Optimal Value Function Methods in Numerical Optimization
Level Set Methods

James V Burke
Mathematics, University of Washington, (jvburke@uw.edu)

Joint work with
Aravkin (UW), Drusvyatskiy (UW), Friedlander (UBC/Davis), Roy (UW)

The Hong Kong Polytechnic University
Applied Mathematics Colloquium
February 4, 2016
Motivation

Optimization in Large-Scale Inference

• A range of large-scale data science applications can be modeled using optimization:
  - Inverse problems (medical and seismic imaging)
  - High dimensional inference (compressive sensing, LASSO, quantile regression)
  - Machine learning (classification, matrix completion, robust PCA, time series)

• These applications are often solved using side information:
  - Sparsity or low rank of solution
  - Constraints (topography, non-negativity)
  - Regularization (priors, total variation, “dirty” data)

• We need efficient large-scale solvers for nonsmooth programs.
The Prototypical Problem

Sparse Data Fitting:

Find sparse \( x \) with \( Ax \approx b \)

Example: Model Selection

\[ y = a^T x \]

where \( y \in \mathbb{R}^k \) is an observation and \( a \in \mathbb{R}^n \) are covariates.

Suppose \( y \) is a disease classifier and \( a \) is micro-array data (\( n \geq 10^4 \)).

Given data \( \{ (y_i, a_i) \}_{i=1}^m \), find \( x \) so that \( y_i \approx a_i^T x \).

Since \( m \ll n \), one can “always” find \( x \) such that \( y_i = a_i^T x \), \( i = 1, \ldots, m \).

This \( x \) gives little insight into the role of the covariates \( a \) in determining the observations. We prefer the most parsimonious subset of covariates that can be used to explain the observations. That is, we prefer the sparsest model from the \( 2^n \) possible models. Such models are used to further our knowledge of disease mechanisms and to develop efficient disease assays.
Sparse Data Fitting:

Find sparse $x$ with $Ax \approx b$

Example: Model Selection

$y = a^T x$ where $y \in \mathbb{R}^k$ is an observation and $a \in \mathbb{R}^n$ are covariates.

Suppose $y$ is a disease classifier and $a$ is micro-array data $(n \geq 10^4)$. Given data $\{(y_i, a_i)\}_{i=1}^m$, find $x$ so that $y_i \approx a_i^T x$. 
The Prototypical Problem

Sparse Data Fitting:

Find \textit{sparse} \(x\) with \(Ax \approx b\)

Example: Model Selection

\[ y = a^T x \] where \(y \in \mathbb{R}^k\) is an observation and \(a \in \mathbb{R}^n\) are covariates.

Suppose \(y\) is a disease classifier and \(a\) is micro-array data (\(n \geq 10^4\)). Given data \(\{(y_i, a_i)\}_{i=1}^m\), find \(x\) so that \(y_i \approx a_i^T x\).

Since \(m \ll n\), one can “always” find \(x\) such that \(y_i = a_i^T x, \quad i = 1, \ldots, m\).
The Prototypical Problem

Sparse Data Fitting:

\[
\text{Find sparse } x \text{ with } Ax \approx b
\]

Example: Model Selection

\[ y = a^T x \quad \text{where } y \in \mathbb{R}^k \text{ is an observation and } a \in \mathbb{R}^n \text{ are covariates.} \]

Suppose \( y \) is a disease classifier and \( a \) is micro-array data (\( n \geq 10^4 \)).

Given data \( \{(y_i, a_i)\}_{i=1}^m \), find \( x \) so that \( y_i \approx a_i^T x \).

Since \( m << n \), one can “always” find \( \bar{x} \) such that

\[ y_i = a_i^T \bar{x}, \quad i = 1, \ldots, m. \]

This \( \bar{x} \) gives little insight into the role of the covariates \( a \) in determining the observations \( y \). We prefer the most parsimonious subset of covariates that can be used to explain the observations.

That is, we prefer the sparsest model from the \( 2^n \) possible models. Such models are used to further our knowledge of disease mechanisms and to develop efficient disease assays.
The Prototypical Problem

Sparse Data Fitting:

Find sparse \( x \) with \( Ax \approx b \)

There are numerous other applications:

- system identification
- image segmentation
- compressed sensing
- grouped sparsity for remote sensor location
- ...

The Prototypical Problem

Sparse Data Fitting:

Find sparse $x$ with $Ax \approx b$
The Prototypical Problem

Sparse Data Fitting:

\[
\text{Find \textbf{sparse} } x \text{ with } Ax \approx b
\]

Convex approaches: \( \|x\|_1 \) as a sparsity surragate
(Candes-Tao-Donaho)

<table>
<thead>
<tr>
<th>\text{BPDN}</th>
<th>\text{LASSO}</th>
<th>\text{Lagrangian (Penalty)}</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \min_x |x|_1 ]</td>
<td>[ \min_x \frac{1}{2} | Ax - b |_2^2 ]</td>
<td>[ \min_x \frac{1}{2} | Ax - b |_2^2 + \lambda |x|_1 ]</td>
</tr>
<tr>
<td>s.t. [ \frac{1}{2} | Ax - b |_2^2 \leq \sigma ]</td>
<td>s.t. [ |x|_1 \leq \tau ]</td>
<td></td>
</tr>
</tbody>
</table>
Sparse Data Fitting:

Find sparse \( x \) with \( Ax \approx b \)

Convex approaches: \( \| x \|_1 \) as a sparsity surrogate

(Candes-Tao-Donoho)

<table>
<thead>
<tr>
<th></th>
<th>BPDN</th>
<th>LASSO</th>
<th>Lagrangian (Penalty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>min ( x )</td>
<td>( | x |_1 )</td>
<td>( \min_x \frac{1}{2} | Ax - b |_2^2 )</td>
<td>( \min_x \frac{1}{2} | Ax - b |_2^2 + \lambda | x |_1 )</td>
</tr>
<tr>
<td>s.t. ( \frac{1}{2} | Ax - b |_2^2 \leq \sigma )</td>
<td>s.t. ( | x |_1 \leq \tau )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- BPDN: often most natural and transparent.
  (physical considerations guide \( \sigma \))
The Prototypical Problem

Sparse Data Fitting:

\[
\text{Find \textit{sparse} } x \text{ with } Ax \approx b
\]

Convex approaches: \(\|x\|_1\) as a sparsity surrogate

(Candes-Tao-Donaho)

<table>
<thead>
<tr>
<th>BPDN</th>
<th>LASSO</th>
<th>Lagrangian (Penalty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\min_{x} |x|_1)</td>
<td>(\min_{x} \frac{1}{2} |Ax - b|_2^2)</td>
<td>(\min_{x} \frac{1}{2} |Ax - b|_2^2 + \lambda |x|_1)</td>
</tr>
<tr>
<td>s.t. (\frac{1}{2} |Ax - b|_2^2 \leq \sigma)</td>
<td>s.t. (|x|_1 \leq \tau)</td>
<td></td>
</tr>
</tbody>
</table>

- **BPDN**: often most \textit{natural} and \textit{transparent}.
  (physical considerations guide \(\sigma\))
- **Lagrangian**: \textit{ubiquitous} in practice.
  ("no constraints")
The Prototypical Problem

Sparse Data Fitting:

Find sparse $x$ with $Ax \approx b$

Convex approaches: $\|x\|_1$ as a sparsity surragate
(Candes-Tao-Donaho)

<table>
<thead>
<tr>
<th>BPDN</th>
<th>LASSO</th>
<th>Lagrangian (Penalty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\min x |x|_1$</td>
<td>$\min x \frac{1}{2} | Ax - b |_2^2$</td>
<td>$\min x \frac{1}{2} | Ax - b |_2^2 + \lambda | x |_1$</td>
</tr>
<tr>
<td>s.t. $\frac{1}{2} | Ax - b |_2^2 \leq \sigma$</td>
<td>s.t. $| x |_1 \leq \tau$</td>
<td></td>
</tr>
</tbody>
</table>

- BPDN: often most natural and transparent.
  (physical considerations guide $\sigma$)
- Lagrangian: ubiquitous in practice.
  (“no constraints”)

All three are (essentially) equivalent computationally!
The Prototypical Problem

Sparse Data Fitting:

Find sparse $x$ with $Ax \approx b$

Convex approaches: $\|x\|_1$ as a sparsity surrogate
(Candes-Tao-Donaho)

<table>
<thead>
<tr>
<th>BPDN</th>
<th>LASSO</th>
<th>Lagrangian (Penalty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\min_x |x|_1$</td>
<td>$\min_x \frac{1}{2} |Ax - b|_2^2$</td>
<td>$\min_x \frac{1}{2} |Ax - b|_2^2 + \lambda |x|_1$</td>
</tr>
<tr>
<td>s.t. $\frac{1}{2} |Ax - b|_2^2 \leq \sigma$</td>
<td>s.t. $|x|_1 \leq \tau$</td>
<td></td>
</tr>
</tbody>
</table>

- BPDN: often most natural and transparent.
  (physical considerations guide $\sigma$)
- Lagrangian: ubiquitous in practice.
  (“no constraints”)

All three are (essentially) equivalent computationally!

Basis for SPGL1 (van den Berg-Friedlander ’08)
Problem class: Solve

\[
\begin{align*}
\min_{x \in \mathcal{X}} & \quad \phi(x) \\
\text{s.t.} & \quad \rho(Ax - b) \leq \sigma
\end{align*}
\]
Optimal Value or Level Set Framework

Problem class: Solve

\[
\begin{align*}
\min_{x \in \mathcal{X}} & \quad \phi(x) \\
\text{s.t.} & \quad \rho(Ax - b) \leq \sigma
\end{align*}
\]

\(\mathcal{P}(\sigma)\)

Strategy: Consider the “flipped” problem

\[
\begin{align*}
v(\tau) & := \min_{x \in \mathcal{X}} \rho(Ax - b) \\
\text{s.t.} & \quad \phi(x) \leq \tau
\end{align*}
\]

\(\mathcal{Q}(\tau)\)
Problem class: Solve

\[
\begin{align*}
\min_{x \in \mathcal{X}} & \quad \phi(x) \\
\text{s.t.} & \quad \rho(Ax - b) \leq \sigma
\end{align*}
\]

Strategy: Consider the “flipped” problem

\[
\begin{align*}
v(\tau) := \min_{x \in \mathcal{X}} & \quad \rho(Ax - b) \\
\text{s.t.} & \quad \phi(x) \leq \tau
\end{align*}
\]

Then \(\text{opt-val}(\mathcal{P}(\sigma))\) is the minimal root of the equation

\[
v(\tau) = \sigma
\]
The intuition behind the proposed framework has a distinguished history, appearing even in antiquity. Perhaps the earliest instance is Queen Dido’s problem and the fabled origins of Carthage.

In short, the problem is to find the maximum area that can be enclosed by an arc of fixed length and a given line. The converse problem is to find an arc of least length that traps a fixed area between a line and the arc. Although these two problems reverse the objective and the constraint, the solution in each case is a semi-circle.
The intuition behind the proposed framework has a distinguished history, appearing even in antiquity. Perhaps the earliest instance is Queen Dido’s problem and the fabled origins of Carthage.

In short, the problem is to find the maximum area that can be enclosed by an arc of fixed length and a given line. The converse problem is to find an arc of least length that traps a fixed area between a line and the arc. Although these two problems reverse the objective and the constraint, the solution in each case is a semi-circle.

Other historical examples abound. More recently, these observations provide the basis for the Markowitz Mean-Variance Portfolio Theory.
Convex Sets
Let $C \subset \mathbb{R}^n$. We say that $C$ is convex if
\[(1 - \lambda)x + \lambda y \in C \text{ whenever } x, y \in C \text{ and } 0 \leq \lambda \leq 1.\]
The Role of Convexity

Convex Sets
Let $C \subset \mathbb{R}^n$. We say that $C$ is convex if
\[(1 - \lambda)x + \lambda y \in C \quad \text{whenever} \quad x, y \in C \quad \text{and} \quad 0 \leq \lambda \leq 1.\]

Convex Functions
Let $f : \mathbb{R}^n \to \mathbb{\bar{R}} := \mathbb{R} \cup \{+\infty\}$. We say that $f$ is convex if the set
\[\text{epi}(f) := \{(x, \mu) : f(x) \leq \mu\}\]
is a convex set.
Convex Sets
Let $C \subseteq \mathbb{R}^n$. We say that $C$ is convex if

$$(1 - \lambda)x + \lambda y \in C \text{ whenever } x, y \in C \text{ and } 0 \leq \lambda \leq 1.$$ 

Convex Functions
Let $f : \mathbb{R}^n \to \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. We say that $f$ is convex if the set

$$\text{epi}(f) := \{(x, \mu) : f(x) \leq \mu\}$$

is a convex set.

$$f((1 - \lambda)x_1 + \lambda x_2) \leq (1 - \lambda)f(x_1) + \lambda f(x_2)$$
Convex Functions

Convex indicator functions

Let $C \subset \mathbb{R}^n$. Then the function

$$\delta_C(x) := \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{if } x \notin C, \end{cases}$$

is a convex function.
Convex Functions

Convex indicator functions
Let $C \subset \mathbb{R}^n$. Then the function

$$
\delta_C(x) := \begin{cases} 
0, & \text{if } x \in C, \\
+\infty, & \text{if } x \notin C,
\end{cases}
$$

is a convex function.

Addition
Non-negative linear combinations of convex functions are convex: $f_i$ convex and $\lambda_i \geq 0, i = 1, \ldots, k$

$$
f(x) := \sum_{i=1}^{k} \lambda_i f_i(x).
$$
Convex Functions

Convex indicator functions
Let $C \subset \mathbb{R}^n$. Then the function

$$\delta_C(x) := \begin{cases} 
0 & \text{if } x \in C, \\
+\infty & \text{if } x \notin C,
\end{cases}$$

is a convex function.

Addition
Non-negative linear combinations of convex functions are convex: $f_i$ convex and $\lambda_i \geq 0, i = 1, \ldots, k$

$$f(x) := \sum_{i=1}^{k} \lambda_i f_i(x).$$

Infimal Projection
If $f : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ is convex, then so is

$$v(x) := \inf_y f(x, y),$$

since

$$\text{epi}(v) = \{ (x, \mu) : \exists y \in \text{s.t. } f(x, y) \leq \mu \}.$$
Convexity of $v$

When $\mathcal{X}$, $\rho$, and $\phi$ are convex, the optimal value function $v$ is a non-increasing convex function by infimal projection:

$$v(\tau) := \min_{x \in \mathcal{X}} \; \rho(Ax - b) \quad \text{s.t.} \quad \phi(x) \leq \tau$$

$$= \min_{x} \; \rho(Ax - b) + \delta_{\text{epi}(\phi)}(x, \tau) + \delta_{\mathcal{X}}(x)$$
Newton and Secant Methods

For $f$ convex and non-increasing, solve $f(\tau) = 0$. 

Newton and Secant Methods

For $f$ convex and non-increasing, solve $f(\tau) = 0$. 

![Graph showing the behavior of $f(\tau) = 0$.]
Newton and Secant Methods

For $f$ convex and non-increasing, solve $f(\tau) = 0$.

Problem: $f$ is often not differentiable.
Newton and Secant Methods

For $f$ convex and non-increasing, solve $f(\tau) = 0$.

Problem: $f$ is often not differentiable.

Use the convex subdifferential

$$\partial f(x) := \{ z : f(y) \geq f(x) + z^T(y - x) \quad \forall \ y \in \mathbb{R}^n \}$$
\[ \tau_* := \inf\{ \tau : f(\tau) \leq 0 \} \text{ and } g_* := \inf\{ g : g \in \partial f(\tau_*) \} < 0 \text{ (non-degeneracy)} \]
Superlinear Convergence

\[ \tau_* := \inf\{ \tau : f(\tau) \leq 0 \} \] and \[ g_* := \inf \{ g : g \in \partial f(\tau_*) \} < 0 \] (non-degeneracy)

Initialization: \[ \tau_{-1} < \tau_0 < \tau_* \]

\[ \tau_{k+1} := \begin{cases} \tau_k & \text{if } f(\tau_k) = 0, \\ \tau_k - \frac{f(\tau_k)}{g_k} & \text{[for } g_k \in \partial f(\tau_k)\text{]} \quad \text{otherwise;} \end{cases} \] (Newton)

and

\[ \tau_{k+1} := \begin{cases} \tau_k & \text{if } f(\tau_k) = 0, \\ \tau_k - \frac{\tau_k - \tau_{k-1}}{f(\tau_k) - f(\tau_{k-1})} f(\tau_k) & \text{otherwise.} \end{cases} \] (Secant)
Superlinear Convergence

\(\tau_* := \inf \{ \tau : f(\tau) \leq 0 \} \) and \(g_* := \inf \{ g : g \in \partial f(\tau_*) \} < 0\) (non-degeneracy)

Initialization: \(\tau_{-1} < \tau_0 < \tau_*\)

\[
\tau_{k+1} := \begin{cases} 
\tau_k & \text{if } f(\tau_k) = 0, \\
\tau_k - \frac{f(\tau_k)}{g_k} & \text{[for } g_k \in \partial f(\tau_k)\text{]} \text{ otherwise; (Newton)}
\end{cases}
\]

and

\[
\tau_{k+1} := \begin{cases} 
\tau_k & \text{if } f(\tau_k) = 0, \\
\tau_k - \frac{\tau_k - \tau_{k-1}}{f(\tau_k) - f(\tau_{k-1})} f(\tau_k) & \text{otherwise. (Secant)}
\end{cases}
\]

If either sequence terminates finitely at some \(\tau_k\), then \(\tau_k = \tau_*\); otherwise,

\[
|\tau_* - \tau_{k+1}| \leq (1 - \frac{g_*}{\gamma_k})|\tau_* - \tau_k|, \; k = 1, 2, \ldots,
\]

where \(\gamma_k = g_k\) (Newton) and \(\gamma_k \in \partial f(\tau_{k-1})\) (secant). In either case, \(\gamma_k \uparrow g_*\) and \(\tau_k \uparrow \tau_*\) globally \(q\)-superlinearly.
• **Problem:** Find root of the *inexactly known* convex function

\[ v(\cdot) - \sigma. \]
Inexact Root Finding

- **Problem:** Find root of the *inexactly known* convex function $v(\cdot) - \sigma$.

- Bisection is one approach
• **Problem:** Find root of the inexactely known convex function

\[ v(\cdot) - \sigma. \]

• Bisection is one approach
  • nonmonotone iterates (bad for warmstarts)
  • at best linear convergence (with perfect information)
• **Problem:** Find root of the *inexactly known* convex function 

\[ v(\cdot) - \sigma. \]

• **Bisection** is one approach
  • *nonmonotone* iterates (bad for warmstarts)
  • *at best* linear convergence (with perfect information)

• **Solution:**
  • modified secant
  • approximate Newton methods
Inexact Root Finding: Secant
Inexact Root Finding: Secant
Inexact Root Finding: Secant
Inexact Root Finding: Secant
Inexact Root Finding: Newton
Inexact Root Finding: Newton
Inexact Root Finding: Newton
Inexact Root Finding: Newton
Inexact Root Finding: Newton
Question: What precision guarantees convergence?
Question: What precision guarantees convergence?
Answer: We need $1 \leq \frac{u}{l} \leq \alpha$, where $\alpha \in [1, 2)$. 
**Question:** What precision guarantees convergence?

**Answer:** We need \( 1 \leq \frac{u}{l} \leq \alpha \), where \( \alpha \in [1, 2) \).

Then both algorithms return \( \bar{\tau} \) with \( \nu(\bar{\tau}) \leq \epsilon \) in

\[
O \left( \log_{2/\alpha} \left( \frac{C}{\epsilon} \right) \right) \text{ iterations}
\]
Inexact Root Finding: Convergence

**Question:** What precision guarantees convergence?

**Answer:** We need $1 \leq \frac{\nu}{\ell} \leq \alpha$, where $\alpha \in [1, 2)$.

Then both algorithms return $\bar{\tau}$ with $\nu(\bar{\tau}) \leq \epsilon$ in

$$O\left(\log_{2/\alpha}\left(\frac{C}{\epsilon}\right)\right)$$

iterations.

Key observation: $C = C(\tau_0)$ is independent of $\nu'(\tau^*)$. 
Minorants from Duality

\[ v(\tau) = \max_y \Phi(y, \tau) \]

\[ \bar{l} = \Phi(\bar{y}, \bar{\tau}) \]
$\nu(\tau) = \max_y \Phi(y, \tau)$

$\bar{s} \in \partial \Phi(\bar{y}, \cdot)(\bar{\tau}) \implies \bar{l} + \bar{s}(\tau - \bar{\tau}) \leq \Phi(\bar{y}, \tau) \leq \nu(\tau)$

$\bar{l} = \Phi(\bar{y}, \bar{\tau})$
Minorants from Duality

\[ v(\tau) = \max_y \Phi(y, \tau) \]

\[ \bar{s} \in \partial \Phi(\bar{y}, \cdot)(\bar{\tau}) \implies \bar{I} + \bar{s}(\tau - \bar{\tau}) \leq \Phi(\bar{y}, \tau) \leq v(\tau) \]

\[ \bar{I} = \Phi(\bar{y}, \bar{\tau}) \]

\[ L(\tau) = \bar{I} + \bar{s}(\tau - \bar{\tau}) \]
Robustness: $1 \leq u/l \leq \alpha$, where $\alpha \in [1, 2)$ and $\epsilon = 10^{-2}$

Figure: Inexact secant (top) and Newton (bottom) for $f_1(\tau) = (\tau - 1)^2 - 10$ (first two columns) and $f_2(\tau) = \tau^2$ (last column). Below each panel, $\alpha$ is the oracle accuracy, and $k$ is the number of iterations needed to converge, i.e., to reach $f_i(\tau_k) \leq \epsilon = 10^{-2}$.
Given a weighted graph $G = (V, E, d)$ find a realization:

$p_1, \ldots, p_n \in \mathbb{R}^2$ with $d_{ij} = \|p_i - p_j\|^2$ for all $ij \in E$. 
Sensor Network Localization (SNL)

SDP relaxation (Weinberger et al. ’04, Biswas et al. ’06):

\[
\begin{align*}
\text{max} & \quad \text{tr}(X) \\
\text{s.t.} & \quad \| P_E \mathcal{K}(X) - d \|_2^2 \leq \sigma \\
& \quad Xe = 0, \quad X \succeq 0
\end{align*}
\]

where \([\mathcal{K}(X)]_{i,j} = X_{ii} + X_{jj} - 2X_{ij}\).
Sensor Network Localization (SNL)

SDP relaxation (Weinberger et al. ’04, Biswas et al. ’06):

$$\begin{align*}
\max \ & \text{tr}(X) \\
\text{s.t.} \ & \|P_EK(X) - d\|_2^2 \leq \sigma \\
& Xe = 0, \quad X \succeq 0
\end{align*}$$

where $[K(X)]_{i,j} = X_{ii} + X_{jj} - 2X_{ij}$.

**Intuition:** $X = PP^T$ and then $\text{tr}(X) = \frac{1}{n+1} \sum_{i,j=1}^{n} ||p_i - p_j||^2$

with $p_i$ the $i$th row of $P$. 
Sensor Network Localization (SNL)

SDP relaxation (Weinberger et al. ’04, Biswas et al. ’06):

\[
\begin{align*}
\text{max} & \quad \text{tr} (X) \\
\text{s.t.} & \quad \|\mathcal{P}_{E}\mathcal{K}(X) - d\|_2^2 \leq \sigma \\
& \quad X e = 0, \quad X \succeq 0
\end{align*}
\]

where \( [\mathcal{K}(X)]_{i,j} = X_{ii} + X_{jj} - 2X_{ij} \).

Intuition: \( X = P P^T \) and then \( \text{tr} (X) = \frac{1}{n + 1} \sum_{i,j=1}^{n} \|p_i - p_j\|^2 \) with \( p_i \) the \( i \)th row of \( P \).

Flipped problem:

\[
\begin{align*}
\text{min} & \quad \|\mathcal{P}_{E}\mathcal{K}(X) - d\|_2^2 \\
\text{s.t.} & \quad \text{tr} \ X = \tau \\
& \quad X e = 0 \quad X \succeq 0.
\end{align*}
\]
Sensor Network Localization (SNL)

SDP relaxation (Weinberger et al. ’04, Biswas et al. ’06):

\[
\begin{align*}
\text{max} & \quad \text{tr} (X) \\
\text{s.t.} & \quad \| \mathcal{P} \mathcal{K}(X) - d \|_2^2 \leq \sigma \\
& \quad X e = 0, \quad X \succeq 0
\end{align*}
\]

where \([\mathcal{K}(X)]_{i,j} = X_{ii} + X_{jj} - 2X_{ij}\).

Intuition: \(X = PP^T\) and then \(\text{tr} (X) = \frac{1}{n+1} \sum_{i,j=1}^{n} \| p_i - p_j \|^2\) with \(p_i\) the \(i\)th row of \(P\).

Flipped problem:

\[
\begin{align*}
\text{min} & \quad \| \mathcal{P} \mathcal{K}(X) - d \|_2^2 \\
\text{s.t.} & \quad \text{tr} X = \tau \\
& \quad X e = 0 \quad X \succeq 0.
\end{align*}
\]

- Perfectly adapted for the Frank-Wolfe method.
Sensor Network Localization (SNL)

SDP relaxation (Weinberger et al. ’04, Biswas et al. ’06):

\[
\begin{align*}
\text{max } & \quad \text{tr } (X) \\
\text{s.t. } & \quad \| \mathcal{P}_E \mathcal{K}(X) - d \|_2^2 \leq \sigma \\
& \quad Xe = 0, \quad X \succeq 0
\end{align*}
\]

where \([\mathcal{K}(X)]_{i,j} = X_{ii} + X_{jj} - 2X_{ij}\).

**Intuition:** \(X = PP^T\) and then \(\text{tr } (X) = \frac{1}{n+1} \sum_{i,j=1}^{n} \| p_i - p_j \|^2\) with \(p_i\) the \(i\)th row of \(P\).

**Flipped problem:**

\[
\begin{align*}
\text{min } & \quad \| \mathcal{P}_E \mathcal{K}(X) - d \|_2^2 \\
\text{s.t. } & \quad \text{tr } X = \tau \\
& \quad Xe = 0 \quad X \succeq 0.
\end{align*}
\]

- Perfectly adapted for the Frank-Wolfe method.

**Key point:** Slater failing (always the case) is irrelevant.
Approximate Newton

Figure: $\sigma = 0.25$
Figure: $\sigma = 0.25$

Figure: $\sigma = 0$
Max-trace

Newton_max

Refine positions

Newton_min

Refine positions
Observations

- Simple strategy for optimizing over complex domains
- Rigorous convergence guarantees
- Insensitivity to ill-conditioning
- Many applications
  - Sensor Network Localization
    (Drusvyatskiy-Krislock-Voronin-Wolkowicz ’15)
  - Sparse/Robust Estimation and Kalman Smoothing
    (Aravkin-B-Pillonetto ’13)
  - Large scale SDP and LP (cf. Renegar ’14)
  - Chromosome reconstruction
    (Aravkin-Becker-Drusvyatskiy-Lozano ’15)
  - Phase retrieval (Aravkin-B-Drusvyatskiy-Friedlander-Roy ’16)
  - Generalized linear models
    (Aravkin-B-Drusvyatskiy-Friedlander-Roy ’16)
- ...
Conjugate Functions and Duality

Convex Indicator
For any convex set $C$, the convex indicator function for $C$ is

$$\delta(x \mid C) := \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C. \end{cases}$$
Conjugate Functions and Duality

Convex Indicator
For any convex set $C$, the convex indicator function for $C$ is

$$\delta (x \mid C) := \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C. \end{cases}$$

Support Functionals
For any set $C$, the support functional for $C$ is

$$\delta^* (x \mid C) := \sup_{z \in C} \langle x, z \rangle .$$
Conjugate Functions and Duality

Convex Indicator
For any convex set $C$, the convex indicator function for $C$ is

$$\delta (x \mid C) := \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C. \end{cases}$$

Support Functionals
For any set $C$, the support functional for $C$ is

$$\delta^* (x \mid C) := \sup_{z \in C} \langle x, z \rangle .$$

Convex Conjugates
For any convex function $g(x)$, the convex conjugate is given by

$$g^*(y) := \delta^* ((y, -1) \mid \text{epi}(g)) = \sup_x [\langle x, y \rangle - g(x)] .$$
Conjugate’s and the Subdifferential

\[ g^*(y) = \sup_x [\langle x, y \rangle - g(x)] . \]
Conjugate’s and the Subdifferential

\[ g^*(y) = \sup_x [\langle x, y \rangle - g(x)] . \]

**The Bi-Conjugate Theorem**

If \( \text{epi}(g) \) is closed and \( \text{dom}(g) \neq \emptyset \), then \( (g^*)^* = g \).
Conjugate’s and the Subdifferential

\[ g^*(y) = \sup_{x} [\langle x, y \rangle - g(x)] . \]

The Bi-Conjugate Theorem
If \( \text{epi}(g) \) is closed and \( \text{dom}(g) \neq \emptyset \), then \( (g^*)^* = g \).

The Young-Fenchel Inequality
\[ g(x) + g^*(z) \geq \langle z, x \rangle \quad \text{for all} \quad x, y \in \mathbb{R}^n \quad \text{with equality if and only if} \]
\[ z \in \partial g(x) \quad \text{and} \quad x \in \partial g^*(z). \]
In particular, \( \partial g(x) = \arg \max_z [\langle z, x \rangle - g^*(z)] \).
Conjugate’s and the Subdifferential

\[ g^*(y) = \sup_x [\langle x, y \rangle - g(x)] . \]

The Bi-Conjugate Theorem
If \( \text{epi}(g) \) is closed and \( \text{dom}(g) \neq \emptyset \), then \( (g^*)^* = g \).

The Young-Fenchel Inequality
\( g(x) + g^*(z) \geq \langle z, x \rangle \) for all \( x, y \in \mathbb{R}^n \) with equality if and only if \( z \in \partial g(x) \) and \( x \in \partial g^*(z) \).

In particular, \( \partial g(x) = \arg\max_z [\langle z, x \rangle - g^*(z)] \).

Maximal Montone Operator
If \( \text{epi}(g) \) is closed and \( \text{dom}(g) \neq \emptyset \), then \( \partial g \) is a maximal monotone operator with \( \partial g^{-1} = \partial g^* \).
Conjugate’s and the Subdifferential

\[ g^*(y) = \sup_x [\langle x, y \rangle - g(x)] . \]

**The Bi-Conjugate Theorem**

If \( \text{epi}(g) \) is closed and \( \text{dom}(g) \neq \emptyset \), then \( (g^*)^* = g \).

**The Young-Fenchel Inequality**

\[ g(x) + g^*(z) \geq \langle z, x \rangle \text{ for all } x, y \in \mathbb{R}^n \text{ with equality if and only if } z \in \partial g(x) \text{ and } x \in \partial g^*(z). \]

In particular, \( \partial g(x) = \text{argmax}_z [\langle z, x \rangle - g^*(z)] \).

**Maximal Montone Operator**

If \( \text{epi}(g) \) is closed and \( \text{dom}(g) \neq \emptyset \), then \( \partial g \) is a maximal monotone operator with \( \partial g^{-1} = \partial g^* \).

**Note:** The lsc hull of \( g \) is \( \text{cl} \ g := g^{**} \).
The perspective function

$$\text{epi} (g^\pi) := \text{cl cone} (\text{epi} (g)) = \text{cl} \left( \bigcup_{\lambda > 0} \lambda \text{epi} (g) \right)$$
The perspective function

\[ \text{epi} (g^\pi) := \text{cl cone} (\text{epi} (g)) = \text{cl} (\bigcup_{\lambda > 0} \lambda \text{epi} (g)) \]

\[ g^\pi (z, \lambda) := \begin{cases} 
\lambda g(\lambda^{-1} z) & \text{if } \lambda > 0, \\
g^\infty (z) & \text{if } \lambda = 0, \\
+\infty & \text{if } \lambda < 0,
\end{cases} \]

where \( g^\infty \) is the *horizon* function of \( g \):

\[ g^\infty (z) := \sup_{x \in \text{dom} g} \left[ g(x + z) - g(x) \right]. \]
Support Functions for $\text{epi}(g)$ and $\text{lev}_g(\tau)$

$g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be closed proper and convex.

Then

$$\delta^* ((y, \mu) \mid \text{epi}(g)) = (g^*)^\pi(y, -\mu)$$

and

$$\delta^* (y \mid [g \leq \tau]) = \text{cl inf}_{\mu \geq 0} \left[ \tau \mu + (g^*)^\pi(y, \mu) \right],$$

where

$$\text{epi}(g) := \{(x, \mu) \mid g(x) \leq \mu\}$$

$$[g \leq \tau] := \{x \mid g(x) \leq \tau\}$$

$$\delta^* (z \mid C) := \sup_{w \in C} \langle z, w \rangle$$
The perturbation function

\[ f(x, b, \tau) := \rho(b - Ax) + \delta((x, \tau) | \text{epi}(\phi)) \]

Its conjugate

\[ f^*(y, u, \mu) = (\phi^*)^\pi(y + A^T u, -\mu) + \rho^*(u) . \]
The perturbation function
\[ f(x, b, \tau) := \rho(b - Ax) + \delta((x, \tau) \mid \text{epi}(\phi)) \]

Its conjugate
\[ f^*(y, u, \mu) = (\phi^*)^\pi(y + A^T u, -\mu) + \rho^*(u) . \]

The Primal Problem infimal projection in \( x \)
\[ \mathcal{P}(b, \tau) : \quad v(b, \tau) := \min_x f(x, b, \tau) . \]
The perturbation function

\[ f(x, b, \tau) := \rho(b - Ax) + \delta((x, \tau) \mid \text{epi}(\phi)) \]

Its conjugate

\[ f^*(y, u, \mu) = (\phi^*)^\pi(y + A^T u, -\mu) + \rho^*(u). \]

The Primal Problem

\[ \mathcal{P}(b, \tau) : \quad v(b, \tau) := \min_x f(x, b, \tau). \]

The Dual Problem

\[ \mathcal{D}(b, \tau) : \quad \hat{v}(b, \tau) := \sup_{u, \mu} \langle b, u \rangle + \tau \mu - f^*(0, u, \mu) \]

(reduced dual)

\[ = \sup_u \langle b, u \rangle - \rho^*(u) - \delta^* (A^T u \mid \phi \leq \tau). \]
The perturbation function

$$f(x, b, \tau) := \rho(b - Ax) + \delta((x, \tau) \mid \text{epi}(\phi))$$

Its conjugate

$$f^*(y, u, \mu) = (\phi^*)^\pi(y + A^T u, -\mu) + \rho^*(u).$$

The Primal Problem

$$\mathcal{P}(b, \tau) : \quad v(b, \tau) := \min_x f(x, b, \tau).$$

The Dual Problem

$$\mathcal{D}(b, \tau) : \quad \hat{v}(b, \tau) := \sup_{u, \mu} \langle b, u \rangle + \tau \mu - f^*(0, u, \mu)$$

(reduced dual)

$$= \sup_u \langle b, u \rangle - \rho^*(u) - \delta^* (A^T u \mid [\phi \leq \tau]).$$

The Subdifferential: If $(b, \tau) \in \text{int}(\text{dom} v)$, then $v(b, \tau) = \hat{v}(b, \tau)$ and

$$\emptyset \neq \partial v(b, \tau) = \argmax_{u, \mu} \mathcal{D}(b, \tau).$$
Piecewise Linear-Quadratic Penalties

\[ \phi(x) := \sup_{u \in U} [\langle x, u \rangle - \frac{1}{2} u^T B u] \]

\( U \subset \mathbb{R}^n \) is nonempty, closed and convex with \( 0 \in U \) (not nec. poly.)
\( B \in \mathbb{R}^{n \times n} \) is symmetric positive semi-definite.

**Examples:**

1. Support functionals: \( B = 0 \)

2. Gauge functionals: \( \gamma(\cdot \mid U^\circ) = \delta^*(\cdot \mid U) \)

3. Norms: \( B = \text{closed unit ball}, \|\cdot\| = \gamma(\cdot \mid B) \)

4. Least-squares: \( U = \mathbb{R}^n, B = I \)

5. Huber: \( U = [-\epsilon, \epsilon]^n, B = I \)
PLQ Densities: Gauss, Laplace, Huber, Vapnik

\[
V(x) = \frac{1}{2} x^2
\]

Gauss

\[
V(x) = |x|
\]

\[
V(x) = \begin{cases} 
-Kx - \frac{1}{2}K^2; & x < -K \\
\frac{1}{2}x^2; & -K \leq x \leq K \\
Kx - \frac{1}{2}K^2; & K < x
\end{cases}
\]

Huber

\[
V(x) = \begin{cases} 
-x - \epsilon; & x < -\epsilon \\
0; & -\epsilon \leq x \leq \epsilon \\
x - \epsilon; & \epsilon \leq x
\end{cases}
\]

Vapnik
Computing $v'$ for PLQ Penalties $\phi$

$$\phi(x) := \sup_{u \in U} [\langle x, u \rangle - \frac{1}{2}u^T Bu]$$

$$\mathcal{P}(b, \tau) : \quad v(b, \tau) := \min \rho(b - Ax) \quad \text{st} \quad \phi(x) \leq \tau$$

$$\partial v(b, \tau) = \left\{ \left( \begin{array}{c} \bar{u} \\ -\bar{\mu} \end{array} \right) \middle| \exists x \text{ s.t. } 0 \in -A^T \partial \rho(b - Ax) + \bar{\mu}^+ \partial \phi(x) \text{ and } \bar{\mu} = \max \left\{ \gamma \left( A^T \bar{u} \mid U \right), \sqrt{\bar{u}^T ABA^T \bar{u}} / \sqrt{2\tau} \right\} \right\}.$$
A Few Special Cases

\[ v(\tau) := \min \frac{1}{2} \| b - Ax \|_2^2 \quad \text{st} \quad \phi(x) \leq \tau \]

Optimal Solution: \( \bar{x} \) \quad \text{Optimal Residual:} \( \bar{r} := A\bar{x} - b \)

1. **Support functionals:** \( \phi(x) = \delta^* (x \mid U), \ 0 \in U \implies \)
   \[ v'(\tau) = -\delta^* (A^T \bar{r} \mid U^\circ) = -\gamma (A^T \bar{r} \mid U) \]

2. **Gauge functionals:** \( \phi(x) = \gamma (x \mid U), \ 0 \in U \implies \)
   \[ v'(\tau) = -\gamma (A^T \bar{r} \mid U^\circ) = -\delta^* (A^T \bar{r} \mid U) \]

3. **Norms:** \( \phi(x) = \|x\| \implies v'(\tau) = -\|A^T \bar{r}\|_* \)

4. **Huber:** \( \phi(x) = \sup_{u \in [-\epsilon, \epsilon]^n} [\langle x, u \rangle - \frac{1}{2} u^T u] \implies \)
   \[ v'(\tau) = -\max\{\epsilon \|A^T \bar{r}\|_\infty, \|A^T \bar{r}\|_2 / \sqrt{2\tau}\} \]

5. **Vapnik:** \( \phi(x) = \|(x - \epsilon)_+\|_1 + \|(-x - \epsilon)_+\|_1 \implies \)
   \[ v'(\tau) = -\left(\|A^T \bar{r}\|_\infty + \epsilon \|A^T \bar{r}\|_2\right) \]
Basis Pursuit with Outliers

$$\text{BP}_\sigma: \min \|x\|_1 \text{ st } \rho(b - Ax) \leq \sigma$$

Standard least-squares: $$\rho(z) = \|z\|_2$$ or $$\rho(z) = \|z\|_2^2.$$
Basis Pursuit with Outliers

\[ \text{BP}_{\sigma}: \min \| x \|_1 \text{ st } \rho(b - Ax) \leq \sigma \]

Standard least-squares: \( \rho(z) = \| z \|_2 \) or \( \rho(z) = \| z \|_2^2 \).

Quantile Huber:

\[
\rho_{\kappa, \tau}(r) = \begin{cases} 
\tau |r| - \frac{\kappa \tau^2}{2} & \text{if } r < -\tau \kappa, \\
\frac{1}{2\kappa} r^2 & \text{if } r \in [-\kappa \tau, (1 - \tau) \kappa], \\
(1 - \tau) |r| - \frac{\kappa (1-\tau)^2}{2} & \text{if } r > (1 - \tau) \kappa.
\end{cases}
\]

Standard Huber when \( \tau = 0.5 \).
Huber
Sparse and Robust Formulation

\[ \text{HBP}_\sigma: \min \|x\|_1 \ \text{st} \ \rho(b - Ax) \leq \sigma \]

Problem Specification

- \( x \) 20-sparse spike train in \( \mathbb{R}^{512} \)
- \( b \) measurements in \( \mathbb{R}^{120} \)
- \( A \) Measurement matrix satisfying RIP
- \( \rho \) Huber function
- \( \sigma \) error level set at .01
- 5 outliers

Results

In the presence of outliers, the robust formulation recovers the spike train, while the standard formulation does not.
Sparse and Robust Formulation

\[ \text{HBP}_\sigma: \min_{0 \leq x} \|x\|_1 \ \text{st} \ \rho(b - Ax) \leq \sigma \]

Problem Specification

- \( x \) 20-sparse spike train in \( \mathbb{R}^{512} \)
- \( b \) measurements in \( \mathbb{R}^{120} \)
- \( A \) Measurement matrix satisfying RIP
- \( \rho \) Huber function
- \( \sigma \) error level set at .01
- 5 outliers

Results

In the presence of outliers, the robust formulation recovers the spike train, while the standard formulation does not.
References

• “Probing the pareto frontier for basis pursuit solutions”
  van der Berg - Friedlander

• “Sparse optimization with least-squares constraints”
  van der Berg - Friedlander

• “Variational Properties of Value Functions.”
  Aravkin - B - Friedlander
  SIOPT 23(2013), 1689–1717.

• “Level-set methods for convex optimization”
  Aravkin - B - Drusvyatskiy - Friedlander - Roy
  Preprint, 2016