A NONLINEAR REGRESSION PERSPECTIVE ON A PRIMAL-DUAL AUGMENTED LAGRANGIAN

sas THE POWER TO KNOW.

Southern California Optimization Day

May 23, 2014

Josh Griffin
Wenwen Zhou

BACKGROUND | TALK PURPOSE

- *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

BACKGROUND

LARGE-SCALE NONLINEAR NONCONVEX 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)

$$\begin{array}{lrcl} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f(x) & & \\ \text{subject to} & c(x) & = & 0 \\ & x & \geq & 0. \end{array}$$

- $c(x) \in \mathbb{R}^m$
- $f(x), c(x)$ are twice continuously differentiable

BACKGROUND NOTATION

- 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: $\pi = y_e - c(x)/\mu$

PDAL MERIT FUNCTION

PRIMAL-DUAL AUGMENTED LAGRANGIAN MERIT FUNCTION

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, μ fixed)

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n, y\in\mathbb{R}^m}{\text{minimize}} & M(x, y) \\ \text{subject to} & x \geq 0. \end{array}$$

PDAL MERIT FUNCTION

KKT SYSTEMS

Approximate Newton's system for $\nabla^2 M \Delta v = -\nabla M$:

$$\underbrace{\begin{pmatrix} H(x,y) + \frac{1}{2\mu} J^T J & J^T \\ J & \mu I \end{pmatrix} \begin{pmatrix} p_x \\ p_y \end{pmatrix}}_{B \approx \nabla^2 M} = -\underbrace{\begin{pmatrix} g - J^T(2\pi - y) \\ c(x) + \mu(y - y_e) \end{pmatrix}}_{\nabla M}$$

Sparse equivalent formulation:

$$\begin{pmatrix} H(x,y) & J^T \\ J & -\mu I \end{pmatrix} \begin{pmatrix} p_x \\ -p_y \end{pmatrix} = -\begin{pmatrix} g - J^T y \\ c(x) + \mu(y - y_e) \end{pmatrix}$$

Compare to classical equations

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

PDAL MERIT FUNCTION

QP SUBPROBLEM EQUIVALENCE

Ill-conditioned QP

$$\begin{array}{ll} \underset{v \in \mathbb{R}^{n+m}}{\text{minimize}} & (v - v_k)^T \nabla M + \frac{1}{2}(v - v_k)^T B (v - v_k) \\ \text{subject to} & x \geq 0, v = (x, y). \end{array}$$

Dual regularized QP (Gill, Kungurtsev, Robinson 2013)

$$\begin{array}{ll} \underset{v \in \mathbb{R}^{n+m}}{\text{minimize}} & g^T(x - x_k) + \frac{1}{2}(x - x_k) H (x - x_k) + \frac{1}{2}\mu \|y\|_2^2 \\ \text{subject to} & c + J(x - x_k) + \mu(y - y_e^k) = 0, x \geq 0. \end{array}$$

PDAL MERIT FUNCTION | USING TRUST-REGIONS FOR NONCONVEX CASE

Trust-region subproblem

$$\begin{array}{ll} \underset{v \in \mathbb{R}^{n+m}}{\text{minimize}} & (v - v_k)^T \nabla M + \frac{1}{2}(v - v_k)^T B (v - v_k) \\ \text{subject to} & \|v\| \leq \delta, x \geq 0 \end{array}$$

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

$$P_K = \begin{pmatrix} I & J^T \\ J & -\mu \end{pmatrix} \text{ equivalently } P_B = \begin{pmatrix} I + \frac{1}{2\mu} J^T J & J^T \\ J & \mu \end{pmatrix}$$

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

PDAL MERIT FUNCTION

FGR STRENGTHS AND CHALLENGES

Strengths

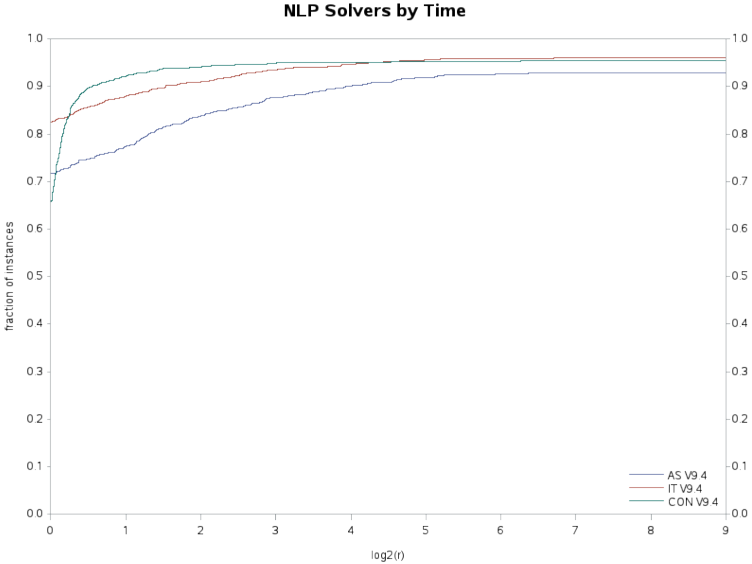
- Primal and dual variables treated nearly identically
- Regularized subproblem
- Potentially locally quadratic convergence rate
- If $y_e \to y^*$, μ need not converge to 0
- Preconditioning optional when μ is large
- Natural constraint preconditioner available

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
- μ often much smaller than classical approaches

PDAL MERIT FUNCTION

SAS TEST SUITE (1097 TEST PROBLEMS)



PDAL MERIT FUNCTION

PRIMARY GOAL

- Preference for minimal algorithmic changes
- Improve constraint handling
- Secondary purpose of μ is regularization
- Primary purpose of μ is counter balance to objective
  - Can μ remain constant if objective is constant?
  - Can μ remain constant if approaching vertex solution?
- $y_e \to y^*$ no longer critical for performance
- Can y_e, y remain bounded for infeasible problems?

$$y_e^{k+1} \quad \to \quad y^k \to y_e^k - \frac{c(x_k)}{\mu_k} \tag{1}$$

$$\Rightarrow \quad \|y_e^k\|, \|y^k\| \to \infty \tag{2}$$

if infeasible and $\mu_k \to 0$

PDAL MERIT FUNCTION

WHY LAGRANGE ESTIMATE IS CRUCIAL

$$\begin{array}{ll} \underset{x\in\mathbb{R}}{\text{minimize}} & -10^5 x \\ \text{subject to} & 10^{-5}x = 0. \end{array}$$

Assume $y_e = 0$, then

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

μ	$x(\mu)$	$c(x(\mu))$
1	10^{15}	10^{10}
10^{-6}	10^{9}	10^{4}
10^{-16}	.1	10^{-6}

Linearized constraint approach of course solves in 1 step.

FULL ROW RANK ASSUMPTION

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:

$$\begin{array}{ll} \underset{s \in \mathbb{R}^n}{\text{minimize}} & \frac{1}{2}\|s\|_M^2 \\ \text{subject to} & Js + c = 0. \end{array}$$

CONSTANT OBJECTIVE INTERIOR

KKT INTERPRETATION

$$\begin{array}{ll} \underset{s_x \in \mathbb{R}^n}{\text{minimize}} & \|s_x\|_M^2 \\ \text{subject to} & J s_x + c = 0. \end{array}$$

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

NO ROW RANK ASSUMPTION

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:

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

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

SPARSE EQUATIONS

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

Can be found as solution to sparse system

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

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 KKT INTERPRETATION

Regularized Newton-systems have the form:

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

where

- $H(x,y) = -\sum_{i=1}^{m} y_i \nabla^2 c_i(x)$
- λI missing (sometimes added as part of trust-region solver)
- Intermediate y can grow large
- Negligible second-order term from LM starts to dominate

REGULARIZED KKT INTERPRETATION (WITH PRIMAL PROXIMITY TERM)

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 μ.

FEASIBILITY CONTROL | PROXIMAL-POINT PRIMAL-DUAL MERIT FUNCTION

Transformation steps:

- Scale $\mathcal{M}(x, y; y_e, \mu)$ by μ
- Redefine $y = \mu y$, $y_e = \mu y_e$
- Add proximity term

Proximal-point Primal-Dual Augmented Lagrangian:

$$\begin{aligned}\mathcal{P}(x, y; \mu, \lambda, y_e) &= \mu f(x) - y_e^T c(x) + \frac{1}{2}\|c(x)\|^2 \\ &\quad + \frac{1}{2}\|c(x) + y - y_e\|^2 + \frac{\lambda}{2}\|x - x_e\|_2^2\end{aligned}$$

- μ placement similar to Byrd, Curtis, Nocedal 2008.
- λ proximity term similar to Friedlander, Orban 2012
- y in $H(x, y)$ replaced with γy (original is approximation)

FEASIBILITY CONTROL | PROXIMAL-POINT PRIMAL-DUAL MERIT FUNCTION

Alternative derivation:

- Hard-code $\mu = 1$
- Add scale term ν to objective
- Add proximity term

Proximal-point Primal-Dual Augmented Lagrangian:

$$\begin{aligned}\mathcal{P}(x, y; \nu, 1, \lambda, y_e) &= \nu f(x) - y_e^T c(x) + \frac{1}{2}\|c(x)\|^2 \\ &+ \frac{1}{2}\|c(x) + y - y_e\|^2 + \frac{\lambda}{2}\|x - x_e\|_2^2\end{aligned}$$

PRIMAL-DUAL REGULARIZED QP SYMMETRY

Primal-Dual regularized QP

$$\begin{array}{ll} \underset{x,y}{\text{minimize}} & g^T x + \frac{1}{2} x^T H x + \frac{\mu}{2}\|y\|_2^2 + \frac{\lambda}{2}\|x - x_e\|_2^2 \\ \text{subject to} & c + Jx + \mu(y - y_e) = 0, x \geq 0, \end{array}$$

Dual of Primal-Dual regularized QP

$$\begin{array}{ll} \underset{y,x}{\text{minimize}} & -c^T y + \frac{1}{2} x^T H x + \frac{\lambda}{2}\|x\|_2 + \frac{\mu}{2}\|y - y_e\|_2^2 \\ \text{subject to} & g + Hx - J^T y + \lambda(x - x_e) \geq 0. \end{array}$$

Friedlander, Orban (2012)

NUMERICAL RESULTS | SIMPLIFICATION FOR FEASIBILITY RESTORATION MODE

Simplifications for feasibility restoration:

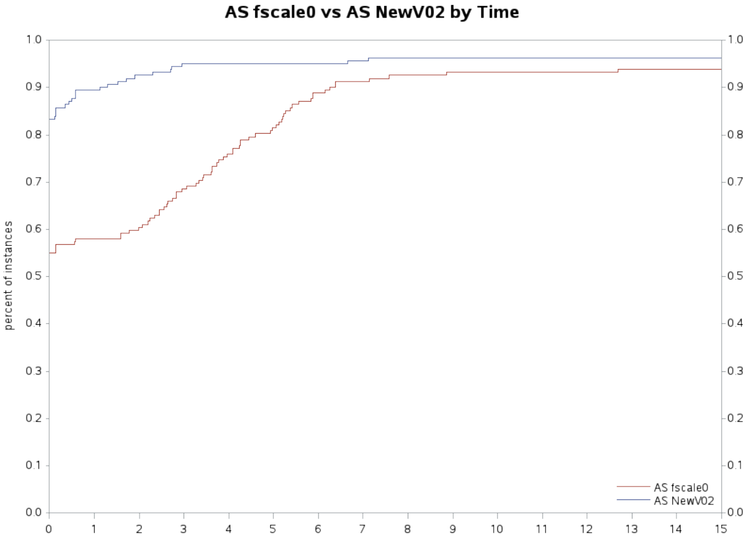
- $y_e = 0$
- $y = \pi = -c(x_k)$
- $x_e = x_k$
- $\gamma = 0$
- $\mu = 0$ (is now "fscale")
- λ increase/decrease like trust-region algorithm

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 $1, \ldots, m$

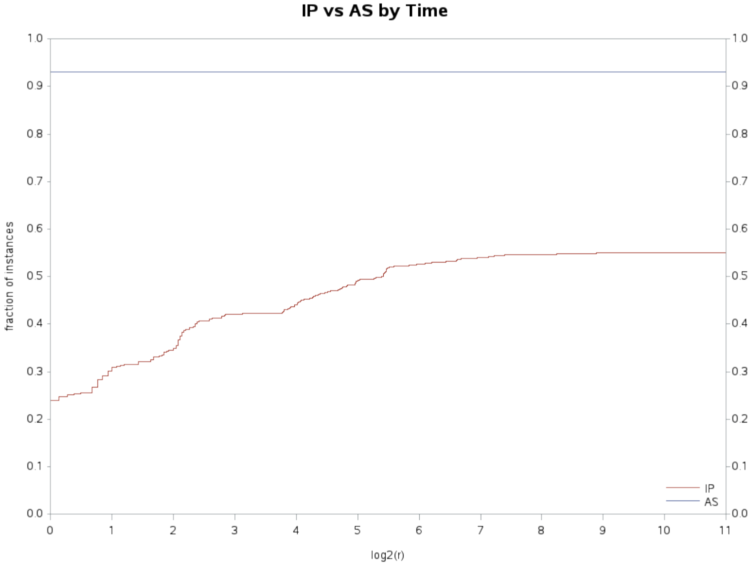
NUMERICAL RESULTS

NUMERICAL RESULTS COMPARISON FOR HARDER SAS TEST SUITE



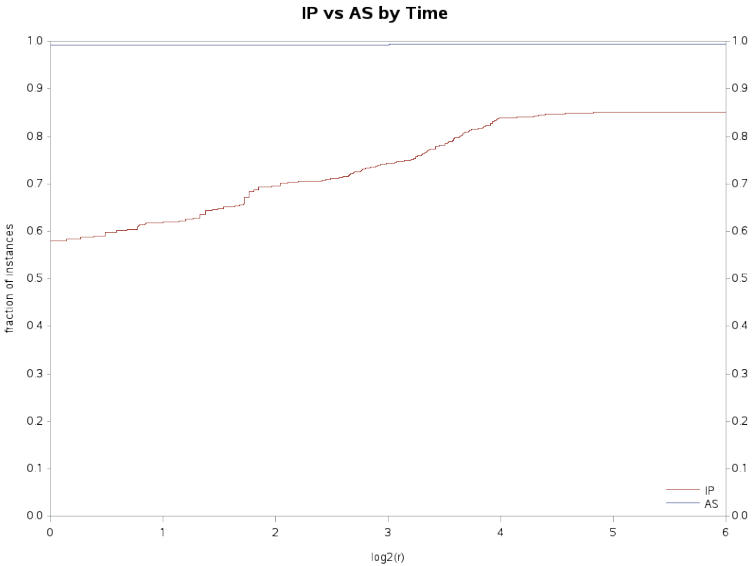
NUMERICAL RESULTS

RANDOMLY GENERATED TEST SUITE I



NUMERICAL RESULTS

RANDOMLY GENERATED TEST SUITE II



CONCLUSIONS | FEASIBILITY CONTROL RECOVERED

- Works quite well for most test-problems we've tried
- Need to refine γ (y-scale for H) and ν (f-scale)heuristics
- Repeat modification to Interior-Point
- Revise convergence proofs with proximity term present

REFERENCES AND DOCUMENTATION

SAS/OR 13.1 User's Guide
Mathematical Programming

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

A Nonlinear Regression Perspective on a Primal-Dual Augmented Lagrangian