A NONLINEAR REGRESSION PERSPECTIVE ON A PRIMAL-DUAL AUGMENTED LAGRANGIAN

Southern California Optimization Day
May 23, 2014

Josh Griffin
Wenwen Zhou
1  Background

2  PDAL Merit Function

3  Constant Objective Interior

4  Feasibility Control

5  Numerical Results

6  Conclusions
Motivate addition of a primal-proximity term to a primal-dual augmented Lagrangian merit function

- Forsgren, Gill (1998)
- Gill, Robinson (2010)

Proximity term similar to Friedlander, Orban 2012

We’ll show

- the proximity term restores primary purpose of penalty term
- search directions have strong correspondence to standard nonlinear regression approaches
- improved performance for infeasible problems
Two primal-dual merit based solvers in PROC OPTMODEL:

1. Interior-point
2. Active-set

for nonlinear (possibly nonconvex) optimization problems:

**NLP (Nonlinear Programming Problem)**

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad c(x) = 0 \\
& \quad x \geq 0.
\end{align*}
\]

- \(c(x) \in \mathbb{R}^m\)
- \(f(x), c(x)\) are twice continuously differentiable
Gradient of objective: \( g = g(x) = \nabla f(x) \)

Jacobian of constraints: \( J = J(x) = c'(x) \)

Lagrangian: \( \mathcal{L}(x, y) = f(x) - c(x)^T y \)

Hessian of Lagrangian: \( H = \nabla^2_{xx} \mathcal{L}(x, y) \)

Augmented Lagrangian:

\[
\mathcal{P}(x; y_e, \mu) = f(x) - y_e^T c(x) + \frac{1}{2\mu} \|c(x)\|^2
\]

Augmented Lagrangian Gradient:

\[
\nabla_x \mathcal{P}(x; y_e, \mu) = g - J^T(y_e - c(x)/\mu)
\]

Primal multipliers: \( \pi = y_e - c(x)/\mu \)
Classical augmented Lagrangian merit function:

$$\mathcal{P}(x; y_e, \mu) = f(x) - y_e^T c(x) + \frac{1}{2\mu} \|c(x)\|^2$$

Both solvers use FGR (Forsgren, Gill, Robinson) merit function:

$$M(x, y; y_e, \mu) = \mathcal{P}(x; y_e, \mu) + \frac{1}{2\mu} \|c(x) + \mu(y - y_e)\|^2$$

Simplifies to sequence of bound constrained subproblems

**Bound-constrained subproblem (y_e, \mu fixed)**

minimize $M(x, y)$

subject to $x \geq 0$. 
Approximate Newton’s system for $\nabla^2 M \Delta \nu = -\nabla M$:

$$
\begin{pmatrix}
H(x, y) + \frac{1}{2\mu} J^T J & J^T \\
J & \mu I
\end{pmatrix}
\begin{pmatrix}
\nu_x \\
\nu_y
\end{pmatrix}
= -
\begin{pmatrix}
g - J^T (2\pi - y) \\
\mu (y - \gamma e)
\end{pmatrix}

\approx \nabla^2 M

$$

Sparse equivalent formulation:

$$
\begin{pmatrix}
H(x, y) & J^T \\
J & -\mu I
\end{pmatrix}
\begin{pmatrix}
\nu_x \\
\nu_y
\end{pmatrix}
= -
\begin{pmatrix}
g - J^T y \\
c(x) + \mu (y - \gamma e)
\end{pmatrix}

$$

Compare to classical equations

$$
\begin{pmatrix}
H(x, y) & J^T \\
J & 0
\end{pmatrix}
\begin{pmatrix}
\hat{\nu}_x \\
\hat{\nu}_y
\end{pmatrix}
= -
\begin{pmatrix}
g - J^T y \\
c(x)
\end{pmatrix}
$$
Ill-conditioned QP

minimize \( (v - v_k)^T \nabla M + \frac{1}{2} (v - v_k)^T B (v - v_k) \)
subject to \( x \geq 0, v = (x, y) \).

Dual regularized QP (Gill, Kungurtsev, Robinson 2013)

minimize \( g^T (x - x_k) + \frac{1}{2} (x - x_k)^T H (x - x_k) + \frac{1}{2} \mu \| y \|_2^2 \)
subject to \( c + J (x - x_k) + \mu (y - y^k_{\theta}) = 0, x \geq 0 \).
Trust-region subproblem

\[
\begin{align*}
\text{minimize} & \quad (v - v_k)^T \nabla M + \frac{1}{2} (v - v_k)^T B (v - v_k) \\
\text{subject to} & \quad \|v\| \leq \delta, x \geq 0
\end{align*}
\]

- We apply an SSM that extends Steihaug-Toint
- Constraint Preconditioner handles inherent ill-conditioning

\[
P_K = \begin{pmatrix} I & J^T \\ J & -\mu \end{pmatrix}
\]
equivalently

\[
P_B = \begin{pmatrix} I + \frac{1}{2\mu} J^T J & J^T \\ J & \mu \end{pmatrix}
\]

- Interior uses Forsgren, Gill (1998) for inequalities
- B can be indefinite
Strengths

- Primal and dual variables treated nearly identically
- Regularized subproblem
- Potentially locally quadratic convergence rate
- If $y_e \rightarrow y^*$, $\mu$ need not converge to 0
- Preconditioning optional when $\mu$ is large
- Natural constraint preconditioner available

Challenges (modifications/safe-guards needed)

- No longer constraint scale invariant
- Less aggressive at reducing constraint violation
- Intermediate values of $y$, $y_e$ grow quickly towards bounds
- $\mu$ often much smaller than classical approaches
Preference for minimal algorithmic changes
Improve constraint handling
Secondary purpose of $\mu$ is regularization
Primary purpose of $\mu$ is counter balance to objective
  - Can $\mu$ remain constant if objective is constant?
  - Can $\mu$ remain constant if approaching vertex solution?

$y_e \rightarrow y^*$ no longer critical for performance
Can $y_e, y$ remain bounded for infeasible problems?

\[
\begin{align*}
y_e^{k+1} & \rightarrow y^k \rightarrow y_e^k - \frac{c(x_k)}{\mu_k} \\
\Rightarrow \quad \|y_e^k\|, \|y^k\| & \rightarrow \infty
\end{align*}
\]

if infeasible and $\mu_k \rightarrow 0$
minimize \quad -10^5 x

subject to \quad 10^{-5} x = 0.

Assume \( y_e = 0 \), then

\[ M(x, y) = -10^5 x + \frac{1}{2\mu} \left( (10^{-5} x)^2 + (10^{-5} x + \mu y)^2 \right) \]

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>( x(\mu) )</th>
<th>( c(x(\mu)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10^{15}</td>
<td>10^{10}</td>
</tr>
<tr>
<td>10^{-6}</td>
<td>10^{9}</td>
<td>10^{4}</td>
</tr>
<tr>
<td>10^{-16}</td>
<td>.1</td>
<td>10^{-6}</td>
</tr>
</tbody>
</table>

Linearized constraint approach of course solves in 1 step.
Let $J = c'(x)$ and assume full row rank.

**Newton’s method on $c(x) = 0$**

while not converged do:

1. Find $s$ such that $J(x)s = -c(x)$
2. Perform line-search on $\|c(x + \alpha s)\|_2^2$

Could choose min-$M$ norm:

$$\minimize_{s \in \mathbb{R}^n} \frac{1}{2} \|s\|_M^2$$

subject to $Js + c = 0$. 

Copyright 2014, SAS Institute Inc. All rights reserved.
minimize \( s_x \in \mathbb{R}^n \) \( \|s_x\|^2_M \)
subject to \( Js_x + c = 0 \).

Can be found as solution to

\[
\begin{pmatrix}
M & J^T \\
J & 0
\end{pmatrix}
\begin{pmatrix}
s_x \\
-s_y
\end{pmatrix}
= -
\begin{pmatrix}
J^T y \\
c
\end{pmatrix}
\]

If \( M \) denotes positive-definite approximation to \( H \):

- Note, if \( M = I \), \( s_x = J^c = -J^T(JJ^T)^{-1}c \)
- Addition of objective simply select different \( s_x \) sequence
- classic KKT equations for NLP
- Newton’s method on \( c(x) = 0 \) always in background
Let $J = c'(x)$.

**Levenberg-Marquardt on** $c(x) = 0$

while not converged do:

1. Solve $(\sigma I + J^T J)s = -J^T c$
2. Perform line-search on $\|c(x + \alpha s)\|_2^2$

Can show $s$ is solution to:

$$\min_{s \in \mathbb{R}^n} \frac{1}{2} \left( \sigma \|s\|_2^2 + \|Js + c\|_2^2 \right)$$

Typically LM assumes $\sum_{i=1} \nabla^2 c_i(x) \to 0$
CONSTANT OBJECTIVE INTERIOR

SPARSE EQUATIONS

\[
\text{minimize } \quad \frac{1}{2} \left( \sigma \| s \|_2^2 + \| Js + c \|_2^2 \right)
\]

Can be found as solution to sparse system

\[
\begin{pmatrix}
\lambda I & J^T \\
J & -\mu I
\end{pmatrix}
\begin{pmatrix}
s_x \\
-s_y
\end{pmatrix}
= -\begin{pmatrix}
J^Ty \\
c + \mu y
\end{pmatrix}
\]

where

- \( \sigma = \lambda \mu \)
- \( y \) can be anything
- As \( \sigma \to 0 \)
  - full row rank: \( s_x \to -J^T(JJ^T)^{-1}c \) (min two-norm)
  - full col rank: \( s_x \to -(J^TJ)^{-1}J^Tc \) (least-squares)
Regularized Newton-systems have the form:

\[
\begin{pmatrix}
H(x, y) & J^T \\
J & -\mu I
\end{pmatrix}
\begin{pmatrix}
s_x \\
-s_y
\end{pmatrix}
=-
\begin{pmatrix}
J^Ty \\
c + \mu y
\end{pmatrix}
\]

where

- \( H(x, y) = -\sum_{i=1}^{m} y_i \nabla^2 c_i(x) \)
- \( \lambda I \) missing (sometimes added as part of trust-region solver)
- Intermediate \( y \) can grow large
- Negligible second-order term from LM starts to dominate
Regularized Newton-systems have the form:

\[
\begin{pmatrix}
H(x, \gamma y) + \lambda I & J^T \\
J & -\mu I
\end{pmatrix}
\begin{pmatrix}
s_x \\
-s_y
\end{pmatrix}
= -
\begin{pmatrix}
J^T y \\
c + \mu y
\end{pmatrix}
\]

If \( y \) converges to \( \pi = -c/\mu \) then

\[
(\lambda \mu I + J^T J + \gamma \sum_{i=1}^{m} c_i \nabla^2 c_i) s_x = -J^T c
\]

**Results:**

1. If \( \gamma = 0 \) is Levenberg-Marquardt
2. If \( \gamma = 1 \) is regularized Newton on \( r(x) = \|c(x)\|^2_2 \)
3. Send \( \lambda \to 0 \) not \( \mu \).
Transformation steps:
- Scale $\mathcal{M}(x, y, y_e, \mu)$ by $\mu$
- Redefine $y = \mu y$, $y_e = \mu y_e$
- Add proximity term

Proximal-point Primal-Dual Augmented Lagrangian:

$$\mathcal{P}(x, y; \mu, \lambda, y_e) = \mu f(x) - y_e^T c(x) + \frac{1}{2} \|c(x)\|^2 + \frac{1}{2} \|c(x) + y - y_e\|^2 + \frac{\lambda}{2} \|x - x_e\|^2$$

- $\mu$ placement similar to Byrd, Curtis, Nocedal 2008.
- $\lambda$ proximity term similar to Friedlander, Orban 2012
- $y$ in $H(x, y)$ replaced with $\gamma y$ (original is approximation)
Alternative derivation:
- Hard-code $\mu = 1$
- Add scale term $\nu$ to objective
- Add proximity term

Proximal-point Primal-Dual Augmented Lagrangian:

$$
\mathcal{P}(x, y; \nu, 1, \lambda, y_e) = \nu f(x) - y_e^T c(x) + \frac{1}{2} \left\| c(x) \right\|^2
+ \frac{1}{2} \left\| c(x) + y - y_e \right\|^2 + \frac{\lambda}{2} \left\| x - x_e \right\|^2
$$
Primal-Dual regularized QP

\[
\begin{align*}
\text{minimize} \quad & g^T x + \frac{1}{2} x^T H x + \frac{\mu}{2} \|y\|^2_2 + \frac{\lambda}{2} \|x - x_e\|^2_2 \\
\text{subject to} \quad & c + J x + \mu (y - y_e) = 0, \quad x \geq 0,
\end{align*}
\]

Dual of Primal-Dual regularized QP

\[
\begin{align*}
\text{minimize} \quad & -c^T y + \frac{1}{2} x^T H x + \frac{\lambda}{2} \|x\|^2_2 + \frac{\mu}{2} \|y - y_e\|^2_2 \\
\text{subject to} \quad & g + H x - J^T y + \lambda (x - x_e) \geq 0.
\end{align*}
\]

Friedlander, Orban (2012)
Simplifications for feasibility restoration:

- \( y_e = 0 \)
- \( y = \pi = -c(x_k) \)
- \( x_e = x_k \)
- \( \gamma = 0 \)
- \( \mu = 0 \) (is now "fscale")
- \( \lambda \) increase/decrease like trust-region algorithm

Preliminary results:

- Old: SAS Test suite with easy problem filtered out
- New: Randomly generated two sets of 900 feasible/infeasible problems
  - \( \ell \leq Ax \leq u \) with \( m \gg n \)
  - \( (a_i^T x - b_i)^2 \leq u_i, \) for \( 1, \ldots, m \)
NUMERICAL RESULTS

NUMERICAL RESULTS COMPARISON FOR HARDER SAS TEST SUITE

AS fscale0 vs AS NewV02 by Time
CONCLUSIONS

FEASIBILITY CONTROL RECOVERED

- Works quite well for most test-problems we’ve tried
- Need to refine $\gamma$ (y-scale for H) and $\nu$ (f-scale) heuristics
- Repeat modification to Interior-Point
- Revise convergence proofs with proximity term present
SAS/OR 13.1 User’s Guide
Mathematical Programming

66851/PDF/default/ormpug.pdf
http://support.sas.com/or

A Nonlinear Regression Perspective on a Primal-Dual Augmented Lagrangian