What's New in Active-Set Methods for Nonlinear Optimization?

Philip E. Gill

Advances in Numerical Computation, Manchester University, July 5, 2011

A Workshop in Honor of Sven Hammarling

What's it all about?

From Wikipedia, 2011

> *"Sequential quadratic programming (SQP) is one of the most popular and robust algorithms for nonlinear continuous optimization. The method is based on solving a series of subproblems designed to minimize a quadratic model of the objective subject to a linearization of the constraints . . . "*

Continuous nonlinear optimization

Given functions that define $f(x)$ and $c(x)$ (and their derivatives) at any x, solve

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & \ell \le \left\{ \begin{array}{c} x \\ c(x) \\ Ax \end{array} \right\} \le u \end{array}$$

The ground rules:

- f and c are arbitrary, but *smooth* functions
- Large number of variables
- Local solutions

A trick learned from LP—add slack variables

$$
\begin{array}{ll}
\underset{x,\, s_A,\, s_C}{\text{minimize}} & f(x) \\
\text{subject to} & c(x) - s_C = 0, \quad Ax - s_A = 0 \\
& \ell \le \begin{Bmatrix} x \\ s_C \\ s_A \end{Bmatrix} \le u
\end{array}
$$

The slacks s_A, s_C provide a constraint Jacobian of full rank.

Prototype problem

Without loss of generality, we consider the problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & c(x) = 0, \quad x \geq 0 \end{array}$$

The $m \times n$ constraint Jacobian has rank m.

Some events in the development of SQP methods

1963	Wilson
1972	MINOS, Murtagh & Saunders
1975	Han & Powell '76
1975–84	the SQP "salad days"
1982	NPSOL, G, Murray, Saunders & Wright (and Sven!)
1984	Karmarkar and the interior-point (IP) "revolution"
1985–	barrier methods, G, Murray, Saunders, Tomlin & Wright '86
1992–	SNOPT, G, Murray & Saunders '97
1997–	AMPL, GAMS introduce automatic differentiation
2008–	the SQP renaissance

Outline

Overview of SQP methods

First, consider the *equality constrained problem:*

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \ f(x) \quad \text{subject to} \quad c(x) = 0$$

- The *objective gradient and Hessian:*
  $$g(x) \triangleq \nabla f(x), \qquad\qquad H(x) \triangleq \nabla^2 f(x)$$
- The $m \times n$ *constraint Jacobian:* $A(x) \triangleq c'(x)$
- The *Lagrangian* $\mathcal{L}(x,\pi) = f(x) - c(x)^T\pi$
- The *Lagrangian gradient and Hessian:*
  $$\nabla_x \mathcal{L}(x,\pi), \qquad H(x,\pi) \triangleq \nabla^2_{xx}\mathcal{L}(x,\pi)$$
- A *local optimal solution* (x^*,π^*)

The gradient of the Lagrangian with respect to both x and π is:

$$\nabla\mathcal{L}(x,\pi) = \begin{pmatrix} g(x) - A(x)^T\pi \\ -c(x) \end{pmatrix}$$

An optimal solution (x^*, π^*) is a stationary point of $\mathcal{L}(x,\pi)$, i.e.,

$$\nabla\mathcal{L}(x^*,\pi^*) = 0$$

The vector (x^*, π^*) solves the nonlinear equations

$$\nabla\mathcal{L}(x,\pi) = \begin{pmatrix} g(x) - A(x)^T\pi \\ -c(x) \end{pmatrix} = 0$$

$n+m$ nonlinear equations in the $n+m$ variables x and π.

Apply *Newton's method* to find a solution of $\nabla\mathcal{L}(x,\pi) = 0$.

Newton's method converges at a *second-order rate*.

$$\left(\text{“Jacobian”}\right)\begin{pmatrix}\text{“Change in}\\ \text{variables”}\end{pmatrix} = -\left(\text{“Residual”}\right)$$

The $(n+m)\times(n+m)$ Jacobian is

$$\begin{pmatrix} H(x,\pi) & -A(x)^T \\ -A(x) & 0 \end{pmatrix}$$

with $H(x,\pi) = \nabla^2 f(x) - \sum_{i=1}^{m} \pi_i \nabla^2 c_i(x)$, the *Lagrangian Hessian*.

Suppose we are given a primal-dual estimate (x_0, π_0).

The *Newton equations* for (p, q), the change to (x_0, π_0), are:

$$\begin{pmatrix} H(x_0, \pi_0) & -A(x_0)^T \\ -A(x_0) & 0 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = - \begin{pmatrix} g(x_0) - A(x_0)^T \pi_0 \\ -c(x_0) \end{pmatrix}$$

These are just the Karush-Kuhn-Tucker *KKT* equations

$$\begin{pmatrix} H_0 & A_0^T \\ A_0 & 0 \end{pmatrix} \begin{pmatrix} p \\ -q \end{pmatrix} = - \begin{pmatrix} g_0 - A_0^T \pi_0 \\ c_0 \end{pmatrix}$$

Set $x_1 = x_0 + p$, and $\pi_1 = \pi_0 + q$.

Wilson's light-bulb moment!

$(x_0 + p, \pi_0 + q)$ is the primal-dual solution of the *quadratic subproblem*:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & g_0^T(x - x_0) + \frac{1}{2}(x - x_0)^T H_0 (x - x_0) \\ \text{subject to} & c_0 + A_0(x - x_0) = 0 \end{array}$$

The sequence $\{(x_k, \pi_k)\}$ converges at a *second-order rate*.

Now consider the *inequality constrained problem*

Given (x_0, π_0), the "Wikipedia" SQP subproblem is:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & g_0^T(x - x_0) + \frac{1}{2}(x - x_0)^T H_0(x - x_0) \\ \text{subject to} & c_0 + A_0(x - x_0) = 0, \quad x \geq 0 \end{array}$$

The QP must be solved by iteration.

$\Rightarrow$ inner/outer iteration structure.

QP solution (x_k^*, π_k^*).

Given any x $(x \geq 0)$, the *active set* is $\mathcal{A}(x) = \{\, i : x_i = 0 \,\}$.

The ϵ-active set is $\mathcal{A}_\epsilon(x) = \{\, i : x_i \leq \epsilon \,\}$.

If $x_k \to x^*$, then

$$\mathcal{A}_\epsilon(x_k) = \mathcal{A}(x^*) \quad \text{for } k \text{ sufficiently large}$$

Define the *free* variables as those with indices *not* in $\mathcal{A}_\epsilon(x)$.

If $\mathcal{A}_\epsilon(x_k) = \mathcal{A}(x^*)$, the QP optimality conditions imply

$$\begin{pmatrix} H_F & A_F^T \\ A_F & 0 \end{pmatrix} \begin{pmatrix} p_F \\ -\pi_k^* \end{pmatrix} = - \begin{pmatrix} g_F \\ c_k \end{pmatrix}$$

where

- p_F is the vector of free components of $x_k^* - x_k$
- A_F is the matrix of free columns of $A(x_k)$
- H_F is the matrix of free rows and columns of $H(x_k, \pi_k)$
- g_F is the vector of free components of $g(x_k)$

If x^* is *nondegenerate*, then A_F has full row rank.

If (x^*, π^*) satisfies the second-order sufficient conditions, then

$$\begin{pmatrix} H_F & A_F^T \\ A_F & 0 \end{pmatrix} \begin{pmatrix} p_F \\ -\pi_k^* \end{pmatrix} = - \begin{pmatrix} g_F \\ c_k \end{pmatrix} \quad \text{is } \textit{nonsingular}$$

$\Rightarrow$ eventually, "Wikipedia SQP" is Newton's method applied to the problem in the free variables.

Two-phase active-set methods

A sequence of equality-constraint QPs is solved, each defined by fixing a subset of the variables on their bounds.

Sequence of related KKT systems with matrix

$$K = \begin{pmatrix} H_F & A_F^T \\ A_F & \end{pmatrix}$$

- A_F has column a_s *added*, or column a_t *deleted*
- H_F has a row and column *added* or *deleted*

These changes are reflected in some factorization of K.

If the fixed set from one QP is used to start the next QP, the subproblems usually require one QP iteration near the solution.

With a good starting point, SQP requires few QP iterations

Four fundamental issues associated with “Wikipedia SQP”:

- Global convergence
  - *Is* (x_{k+1}, π_{k+1}) *“better” than* (x_k, π_k) ?
- Ill-posed QP subproblems near (x^*, π^*)
  - QP subproblem may be infeasible
  - *Ill-conditioned or singular equations*
- Computational efficiency
  - Sequence of linear equations with changing structure
  - *Need to use efficient software for linear equations*
- Nonconvex QP subproblems
  - Indefinite QP is difficult!

Global convergence

Line-search and *trust-region* methods force convergence by ensuring that $\mathcal{M}(x_{k+1}, \pi_{k+1}) < \mathcal{M}(x_k, \pi_k)$ for some *merit function* $\mathcal{M}(x, \pi)$.

Two popular merit functions are:

- the ℓ_1 penalty function:

$$\mathcal{M}(x) = f(x) + \frac{1}{\mu} \sum_{i=1}^{m} |c_i(x)|$$

- the augmented Lagrangian merit function

$$\mathcal{M}(x, \pi) = f(x) - \pi^T c(x) + \frac{1}{2\mu} \sum_{i=1}^{m} c_i(x)^2$$

μ is the *penalty parameter*.

Ill-Conditioning and Singularity

At a *degenerate* QP solution, the rows of A_F are linearly dependent

$$\Rightarrow \qquad \begin{pmatrix} H_F & A_F^T \\ A_F & 0 \end{pmatrix} \begin{pmatrix} p_k \\ -\pi_k^* \end{pmatrix} = -\begin{pmatrix} g_F \\ c_k \end{pmatrix} \quad \text{is } \textit{singular}$$

Almost all practical optimization problems are degenerate

Options:

- Identify an A_F with linearly independent rows e.g., SNOPT. G, Murray & Saunders '05.
- Regularize the KKT system. Hager '99, Wright '05.

Where does SNOPT fit in this discussion?

- Positive-definite $H \Rightarrow$ the subproblem is a convex program
  - H is approximated by a *limited-memory quasi-Newton method*
- A *two-phase active-set method* is used for the convex QP
  - *Elastic mode* is entered if the QP is infeasible or the multipliers are large
- The KKT equations are solved by updating factors of A_F and the reduced Hessian

Interest in SQP methods declines. . .

In the late 1980s/early 1990's, research on SQP methods declined.

Three reasons (but interconnected):

- The rise of interior-point methods
- The rise of automatic differentiation packages
  - modeling languages such as AMPL and GAMS started to provide second derivatives automatically.
- Computer architecture evolved

The "Wikipedia" QP

$$\begin{array}{ll} \underset{x}{\text{minimize}} & g_k^T(x - x_k) + \frac{1}{2}(x - x_k)^T H_k (x - x_k) \\ \text{subject to} & c_k + A_k(x - x_k) = 0, \quad x \geq 0 \end{array}$$

is *NP hard* when H_k is indefinite.

Methods based on solving indefinite QP's are problematic.

Efficient software for linear equations

Computer hardware is changing

- Moore's Law is fading

  "The number of transistors on a microchip will double every 18 months"

- Moore's Law has been "updated":

  "the number of cores (cpus) on a processor will double every 18 months"

- it's already happening. . .
  - 2008 Mac G5: 4 quad-core processors = 16 cpus
  - 2011 Mac Book: dual 16-core processors = 32 cpus
  - 2013 dual 132-core = 264 cpus
  - $>$ 2008 potentially hundreds of cpus using GPUs

20 years of progress

Linear programming with MINOS

PILOT 1442 rows, 3652 columns, 43220 nonzeros

Year	Itns	Cpu secs	Architecture
1987	–	8.7×10^4	DEC Vaxstation II
⋮	⋮	⋮	⋮
2005	17738	22.2	dual-core Xeon
2006	16865	9.7	dual-core Opteron 2.4Ghz
2007	16865	8.1	dual-core Opteron 3.1Ghz
2008	16865	8.7	quad-core Opteron 3.1Ghz

The **nice** features of IP methods

IP methods . . .

- work best when second derivatives are provided
- solve a sequence of systems with *fixed structure*
  - they can exploit solvers designed for modern computer architectures
- IP methods are blazingly fast on one-off problems

The SQP renaissance

Then, things started to change. . .

Many important applications require the solution of a *sequence of related optimization problems*

- ODE and PDE-based optimization with mesh refinement
- Mixed-integer nonlinear programming
  - infeasible constraints are likely to occur

The common feature is that we would like to benefit from *good approximate solutions*.

The **not-so-nice** features of IP methods

IP methods ...

- have difficulty exploiting a good solution
- have difficulty certifying infeasible constraints
- have difficulty exploiting linear constraints
- factor a KKT matrix with every constraint present

IP methods are fast on one-off problems *that aren't too hard*

SQP vs IP *Ying vs Yang*? or is it *Yang vs Ying*?

Modern SQP methods

(Joint work with Daniel Robinson)

Modern SQP Methods

Aims:

- to define an SQP method that exploits second derivatives.
- to provide a globally convergent method that is provably effective for degenerate problems
  - perform stabilized SQP near a solution
- allow the use of modern *sparse matrix* packages
  - "black-box" linear equation solvers
- *Do all of the above as seamlessly as possible!*

When formulating methods, how may we best exploit modern computer architectures?

- Methods based on sparse updating are hard to speed up
- Reformulate methods to shift the emphasis from *sparse matrix updating* to *sparse matrix factorization*
  - Thereby exploit state-of-the-art linear algebra software
  - Less reliance on specialized "home grown" software

Shifting from updating to factorization

An SQP example

Given $K = \begin{pmatrix} H & A_F^T \\ A_F & \end{pmatrix}$, quantities for the next QP iteration may be found by solving a *bordered system* with matrices:

$$\left(\begin{array}{cc|c} H & A_F^T & h_t \\ A_F & & a_t \\ \hline h_t^T & a_t^T & h_{tt} \end{array}\right) \qquad \text{(add column } a_t\text{)}$$

$$\left(\begin{array}{cc|c} H & A_F^T & e_s \\ A_F & & 0 \\ \hline e_s^T & 0 & 0 \end{array}\right) \qquad \text{(delete column } a_s\text{)}$$

Schur complement QP method

G, Murray, Saunders & Wright 1990

In general,

$$K_j v = f \qquad \equiv \qquad \begin{pmatrix} K_0 & W \\ W^T & D \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

1 solve with dense Schur-complement $C = D - W^T K_0^{-1} W$

2 solves with K_0

Used in GALAHAD's QPA, Gould & Toint '04

Block-LU updates G, Murray, Saunders & Wright '84, Huynh '08

Infeasibility, ill-conditioning and all that. . .

Given fixed $\pi_E \approx \pi^*$, and fixed $\mu > 0$, consider the *generalized augmented Lagrangian*

$$\mathcal{M}(x, \pi; \pi_E, \mu) = f(x) - c(x)^T \pi_E + \frac{1}{2\mu} \|c(x)\|_2^2 + \frac{1}{2\mu} \|c(x) + \mu(\pi - \pi_E)\|_2^2$$

G & Robinson '10.

$\mathcal{M}$ involves $n+m$ variables and has gradient

$$\nabla\mathcal{M}(x,\pi;\pi_E,\mu) = \begin{pmatrix} g(x) - A(x)^T\big(\pi_A - (\pi - \pi_A)\big) \\ \mu\big(\pi - \pi_A\big) \end{pmatrix}$$

where $\pi_A \equiv \pi_A(x) = \pi_E - c(x)/\mu$.

The Hessian of $\mathcal{M}$ is

$$\nabla^2\mathcal{M}(x,\pi;\pi_E,\mu) = \begin{pmatrix} H + \frac{2}{\mu}A^TA & A^T \\ A & \mu I \end{pmatrix}$$

with $H = H\big(x, \pi_A - (\pi - \pi_A)\big)$.

Result I

Theorem

Consider the bound constrained problem

$$\underset{x,\pi}{\text{minimize}} \;\; \mathcal{M}(x,\pi\,;\pi^*,\mu) \quad \text{subject to} \quad x \geq 0 \qquad \text{(BC)}$$

where π^* is a Lagrange multiplier vector.

If (x^*,π^*) satisfies the second-order sufficient conditions for the problem:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; f(x) \quad \text{subject to} \quad c(x) = 0, \;\; x \geq 0$$

then there is a $\bar{\mu} > 0$ such that (x^*,π^*) is a minimizer of (BC) for all $0 < \mu < \bar{\mu}$.

[0]. Choose initial μ and π_E, an estimate of π^*;

[1]. Find an approximate solution of

$$\underset{x,\pi}{\text{minimize}} \;\; \mathcal{M}(x,\pi\,;\pi_E,\mu) \quad \text{subject to} \quad x \ge 0$$

[2]. Update π_E and μ; Repeat at [1].

The problem

$$\underset{x,\pi}{\text{minimize}} \;\; \mathcal{M}(x, \pi\,; \pi_E, \mu) \quad \text{subject to} \quad x \geq 0$$

is solved using a line-search method that minimizes a sequence of quadratic models:

$$Q_{\mathcal{M}}(x, \pi\,; \pi_E, \mu) \approx \mathcal{M}(x, \pi\,; \pi_E, \mu)$$

Two different values of μ are maintained:

- For the line search on $\mathcal{M}$: $\mu = \mu_k$ with "large" μ_k
- For the QP subproblem with $Q_{\mathcal{M}}$: $\mu = \mu_R$ with $\mu_R \ll \mu_k$

We solve a sequence of *convex* QPs:

$$\begin{array}{ll} \underset{v=(x,\pi)}{\text{minimize}} & Q_{\mathcal{M}}(v) = g_{\mathcal{M}}^T(v - v_k) + \frac{1}{2}(v - v_k)^T H_{\mathcal{M}}(v - v_k) \\ \text{subject to} & x \geq 0 \end{array}$$

where $v_k = (x_k, \pi_k)$, and

$$g_{\mathcal{M}} = \nabla\mathcal{M}(x_k, \pi_k; \mu_R), \qquad H_{\mathcal{M}} \approx \nabla^2\mathcal{M}(x_k, \pi_k; \mu_R)$$

We define

$$H_{\mathcal{M}} = \begin{pmatrix} \bar{H}_k + \frac{2}{\mu} A_k^T A_k & A_k^T \\ A_k & \mu I \end{pmatrix}$$

where

- $\bar{H}_k = H(x_k, \pi_k) + D_k$, where D_k is a sparse diagonal.
- D_k is chosen so that $\bar{H}_k + \frac{1}{\mu} A_k^T A_k$ positive definite.

Result II

Theorem (G & Robinson '11)

The bound constrained QP

$$\underset{\Delta v=(p,q)}{\text{minimize}} \;\; g_{\mathcal{M}}^T \Delta v + \tfrac{1}{2}\Delta v^T H_{\mathcal{M}} \Delta v \quad \text{subject to} \quad x + p \geq 0$$

is equivalent to the QP problem

$$\begin{aligned} \underset{p,q}{\text{minimize}} \quad & g^T p + \tfrac{1}{2} p^T \bar{H} p + \tfrac{1}{2}\mu \|\pi + q\|^2 \\ \text{subject to} \quad & c + Ap + \mu(\pi + q - \pi_E) = 0, \quad x + p \geq 0. \end{aligned}$$

(known as the "stabilized" SQP subproblem).

At QP iteration j, a direction $(\Delta p_j, \Delta q_j)$ is found satisfying

$$\begin{pmatrix} \bar{H}_F & -A_F^T \\ A_F & \mu I \end{pmatrix} \begin{pmatrix} \Delta p_F \\ \Delta q_j \end{pmatrix} = - \begin{pmatrix} (\widehat{g}(x_j) - A_k^T \pi_j)_F \\ \widehat{c}(x_j) + \mu(\pi_j - \pi_E) \end{pmatrix},$$

with $\widehat{g}(x) = g_k + \bar{H}_k(x - x_k)$ and $\widehat{c}(x) = c_k + A_k(x - x_k)$

- This system is nonsingular for $\mu > 0$
- If $\mu = \mu_R$ (small), then this is a "stabilized" SQP step
- "Black-box" symmetric indefinite solvers may be used

- No "phase-one" procedure is needed for the QP
- The QP subproblem is *always* feasible
- As the outer iterations converge, the directions (p_k, q_k) satisfy

$$\begin{pmatrix} \bar{H}_F & -A_F^T \\ A_F & \mu I \end{pmatrix} \begin{pmatrix} p_F \\ q_k \end{pmatrix} = - \begin{pmatrix} (g_k - A_k^T \pi_k)_F \\ c_k + \mu(\pi_k - \pi_E) \end{pmatrix}$$

  These equations define $\pi_k + q_k$ as an $O(\mu)$ estimate of the unique least-length Lagrange multipliers.
- A *fixed* sparse matrix is can be factored.

Properties of the modification

If the QP *does not change the active set*, then the final KKT system satisfies

$$\begin{pmatrix} \bar{H}_F & A_F^T \\ A_F & -\mu I \end{pmatrix} = \begin{pmatrix} H_F + D_F & A_F^T \\ A_F & -\mu I \end{pmatrix} = \begin{pmatrix} H_F & A_F^T \\ A_F & -\mu I \end{pmatrix}$$

$\Rightarrow$ the QP step is computed using H_F (unmodified) and A_F.

$\Rightarrow$ in the limit, this is Newton's method wrt the free variables.

$\Rightarrow$ potential second-order convergence rate.

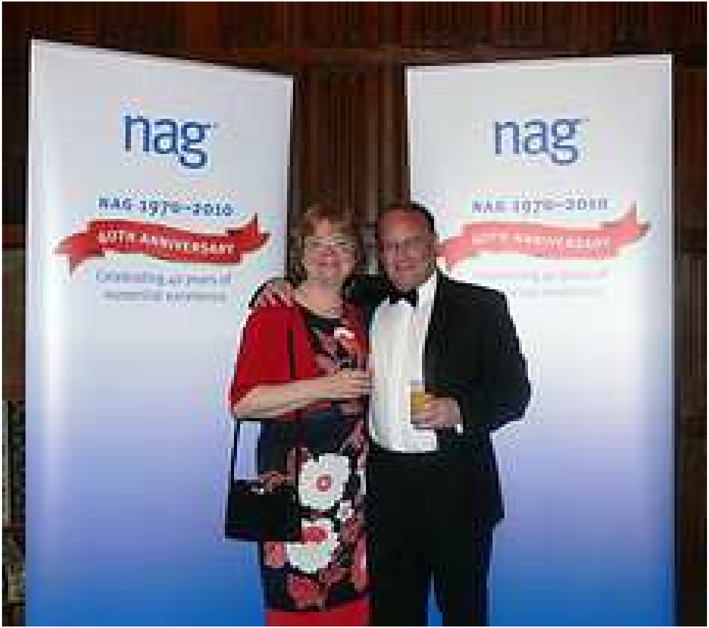
Summary and comments

- Recent developments in MINLP and PDE- and ODE-constrained optimization has sparked renewed interest in second-derivative SQP methods
- Multi-core architectures require new ways of looking at how optimization algorithms are formulated
  - Reliance on state-of-the-art linear algebra software

The method . . .

- involves a convex QP for which the dual variables may be bounded explicitly
- is based on sparse matrix factorization
  - allows the use of some "black-box" indefinite solvers
- is "global" but reduces to stabilized SQP near a solution

Happy Birthday Sven!


nag
nag

References

Philip E. Gill & Daniel Robinson, *A primal-dual augmented Lagrangian*, Computational Optimization and Applications, **47** (2010), 1-25.

Philip E. Gill & Elizabeth Wong, *Methods for convex and general quadratic programming*, Report NA 10-1, Department of Mathematics, University of California, San Diego, 2010.

Philip E. Gill & Elizabeth Wong, *Sequential quadratic programming methods*, in J. Lee & S. Leyffer (eds.), *Mixed-Integer Nonlinear Optimization: Algorithmic Advances and Applications*, The IMA Volumes in Mathematics and its Applications, Springer Verlag,Berlin, Heidelberg and New York, 2011.

Philip E. Gill & Daniel Robinson, *Regularized primal-dual sequential quadratic programming methods*, Report NA 11-1, Department of Mathematics, University of California, San Diego, 2011.