A Survey of Generalized Gauss-Newton and Sequential Convex Programming Methods

Moritz Diehl

Systems Control and Optimization Laboratory
Department of Microsystems Engineering and Department of Mathematics
University of Freiburg, Germany

based on joint work with
Florian Messerer
Robin Verschueren (ABB), and
Joris Gillis (KU Leuven)

Benelux Meeting, Lommel, March 20, 2019

Considered problem class

Nonlinear optimization with convex substructure:

$$\begin{aligned} \underset{w \in \mathbb{R}^{n_w}}{\text{minimize}} \quad & \phi_0(F_0(w)) \\ \text{subject to} \quad & F_i(w) \in \Omega_i, \quad i = 1, \ldots, m, \\ & G(w) = 0 \end{aligned}$$

Assumptions:

- twice continuously differentiable functions $G : \mathbb{R}^{n_w} \to \mathbb{R}^{n_g}$ and $F_i : \mathbb{R}^{n_w} \to \mathbb{R}^{n_{F_i}}$ for $i = 0, 1, \ldots, m$.
- outer function $\phi_0 : \mathbb{R}^{n_{F_0}} \to \mathbb{R}$ convex.
- sets $\Omega_i \subset \mathbb{R}^{n_{F_i}}$ convex for $i = 1, \ldots, m$,
  (possibly $\Omega_i = \{z \in \mathbb{R}^{n_{F_i}} | \phi_i(z) \leq 0\}$ with convex ϕ_i)

Considered problem class

Nonlinear optimization with convex substructure:

$$\begin{aligned} \underset{w \in \mathbb{R}^{n_w}}{\text{minimize}} \quad & \phi_0(F_0(w)) \\ \text{subject to} \quad & F_i(w) \in \Omega_i, \quad i = 1, \ldots, m, \\ & G(w) = 0 \end{aligned}$$

Assumptions:

- twice continuously differentiable functions $G : \mathbb{R}^{n_w} \to \mathbb{R}^{n_g}$ and $F_i : \mathbb{R}^{n_w} \to \mathbb{R}^{n_{F_i}}$ for $i = 0, 1, \ldots, m$.
- outer function $\phi_0 : \mathbb{R}^{n_{F_0}} \to \mathbb{R}$ convex.
- sets $\Omega_i \subset \mathbb{R}^{n_{F_i}}$ convex for $i = 1, \ldots, m$,
  (possibly $\Omega_i = \{z \in \mathbb{R}^{n_{F_i}} | \phi_i(z) \leq 0\}$ with convex ϕ_i)

Idea:

exploit convex substructure via *iterative convex approximation algorithms*

Why is this class of problems and algorithms interesting?

- some nonlinear programming (NLP) problems have second order cone or matrix inequality constraints which cannot be treated by standard NLP solvers
- there exist many reliable and efficient convex optimization solvers

Some application areas:

- nonlinear matrix inequalities for reduced order controller design [Fares, Noll, Apkarian 2002; Tran-Dinh, Gumussov, Michiels, Diehl 2012]
- ellipsoidal terminal regions in nonlinear model predictive control [Chen and Allgöwer 1998; Verschueren 2016]
- robustified inequalities in nonlinear optimization [Nagy and Braatz 2003; Diehl, Bock, Kostina 2006]
- tube-following optimal control problems [Van Duikeren, 2019]
- deep neural network training with convex loss functions [Schraudolph 2002; Martens 2016]

First the bad news

- Iterative convex approximation methods such as sequential convex programming (SCP) have **only linear convergence** in general.
- The rate of convergence cannot be improved to superlinear by any bounded semi-definite Hessian modification [Diehl, Jarre, Vogelbusch 2006]

Nasty example problem with dominant nonconvexities in objective:

$$\begin{array}{ll} \underset{w \in \mathbb{R}^2}{\text{minimize}} & -w_1^2 - (1 - w_2)^2 \\ \text{subject to} & \|w\|_2 \leq 1 \end{array}$$

First the bad news

- Iterative convex approximation methods such as sequential convex programming (SCP) have **only linear convergence** in general.
- The rate of convergence cannot be improved to superlinear by any bounded semi-definite Hessian modification [Diehl, Jarre, Vogelbusch 2006]

Nasty example problem with dominant nonconvexities in objective:

$$\begin{aligned} \underset{w \in \mathbb{R}^2}{\text{minimize}} \quad & -w_1^2 - (1-w_2)^2 \\ \text{subject to} \quad & \|w\|_2 \leq 1 \end{aligned}$$

But: many real-world problems have dominant convexities and SCP shows fast linear convergence.

Overview

PART I

- Smooth unconstrained Problems
  - Sequential Convex Programming (SCP)
  - Generalized Gauss-Newton (GGN)
  - Local convergence analysis
  - Desirable divergence
  - Constrained formulation of unconstrained problems
  - Illustrative example
- Constrained problems
  - Sequential Convex Programming (SCP)
  - Constrained Generalized Gauss-Newton (CGGN)
  - Sequential Convex Quadratic Programming (SCQP)
- Demo 1: CasADi for optimization (Joris Gillis)

PART II

- Dynamic Optimization and Applications
- Exercise / Demo 2: Constrained SCP and GGN (Joris Gillis)

Simplest case: smooth unconstrained problems

Unconstrained minimization of "convex over nonlinear" function:

$$\underset{w \in \mathbb{R}^n}{\text{minimize}} \quad \underbrace{\phi(F(w))}_{=:f(w)}$$

Assumptions:
- Inner function $F : \mathbb{R}^n \to \mathbb{R}^N$ of class C^3
- Outer function $\phi : \mathbb{R}^N \to \mathbb{R}$ of class C^3 and convex

Remark:
$F, \phi, f \in C^3$ avoids technical details that would obfuscate main results.

Notation and Preliminaries

Matrix inequality $A \succ B$ for $A, B \in \mathbb{S}^n$ (symmetric matrices)

Gradient $\nabla f(w) \in \mathbb{R}^n$ and Hessian $\nabla^2 f(w) \in \mathbb{S}^n$ for $f : \mathbb{R}^n \to \mathbb{R}$

Jacobian $J(w) := \frac{\partial F}{\partial w}(w) \in \mathbb{R}^{N \times n}$ for $F : \mathbb{R}^n \to \mathbb{R}^N$

Linearization (first order Taylor series) at $\bar{w} \in \mathbb{R}^n$:

$$F_{\text{lin}}(w; \bar{w}) := F(\bar{w}) \; + \; J(\bar{w}) \; (w - \bar{w})$$

Big $O(\cdot)$ notation: e.g. $F_{\text{lin}}(w; \bar{w}) = F(w) + O(\|w - \bar{w}\|^2)$

Note: gradient of $f(w) = \phi(F(w))$ is $\nabla f(w) = J(w)^\top \nabla \phi(F(w))$

Method 1: Sequential Convex Programming (SCP)

Starting at $w_0 \in \mathbb{R}^n$ we generate sequence $\ldots, w_k, w_{k+1}, \ldots$

by obtaining w_{k+1} as solution of convex subproblem at $\bar{w} = w_k$:

$$\underset{w \in \mathbb{R}^n}{\text{minimize}} \quad \underbrace{\phi\left(F_{\text{lin}}(w; \bar{w})\right)}_{=: f_{\text{SCP}}(w; \bar{w})} \tag{1}$$

(requires possibly expensive calls of a convex optimization solver)

Lemma 1: $f_{\text{SCP}}(w; \bar{w}) = f(w) + O(\|w - \bar{w}\|^2)$

Proof: $\nabla_w f_{\text{SCP}}(\bar{w}; \bar{w}) = J(\bar{w})^\top \nabla\phi(F(\bar{w})) = \nabla f(\bar{w})$ $\square$

Corollary 2: $\nabla f(\bar{w}) = 0 \quad \Leftrightarrow \quad \bar{w}$ fixed point of SCP
(if subproblems (1) have unique solutions)

Tutorial Example

Regard $F(w) = \begin{bmatrix} \sin(w) \\ \exp(w) - 2 \end{bmatrix}$ and $\phi(z) = z_1^4 + z_2^2$

Linearization: $F_{\text{lin}}(w; \bar{w}) = \begin{bmatrix} \sin(\bar{w}) + \cos(\bar{w})(w - \bar{w}) \\ \exp(\bar{w}) - 2 + \exp(\bar{w})(w - \bar{w}) \end{bmatrix}$

SCP objective:

$$f_{\text{SCP}}(w; \bar{w}) = (\sin(\bar{w}) + \cos(\bar{w})(w - \bar{w}))^4 + (\exp(\bar{w}) - 2 + \exp(\bar{w})(w - \bar{w}))^2$$

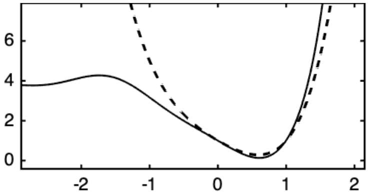


Fig.: $f(w) = \phi(F(w))$ (solid) and $f_{\text{SCP}}(w; \bar{w})$ (dashed) for $\bar{w} = 0$

SCP for Least Squares Problems = Gauss-Newton

With quadratic $\phi(z) = \frac{1}{2}\|z\|_2^2 = \frac{1}{2}z^\top z$, SCP subproblems become

$$\underset{w \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|F(w_k) + J(w_k)(w - w_k)\|_2^2 \tag{2}$$

If $\mathrm{rank}(J) = n$ this is uniquely solvable, giving

$$w_{k+1} = w_k - \Big(\underbrace{J(w_k)^\top J(w_k)}_{=:B_{\mathrm{GN}}(w_k)}\Big)^{-1} \underbrace{J(w_k)^\top F(w_k)}_{=\nabla f(w_k)}$$

SCP=Newton-type method with "Gauss-Newton Hessian approximation"

$$B_{\mathrm{GN}}(w) \approx \nabla^2 f(w)$$

Method 2: Generalized Gauss-Newton

For general convex $\phi(z)$ we have

$$\nabla^2 f(w) = \underbrace{J(w)^\top \, \nabla^2 \phi(F(w)) \, J(w)}_{=: B_{\mathrm{GGN}}(w)} + \underbrace{\textstyle\sum_{j=1}^N \nabla^2 F_j(w) \, \nabla_{z_j} \phi(F(w))}_{=: E_{\mathrm{GGN}}(w)}$$

"GGN Hessian" "Error matrix"

Method 2: Generalized Gauss-Newton

For general convex $\phi(z)$ we have

$$\nabla^2 f(w) = \underbrace{J(w)^\top \nabla^2 \phi(F(w))\, J(w)}_{=: B_{\mathrm{GGN}}(w)} + \underbrace{\textstyle\sum_{j=1}^N \nabla^2 F_j(w)\, \nabla_{z_j} \phi(F(w))}_{=: E_{\mathrm{GGN}}(w)}$$

"GGN Hessian" "Error matrix"

Generalized Gauss-Newton (GGN) method iterates according to

$$w_{k+1} = w_k - B_{\mathrm{GGN}}(w)^{-1} \nabla f(w_k)$$

Method 2: Generalized Gauss-Newton

For general convex $\phi(z)$ we have

$$\nabla^2 f(w) = \underbrace{J(w)^\top \nabla^2\phi(F(w))\, J(w)}_{=:B_{\text{GGN}}(w)} + \underbrace{\sum_{j=1}^N \nabla^2 F_j(w)\, \nabla_{z_j}\phi(F(w))}_{=:E_{\text{GGN}}(w)}$$

"GGN Hessian" "Error matrix"

Generalized Gauss-Newton (GGN) method iterates according to

$$w_{k+1} = w_k - B_{\text{GGN}}(w)^{-1}\nabla f(w_k)$$

Note: GGN solves convex quadratic subproblems

$$\min_{w\in\mathbb{R}^n} \underbrace{f(w_k)+\nabla f(w_k)^\top(w-w_k)+\frac{1}{2}(w-w_k)^\top B_{\text{GGN}}(w_k)(w-w_k)}_{=:f_{\text{GGN}}(w;w_k)}$$

Tutorial Example: GGN and SCP

Regard again $F(w) = \begin{bmatrix} \sin(w) \\ \exp(w) - 2 \end{bmatrix}$ and $\phi(z) = z_1^4 + z_2^2$

Jacobian:

$$J(\bar{w}) = \begin{bmatrix} \cos(\bar{w}) \\ \exp(\bar{w}) \end{bmatrix}$$

GGN objective at $\bar{w} = 0$:

$$f_{\mathrm{GGN}}(w; \bar{w}) = (\exp(\bar{w}) - 2 + \exp(\bar{w})(w - \bar{w}))^2$$

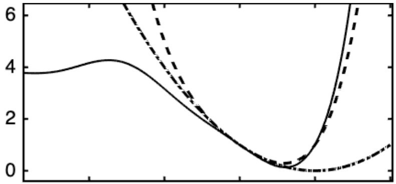


Figure shows f (solid), f_{SCP} (dashed) and f_{GGN} (dash dotted)

Differences and Similarities of SCP and GGN

- both SCP and GGN generalize the Gauss-Newton method (and become equal to it when applied to nonlinear least squares)
- both iteratively solve convex optimization problems
- GGN only solves a positive definite linear system per iteration
- SCP solves a full convex problem per iteration (5x-30x slower)
- SCP often less sensitive to initialization
- both are NOT second order methods
- both converge linearly *with the same rate* (details follow)

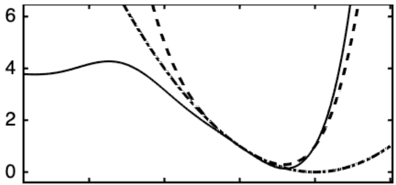
Local Convergence Analysis for SCP and GGN

Recall:
$f_{\mathrm{SCP}}(w;\bar{w}) = \phi\left(F_{\mathrm{lin}}(w;\bar{w})\right)$
$f_{\mathrm{GGN}}(w;\bar{w}) = f(\bar{w}) + \nabla f(\bar{w})^\top (w-\bar{w}) + \frac{1}{2}(w-\bar{w})^\top B_{\mathrm{GGN}}(\bar{w})(w-\bar{w})$
$\nabla_w f_{\mathrm{SCP}}(\bar{w};\bar{w}) = \nabla_w f_{\mathrm{GGN}}(\bar{w};\bar{w}) = \nabla f(\bar{w})$

Lemma 3 (Equality of SCP and GGN up to Second Order)

$f_{\mathrm{SCP}}(w;\bar{w}) = f_{\mathrm{GGN}}(w;\bar{w}) + O(\|w-\bar{w}\|^3)$

Proof: $\nabla_w^2 f_{\mathrm{SCP}}(\bar{w};\bar{w}) = B_{\mathrm{GGN}}(\bar{w}) = \nabla_w^2 f_{\mathrm{GGN}}(\bar{w};\bar{w})$ □



Note: in general $B_{\mathrm{GGN}}(\bar{w}) \neq \nabla^2 f(\bar{w})$

Local Convergence of SCP and GGN

Main Theorem 1 [Diehl and Messerer, submitted to CDC 2019]

Regard w^* with $\nabla f(w^*) = 0$ and $B_{\mathrm{GGN}}(w^*) \succ 0$. Then

- w^* is a fixed point for both the SCP and GGN iterations
- both methods are well-defined in a neighborhood of w^*
- their linear contraction rates are equal and given by

$$\min\{\alpha \in \mathbb{R} \mid -\alpha B_{\mathrm{GGN}}(w^*) \preceq E_{\mathrm{GGN}}(w^*) \preceq \alpha B_{\mathrm{GGN}}(w^*)\}$$

Local Convergence of SCP and GGN

Main Theorem 1 [Diehl and Messerer, submitted to CDC 2019]

Regard w^* with $\nabla f(w^*) = 0$ and $B_{\mathrm{GGN}}(w^*) \succ 0$. Then

- w^* is a fixed point for both the SCP and GGN iterations
- both methods are well-defined in a neighborhood of w^*
- their linear contraction rates are equal and given by

$$\min\{\alpha \in \mathbb{R} \mid -\alpha B_{\mathrm{GGN}}(w^*) \preceq E_{\mathrm{GGN}}(w^*) \preceq \alpha B_{\mathrm{GGN}}(w^*)\}$$

Corollary: Necessary condition for local convergence of both methods is

$$B_{\mathrm{GGN}}(w^*) \succeq \frac{1}{2}\nabla^2 f(w^*)$$

Local Convergence of SCP and GGN

Main Theorem 1 [Diehl and Messerer, submitted to CDC 2019]

Regard w^* with $\nabla f(w^*) = 0$ and $B_{\mathrm{GGN}}(w^*) \succ 0$. Then

- w^* is a fixed point for both the SCP and GGN iterations
- both methods are well-defined in a neighborhood of w^*
- their linear contraction rates are equal and given by

$$\min\{\alpha \in \mathbb{R} \mid -\alpha B_{\mathrm{GGN}}(w^*) \preceq E_{\mathrm{GGN}}(w^*) \preceq \alpha B_{\mathrm{GGN}}(w^*)\}$$

Corollary: Necessary condition for local convergence of both methods is

$$B_{\mathrm{GGN}}(w^*) \succeq \frac{1}{2}\nabla^2 f(w^*)$$

Proof of corollary: Local convergence requires $\alpha \leq 1$ and
$E_{\mathrm{GGN}}(w^*) = \nabla_w^2 f(w^*) - B_{\mathrm{GGN}}(w^*) \preceq \alpha B_{\mathrm{GGN}}(w^*)$ □

Proof of Main Theorem

Define solution operators $w^{\mathrm{sol}}_{\mathrm{SCP}}(\bar{w})$ and $w^{\mathrm{sol}}_{\mathrm{GGN}}(\bar{w})$ at linearization point $\bar{w}$ and apply the implicit function theorem to

$$\nabla_w f_i(w_i^{\mathrm{sol}}(\bar{w}); \bar{w}) = 0 \quad \text{for} \quad i = \mathrm{SCP}, \mathrm{GGN}$$

Well-defined for $\bar{w}$ in neighborhood of w^* because $\nabla_w^2 f_{\mathrm{SCP}}(w^*; w^*) = \nabla_w^2 f_{\mathrm{GGN}}(w^*; w^*) = B_{\mathrm{GGN}}(w^*) \succ 0$, and derivatives are given by

$$\frac{\mathrm{d}w_i^{\mathrm{sol}}}{\mathrm{d}\bar{w}}(w^*) = -(\underbrace{\nabla_w \nabla_w f_i(w^*; w^*)}_{=B_{\mathrm{GGN}}(w^*)=:B_*})^{-1} \underbrace{\nabla_{\bar{w}} \nabla_w f_i(w^*; w^*)}_{=E_{\mathrm{GGN}}(w^*)=:E_*}$$

We used that all second derivatives are equal due to Lemma 3 and that

$$\begin{aligned} \nabla_{\bar{w}} \nabla_w f_{\mathrm{GGN}}(w^*; w^*) &= \nabla_{\bar{w}} \left(\nabla f(\bar{w}) + B_{\mathrm{GGN}}(\bar{w})(w - \bar{w})\right)\big|_{w=\bar{w}=w^*} \\ &= \nabla^2 f(w^*) - B_{\mathrm{GGN}}(w^*) = E_{\mathrm{GGN}}(w^*) \end{aligned}$$

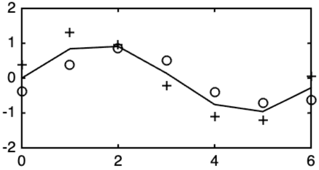
Proof of Main Theorem (continued)

Spectral radius $\rho(-B_*^{-1}E_*) = \rho(B_*^{-1}E_*)$ equals linear contraction rate of SCP and GGN algorithms. Matrix can be transformed to similar, but symmetric matrix: $B_*^{-1}E_* \sim B_*^{-\frac{1}{2}}\, E_*\, B_*^{-\frac{1}{2}}$ with same spectral radius.

Now, $$\rho(B_*^{-\frac{1}{2}}\, E_*\, B_*^{-\frac{1}{2}}) \leq \alpha \quad \Leftrightarrow \quad -\alpha I \preceq B_*^{-\frac{1}{2}}\, E_*\, B_*^{-\frac{1}{2}} \preceq \alpha I$$
$$\Leftrightarrow \quad -\alpha B_* \preceq E_* \preceq \alpha B_* \qquad \square$$

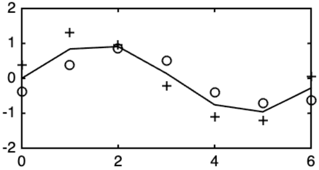
Desirable divergence and mirror problem

SCP and GGN do not converge to every local minimum. This can help to avoid "bad" local minima, as discussed next.



Regard maximum likelihood estimation problem $\min_w \phi(M(w) - y)$ with nonlinear model $M : \mathbb{R}^n \to \mathbb{R}^N$ and measurements $y \in \mathbb{R}^N$. Assume penalty ϕ is symmetric with $\phi(-z) = \phi(z)$ as is the case for symmetric error distributions. At a solution w^*, we can generate "mirror measurements" $y_{\mathrm{mr}} := 2M(w^*) - y$ obtained by reflecting the residuals. From a statistical point of view, y_{mr} should be as likely as y.

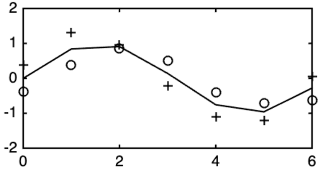
SCP Divergence $\Leftrightarrow$ Minimum unstable under mirroring



Theorem 2 [Diehl and Messerer, submitted] inspired by [Bock 1987]

Regard a local minimizer w^* of $f(w) = \phi(M(w) - y)$ with $\nabla^2 f(w^*) \succ 0$. If the necessary SCP-GGN-convergence condition $B_{\mathrm{GGN}}(w^*) \succeq \frac{1}{2}\nabla^2 f(w^*)$ does not hold, then w^* is a stationary point, but **not** a local minimum for $f_{\mathrm{mr}}(w) := \phi(M(w) - y_{\mathrm{mr}})$.

SCP Divergence $\Leftrightarrow$ Minimum unstable under mirroring



Theorem 2 [Diehl and Messerer, submitted] inspired by [Bock 1987]

Regard a local minimizer w^* of $f(w) = \phi(M(w) - y)$ with $\nabla^2 f(w^*) \succ 0$. If the necessary SCP-GGN-convergence condition $B_{\mathrm{GGN}}(w^*) \succeq \frac{1}{2}\nabla^2 f(w^*)$ does not hold, then w^* is a stationary point, but **not** a local minimum for $f_{\mathrm{mr}}(w) := \phi(M(w) - y_{\mathrm{mr}})$.

Sketch of proof: use $M(w^*) - y_{\mathrm{mr}} = y - M(w^*)$ to show that
$\nabla f_{\mathrm{mr}}(w^*) = J(w^*)^\top (y - M(w^*)) = 0$ and
$\nabla^2 f_{\mathrm{mr}}(w^*) = B_{\mathrm{GGN}}(w^*) - E_{\mathrm{GGN}}(w^*) = 2B_{\mathrm{GGN}}(w^*) - \nabla^2 f(w^*) \not\succeq 0$

Constrained formulation of unconstrained SCP and GGN

We can express $\min_w \phi(F(w)$ as constrained NLP with slacks $z \in \mathbb{R}^N$:

$$\boxed{\min_{w,z} \quad \phi(z) \quad \text{s.t.} \quad z - F(w) = 0}$$

SCP subproblem then becomes

$$\boxed{\min_{w,z} \quad \phi(z) \quad \text{s.t.} \quad z - F_{\text{lin}}(w;\bar{w}) = 0}$$

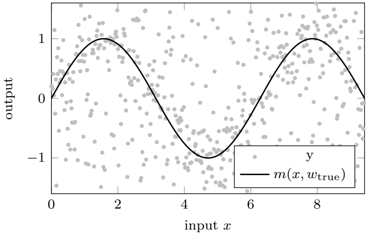
with $\bar{w}$ being the only iteration memory. If initialized at the same w_0, unconstrained and constrained SCP deliver identical iterates $w_1, w_2, \ldots$

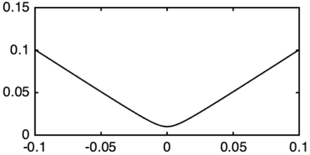
On the other hand, the GGN subproblem becomes

$$\boxed{\min_{w,z} \phi_{\text{GGN}}(z;\bar{z}) \quad \text{s.t.} \quad z - F_{\text{lin}}(w;\bar{w}) = 0}$$

with $\phi_{\text{GGN}}(z;\bar{z}) := \phi(\bar{z}) + \nabla\phi(\bar{z})^\top(z-\bar{z}) + \frac{1}{2}(z-\bar{z})^\top \nabla^2\phi(\bar{z})(z-\bar{z})$
Here, both $\bar{w}$ and $\bar{z}$ form the iteration memory, and unconstrained and constrained GGN methods perform different iterates in general.

Illustrative example (implemented by Florian Messerer)





Regard scalar model

$$m(x, w) := \sin(wx)$$

with input output measurements (x_i, y_i) for $i = 1, \ldots, N$ (left plot).
Define $M_i(w) := m(x_i, w)$ and $F(w) := M(w) - y$
As outer convexity we use a Huber-like penalty (right plot)

$$\phi(z) := \frac{1}{N} \sum_{i=1}^{N} \sqrt{\delta^2 + z_i^2}, \tag{3}$$

Implementation of SCP and GGN

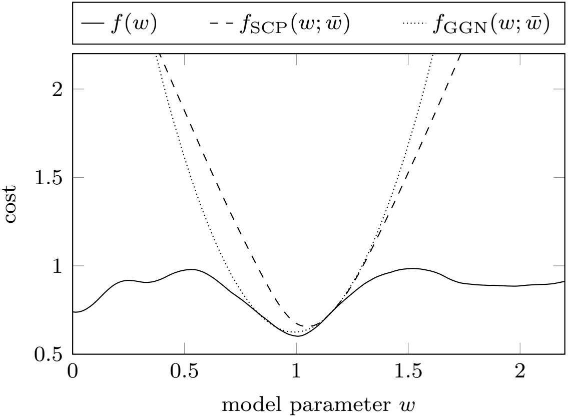
- obtain all needed derivatives from CasADi via MATLAB interface
- use linear solve ("backslash") for GGN
- formulate SCP subproblem as second order cone program (SOCP):

$$\min_{\substack{w \in \mathbb{R}, \\ s \in \mathbb{R}^N}} \quad \sum_{i=1}^{N} s_i$$

$$\text{s.t.} \quad \sqrt{\delta^2 + F_{\mathrm{lin},i}(w; w_k)^2} \le s_i \quad \text{for } i = 1, \dots, N$$

- use Gurobi via CasADi as SOCP solver

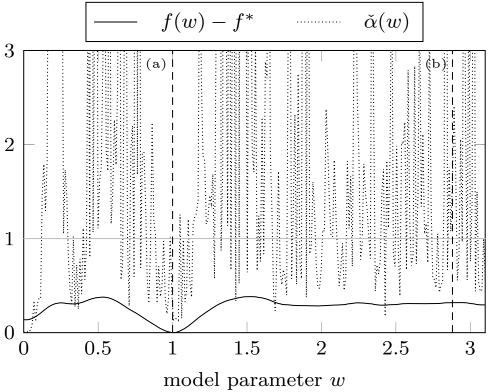
Visualization of SCP and GGN Subproblems



$(\bar{w} = 1.15)$

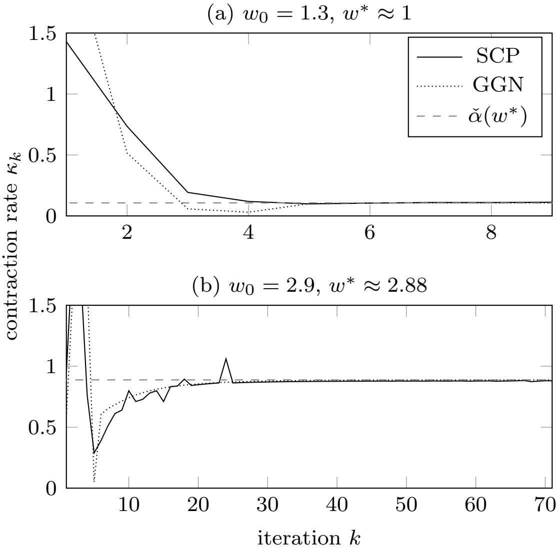
Objective and Local Contraction Rate

$f(w)$ and local rate $\check{\alpha}(w) = \frac{|\nabla^2 f(w) - B_{\mathrm{GGN}}(w)|}{|B_{\mathrm{GGN}}(w)|}$



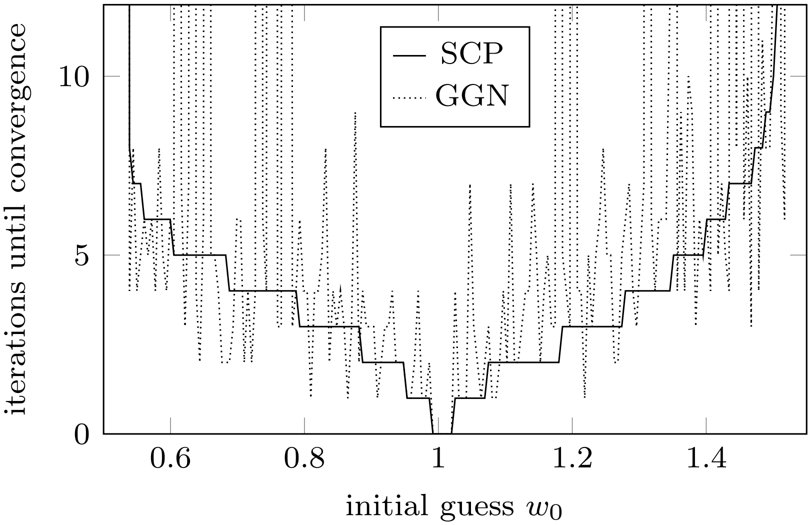
Fast contraction rate at global minimum $w^* \approx 1$, otherwise slower.

Empirical Contraction Rates in Agreement with Theorem 1



$$\kappa_k = \frac{|w_{k+1} - w_k|}{|w_k - w_{k-1}|}$$

Iteration count: SCP more predictable than GGN



Overview

PART I

- Smooth unconstrained Problems
  - Sequential Convex Programming (SCP)
  - Generalized Gauss-Newton (GGN)
  - Local convergence analysis
  - Desirable divergence
  - Constrained formulation of unconstrained problems
  - Illustrative example
- **Constrained problems**
  - Sequential Convex Programming (SCP)
  - Constrained Generalized Gauss-Newton (CGGN)
  - Sequential Convex Quadratic Programming (SCQP)
- Demo 1: CasADi for optimization (Joris Gillis)

PART II

- Dynamic Optimization and Applications
- Exercise / Demo 2: Constrained SCP and GGN (Joris Gillis)

Constrained Problems

Recall:

$$\begin{aligned} \underset{w \in \mathbb{R}^{n_w}}{\text{minimize}} \quad & \phi_0(F_0(w)) \\ \text{subject to} \quad & F_i(w) \in \Omega_i, \quad i = 1, \ldots, m, \\ & G(w) = 0 \end{aligned}$$

class C^2 functions $G : \mathbb{R}^{n_w} \to \mathbb{R}^{n_g}$ and $F_i : \mathbb{R}^{n_w} \to \mathbb{R}^{n_{F_i}}$ for $i = 0, 1, \ldots, m$.
outer function $\phi_0 : \mathbb{R}^{n_{F_0}} \to \mathbb{R}$ convex.
sets $\Omega_i \subset \mathbb{R}^{n_{F_i}}$ convex for $i = 1, \ldots, m$,
(possibly $\Omega_i = \{z \in \mathbb{R}^{n_{F_i}} | \phi_i(z) \le 0\}$ with convex ϕ_i)

SCP generalizes canonically...

SCP Subproblem

$$\begin{aligned} \underset{w \in \mathbb{R}^{n_w}}{\text{minimize}} \quad & \phi_0(F_0^{\text{lin}}(w;\bar{w})) \\ \text{subject to} \quad & F_i^{\text{lin}}(w;\bar{w}) \in \Omega_i, \quad i = 1,\dots,m, \\ & G^{\text{lin}}(w;\bar{w}) = 0 \end{aligned}$$

- obtain w_{k+1} as solution of convex problem at $\bar{w} = w_k$
- only primal variables $\bar{w}$ form "optimizer state", no multipliers needed
- SCP is an affine invariant method (next slide)

Affine Invariance of SCP

Regard affine input transformation $w = b + Bv$ and linear transformation of equality constraint with invertible matrices A and B, to yield affinely transformed problem:

$$\begin{aligned} \underset{v \in \mathbb{R}^{n_w}}{\text{minimize}} \quad & \phi_0(F_0(b + Bv)) \\ \text{subject to} \quad & F_i(b + Bv) \in \Omega_i, \quad i = 1, \ldots, m, \\ & AG(b + Bv) = 0 \end{aligned}$$

One can show that SCP applied to affinely transformed problem and initialized at v_0 performs identical iterates to SCP applied to the original problem initialized at $w_0 = b + Bv_0$, i.e., $v_k = b + Bw_k$ for $k = 1, 2, \ldots$

Smooth NLP formulation

For constrained GGN type algorithms, we need to assume more smoothness than for SCP, and regard:

$$\begin{aligned} \underset{w \in \mathbb{R}^{n_w}}{\text{minimize}} \quad & \underbrace{\phi_0(F_0(w))}_{=:f_0(w)} \\ \text{subject to} \quad & \underbrace{\phi_i(F_i(w))}_{=:f_i(w)} \leq 0, \quad i = 1, \ldots, m, \\ & G(w) = 0 \end{aligned}$$

with convex C^2 functions $\phi_0, \phi_1, \ldots, \phi_m$.

Constrained Generalized Gauss-Newton (CGGN)

Use $B_{\text{CGGN}}(w) := J_0(w)^\top \nabla^2 \phi_0(F_0(w)) J_0(w)$, but now solve

$$\begin{aligned} \underset{w \in \mathbb{R}^{n_w}}{\text{minimize}} \quad & f_0^{\text{lin}}(w; \bar{w}) + \frac{1}{2}(w - \bar{w})^\top B_{\text{CGGN}}(\bar{w})(w - \bar{w}) \\ \text{subject to} \quad & f_i^{\text{lin}}(w; \bar{w}) \leq 0, \quad i = 1, \dots, m, \\ & G^{\text{lin}}(w; \bar{w}) = 0 \end{aligned}$$

- again, the method is multiplier free
- method is also affine invariant
- each iteration requires solution of a convex quadratic program (QP)
- convexity in constraints does not enter CGGN subproblem

Remark: for least-squares objectives, this method is due to [Bock 1987]. In mathematical optimization papers, Bock's method is called "the Generalized Gauss-Newton (GGN) method" because it generalizes the GN method to constrained problems. In order to avoid a notation clash with computer science we prefer to call Bock's method "the constrained Gauss-Newton method".

Sequential Convex Quadratic Programming (SCQP)

Following [Verschueren et al., 2016], use

$$B_{\mathrm{SCQP}}(w,\mu) := J_0(w)^\top \nabla^2 \phi_0(F_0(w)) J_0(w) + \sum_{i=1}^{m} \mu_i J_i(w)^\top \nabla^2 \phi_i(F_i(w)) J_i(w)$$

and solve:

$$\begin{aligned} \underset{w \in \mathbb{R}^{n_w}}{\text{minimize}} \quad & f_0^{\mathrm{lin}}(w;\bar{w}) + \frac{1}{2}(w-\bar{w})^\top B_{\mathrm{SCQP}}(\bar{w},\mu)(w-\bar{w}) \\ \text{subject to} \quad & f_i^{\mathrm{lin}}(w;\bar{w}) \leq 0, \quad i = 1,\ldots,m, \\ & G^{\mathrm{lin}}(w;\bar{w}) = 0 \end{aligned}$$

- "optimizer state" contains $\bar{w}$ but also inequality multipliers $\mu \geq 0$
- affine invariant
- SCQP has same contraction rate as SCP, but only solves QP
- $B_{\mathrm{SCQP}}(w,\mu) \succeq B_{\mathrm{CGGN}}(w)$

Some References

- J. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl. CasADi: a software framework for nonlinear optimization and optimal control. Mathematical Programming Computation, 2018.
- H.G. Bock, E. Kostina, and J.P. Schlöder. Numerical methods for parameter estimation in nonlinear differential algebraic equations. GAMM Mitteilungen, 30/2:376408, 2007.
- J. Martens. Second-Order Optimization for Neural Networks. PhD thesis, University of Toronto, 2016.
- N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. Neural Computation, 14(7):1723–1738, 2002.
- Q. Tran-Dinh. Sequential Convex Programming and Decomposition Approaches for Nonlinear Optimization. PhD thesis, KU Leuven, 2012.
- R. Verschueren, N. van Duijkeren, R. Quirynen, and M. Diehl. Exploiting convexity in direct optimal control: a sequential convex quadratic programming method. In Proceedings of the IEEE Conference on Decision and Control (CDC), 2016.