approximation for the optimal multipliers, which may adversely affect convergence of an unconstrained optimization method applied to the quadratic penalty subproblem.

Outline

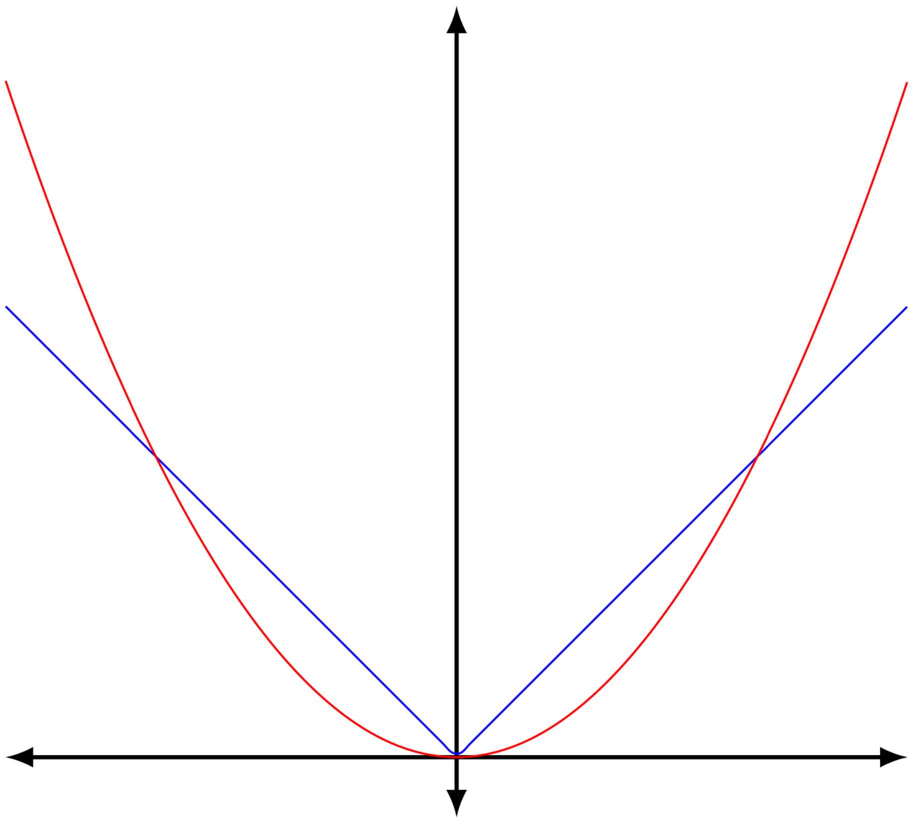
Issue with quadratic penalization

- The quadratic penalty function

$$\phi(x;\nu) = f(x) + \frac{\nu}{2}\|c(x)\|_2^2$$

  places a large penalty on points with large violations in the constraints.
- However, it is also true that small violations are not penalized very much.
- Thus, we are required to have $\nu \to \infty$ in order to truly satisfy the constraints.
- We will find a better trade-off by penalizing large violations less while penalizing small violations more than a quadratic penalty function.
- This will lead to better behavior for finite ν.

$\|c(x)\|^2$ versus $\|c(x)\|$



"Exact" penalization

Definition (Exact Penalty Function)

A penalty function $\phi(x;\nu)$ is exact if there exists ν_* such that for all $\nu > \nu_*$, a local solution of the constrained problem is a local minimizer of $\phi(x;\nu)$.

- The quadratic penalty function is not exact.
- The following ℓ_1 penalty function is exact:

$$\phi(x;\nu) := f(x) + \nu\|c(x)\|_1.$$

Theorem

Suppose that x_* is a strict local solution of the constrained problem at which the KKT conditions are satisfied with Lagrange multipliers λ_*. Then, x_* is a local minimizer of $\phi(x;\nu)$ for all $\nu > \nu_*$ where

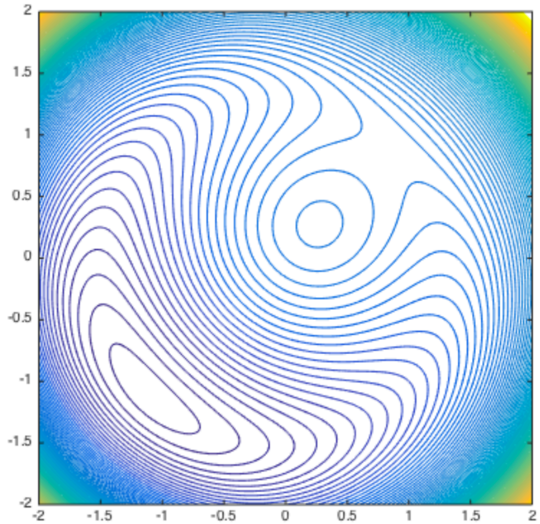
$$\nu_* = \|\lambda_*\|_\infty.$$

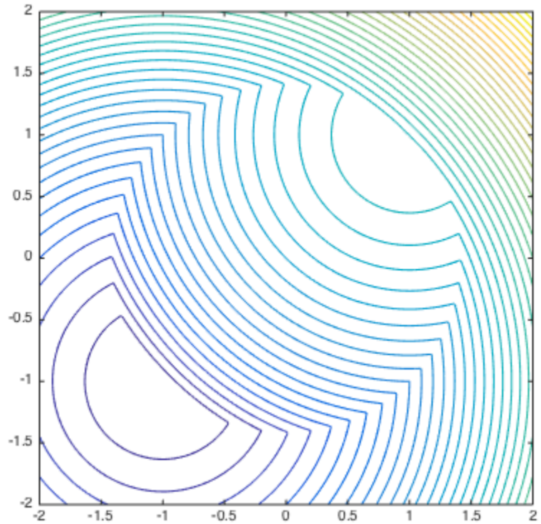
If, in addition, the second order sufficient conditions hold at x_* and $\nu > \nu_*$, then x_* is a strict local minimizer of $\phi(x;\nu)$.

(Note: For $\phi(x;\nu) := f(x) + \nu\|c(x)\|_a$, the same result holds with $\nu_* = \|\lambda_*\|_b$, where $\|\cdot\|_a$ and $\|\cdot\|_b$ are dual norms, e.g., $a = b = 2$, since ℓ_2 norm is self dual.)

Illustration of ℓ_1 penalty

Contours of quadratic penalty (left) and ℓ_1 penalty (right) for $\nu = 1$.





ℓ_1 penalizes small violations more than quadratic penalization.

Minimizing the ℓ_1 penalty function

We can now use a similar strategy as before, except that if things are nice, then for a finite value of ν we'll obtain the true solution to the constrained problem.

- However, there is a catch... $\phi(x;\nu)$ is nonsmooth!

Theorem

A point x_* is stationary for the ℓ_1 penalty function for $\nu = \nu_*$ if the directional derivative of $\phi(x;\nu_*)$ at $x = x_*$ is nonnegative for any d, i.e.,

$$D\phi(d;\nu_*,x_*) = \nabla f(x_*)^T d + \nu_* \sum |\nabla c^i(x_*)^T d| \geq 0 \quad \forall d.$$

(Recalling our subdifferential calculus, we know that $\phi(x;\nu)$ admits a directional derivative for any d. Also, note that this definition for $\nu = 0$ provides a definition for a stationary point for the measure of infeasibility $\|c(x)\|_1$.)

Definition (Infeasible Stationary Point)

If a point is infeasible for the constrained problem, but is stationary for the infeasibility measure $\|c(x)\|_1$, then it is an infeasible stationary point.

ℓ_1 penalty algorithm

Algorithm 2 ℓ_1 Penalty Algorithm (Sketch)

1: Choose $\nu_0 > 0$ and $\epsilon > 0$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Find an approximate solution x_k to

$$\min_x \ \phi(x; \nu_k) = f(x) + \nu_k \|c(x)\|_1.$$

4: If $\|c(x_k)\|_1 \leq \epsilon$, then stop.
5: Choose $\nu_{k+1} > \nu_k$.
6: **end for**

Convergence result

Theorem

Suppose x_k is a stationary point for $\phi(x;\nu)$ for all ν greater than a certain threshold. If x_k is feasible, then there exists λ_k such that (x_k, λ_k) is a KKT point; otherwise, x_k is an infeasible stationary point.

That is, suppose you compute x_k by minimizing $\phi(x;\nu)$ for some "large" ν. If the minimizer doesn't change no matter how much you increase ν, then the result of the above theorem applies.

Drawbacks of ℓ_1 penalization

Do we have the same drawbacks as quadratic penalization?

- We still need to define $\{\nu_k\}$, but the choice is less important as we only need to eventually choose ν_k above a finite threshold.
- However, we may still have a hard time finding x_k that approximately minimizes the penalty function. For example, we may still have $\phi(x;\nu)$ unbounded below, even if f is convex and bounded below on the feasible set.
- Is ill-conditioning a problem? Not as much, again since we only need to eventually choose ν_k above a finite threshold.

Inequality constraints

Corresponding to the generally constrained problem

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & c^{\mathcal{E}}(x) = 0 \\ & c^{\mathcal{I}}(x) \le 0, \end{aligned}$$

we have the following ℓ_1 exact penalty problem

$$\min_x \ \phi(x;\nu) := f(x) + \nu \|c^{\mathcal{E}}(x)\|_1 + \nu \|\max\{c^{\mathcal{I}}(x), 0\}\|_1.$$

Results similar to those discussed hold in the generally constrained case as well.

Outline

Improving quadratic penalization

- The ℓ_1 exact penalty function represents an improvement over the quadratic penalty function since ν need not tend to ∞.
- Unfortunately, however, it is nonsmooth.
- An alternative way to improve the quadratic penalty function while maintaining smoothness is to allow alternative estimates of the optimal Lagrange multipliers.
- Recall that in quadratic penalization:

$$c^i(x_k) \approx \lambda_*^i / \nu_k,$$

so indeed $c^i(x_k) \to 0$ requires $\nu_k \to \infty$.

"Augmenting" the Lagrangian

Suppose that instead of combining a quadratic penalty term to f to get

$$f(x) + \frac{\nu}{2}\|c(x)\|^2,$$

we combine one to the Lagrangian:

$$\begin{aligned}\phi(x,\lambda;\nu) &= L(x,\lambda) + \frac{\nu}{2}\|c(x)\|_2^2 \\ &= f(x) + \lambda^T c(x) + \frac{\nu}{2}\|c(x)\|_2^2.\end{aligned}$$

- Why is this justified?
- The optimality conditions for the minimization of $\phi(\cdot,\cdot;\nu_k)$ are

$$\nabla_x \phi(x_k,\lambda_k;\nu_k) = \nabla f(x_k) + \sum(\lambda_k^i + \nu_k c^i(x_k))\nabla c^i(x_k).$$

- Relating this to the KKT conditions for the constrained problem, we see

$$\lambda_*^i \approx \lambda_k^i + \nu_k c^i(x_k),$$

or, in other words,

$$c^i(x_k) \approx (\lambda_*^i - \lambda_k^i)/\nu_k.$$

Comparison with quadratic penalization

- For quadratic penalization we have

$$c^i(x_k) \approx \lambda_*^i/\nu_k.$$

Thus, $c^i(x_k) \to 0$ requires $\nu_k \to \infty$.

- For the augmented Lagrangian we have

$$c^i(x_k) \approx (\lambda_*^i - \lambda_k^i)/\nu_k.$$

Thus, $c^i(x_k) \to 0$ requires $\nu_k \to \infty$ or $\lambda_k^i \to \lambda_*^i$.

- Handling λ_k as a distinct quantity now, the relationship

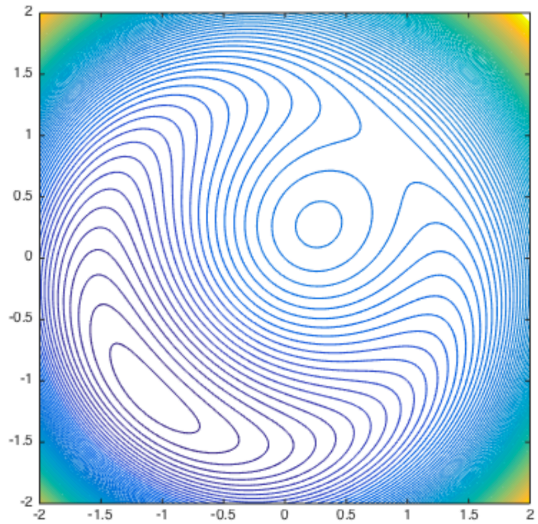
$$\lambda_*^i \approx \lambda_k^i + \nu_k c^i(x_k)$$

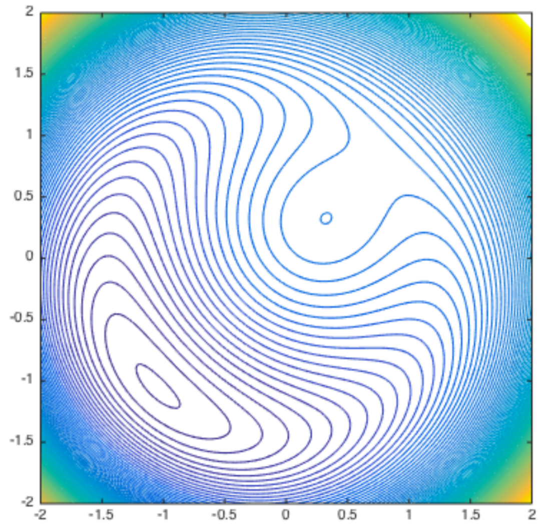
hints that a good update for λ may be

$$\lambda_{k+1}^i \leftarrow \lambda_k^i + \nu_k c^i(x_k).$$

Illustration of augmented Lagrangian

Contours of quadratic penalty (left) and aug. Lagrangian ($\lambda = \frac{2}{5}$, right) for $\nu = 1$:





Minimizer of aug. Lagrangian approaches x_* as $\nu \to \infty$ and/or $\lambda \to \lambda_* = \frac{1}{2}$.

Augmented Lagrangian algorithm (i.e. method of multipliers)

Algorithm 3 Augmented Lagrangian Algorithm (Sketch)

1: Choose $\nu_0 > 0$, $\{\tau_k\} \to 0$, and λ_0.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Find an approximate solution x_k to

$$\min_x \ \phi(x, \lambda_k; \nu_k) = f(x) + \lambda_k^T c(x) + \frac{\nu_k}{2} \|c(x)\|_2^2$$

satisfying

$$\|\nabla \phi(x_k, \lambda_k; \nu_k)\| \leq \tau_k.$$

4: **end for**
5: If an optimality test for the constrained problem is satisfied, then stop.
6: Update $\lambda_{k+1} \leftarrow \lambda_k + \nu_k c^i(x_k)$.
7: Choose $\nu_{k+1} \geq \nu_k$.

(Similar to quadratic penalty algorithm, except λ_k is updated separately.)

Convergence result

Theorem

Suppose x_* is a solution of the constrained problem at which LICQ and the second-order sufficient conditions are satisfied for λ_*. Then, there exists $\nu_* > 0$ such that for all $\nu > \nu_*$, x_* is a strict local minimizer of $\phi(x, \lambda_*; \nu_*)$.

- Nice! But...
- Unfortunately, this requires knowing the optimal multipliers in advance.

Practical convergence result

Theorem

Suppose x_*, λ_*, and ν_* are defined as in the previous theorem. Then, there exist $\{\epsilon, \delta, M\} \subset \mathbb{R}_{++}$ such that the following statements are true.

(a) *For all λ_k and ν_k with $\|\lambda_k - \lambda_*\| \leq \nu_* \delta \leq \nu_k \delta$, the problem to minimize $\phi(x, \lambda_k; \nu_k)$ over x subject to $\|x - x_*\| \leq \epsilon$ has a unique solution x_k satisfying*

$$\|x_k - x_*\| \leq M \|\lambda_k - \lambda_*\| / \nu_k.$$

(b) *For all λ_k and ν_k with $\|\lambda_k - \lambda_*\| \leq \nu_* \delta \leq \nu_k \delta$, we have*

$$\|\lambda_{k+1} - \lambda_*\| \leq M \|\lambda_k - \lambda_*\| / \nu_k.$$

(c) *For all λ_k and ν_k with $\|\lambda_k - \lambda_*\| \leq \nu_* \delta \leq \nu_k \delta$, the Hessian*

$$\nabla^2_{xx} \phi(x_k, \lambda_k; \nu_k) = \nabla^2_{xx} L(x_k, \lambda_k) + \nu \nabla c(x_k) \nabla c(x_k)^T$$

is positive definite and the constraint gradients are linearly independent.

Drawbacks of augmented Lagrangians

Do we have the same drawbacks as quadratic penalization?

- We still need to define $\{\nu_k\}$, but the choice is less important (assuming $\lambda_k \to \lambda_*$!) as we only need to eventually choose ν_k above a finite threshold.
- We still may have a hard time finding x_k that approximately minimizes the augmented Lagrangian function as $\phi(x, \lambda; \nu)$ may be unbounded below, even if f is convex and bounded below on the feasible set. (None of the methods in these notes are able to avoid this difficulty!)
- Is ill-conditioning a problem? Not as much, again since we only need to eventually choose ν_k above a finite threshold.

Inequality constraints

Corresponding to the generally constrained problem

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & c^{\mathcal{E}}(x) = 0 \\ & c^{\mathcal{I}}(x) \le 0, \end{aligned}$$

there are a variety of ways to create an augmented Lagrangian; e.g., through the addition of slacks, we can assume the only inequalities are bounds on x, meaning that we can have the same augmented Lagrangian

$$\phi(x, \lambda; \nu) := f(x) + \lambda^T c(x) + \frac{\nu}{2}\|c(x)\|_2^2$$

and apply bound constrained instead of unconstrained techniques.

Results similar to those discussed hold in the generally constrained case as well.

The algorithms are more complicated; they used to be popular (e.g., `LANCELOT`), but are now less preferred than SQO and IPM.

Outline

Fundamentals

Quadratic Penalization

Exact Penalization

Augmented Lagrangians

Summary

Penalty and augmented Lagrangian methods

The algorithms discussed in these notes aim to solve constrained problems through the use of unconstrained optimization techniques, where the overall theme is to handle constraints through penalization.

- Quadratic penalization
  - Major drawback: Ill-conditioned subproblems.
- ℓ_1 exact penalization
  - Major drawback: Nonsmooth subproblems.
- Augmented Lagrangians
  - Major drawback: Adverse effects of poor multiplier estimates.

Drawbacks of all of the above is the potential to have subproblems that are unbounded below and the algorithms' sensitivities to choices of ν_k (and λ_k).