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:

$$\boxed{\text{Find sparse } x \text{ with } Ax \approx b}$$

The Prototypical Problem

Sparse Data Fitting:

$$\boxed{\text{Find sparse } x \text{ with } Ax \approx b}$$

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 $y_i \approx a_i^T x$.

The Prototypical Problem

Sparse Data Fitting:

$$\boxed{\text{Find sparse } x \text{ with } Ax \approx b}$$

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 $m << n$, one can "always" find $\overline{x}$ such that

$$y_i = a_i^T x, \ \ i = 1, \ldots, m.$$

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 $m << n$, one can "always" find $\overline{x}$ such that

$$y_i = a_i^T x, \;\; i = 1, \ldots, m.$$

This $\overline{x}$ 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:

$$\boxed{\text{Find sparse } x \text{ with } Ax \approx b}$$

There are numerous other applications;

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

The Prototypical Problem

Sparse Data Fitting:

$$\boxed{\text{Find sparse } x \text{ with } Ax \approx b}$$

The Prototypical Problem

Sparse Data Fitting:

Find sparse x with $Ax \approx b$

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

BPDN	LASSO	Lagrangian (Penalty)
$\min_x \|x\|_1$ s.t. $\frac{1}{2}\|Ax-b\|_2^2 \leq \sigma$	$\min_x \frac{1}{2}\|Ax-b\|_2^2$ s.t. $\|x\|_1 \leq \tau$	$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$

The Prototypical Problem

Sparse Data Fitting:

Find sparse x with $Ax \approx b$

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

BPDN	LASSO	Lagrangian (Penalty)
$\min_x \|x\|_1$ s.t. $\frac{1}{2}\|Ax-b\|_2^2 \leq \sigma$	$\min_x \frac{1}{2}\|Ax-b\|_2^2$ s.t. $\|x\|_1 \leq \tau$	$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$

- BPDN: often most natural and transparent.
  (physical considerations guide σ)

The Prototypical Problem

Sparse Data Fitting:

Find sparse x with $Ax \approx b$

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

BPDN	LASSO	Lagrangian (Penalty)
$\min_x \|x\|_1$ s.t. $\frac{1}{2}\|Ax - b\|_2^2 \leq \sigma$	$\min_x \frac{1}{2}\|Ax - b\|_2^2$ s.t. $\|x\|_1 \leq \tau$	$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$

- BPDN: often most natural and transparent.
  (physical considerations guide σ)
- Lagrangian: 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)

BPDN	LASSO	Lagrangian (Penalty)
$\min_x \|x\|_1$ s.t. $\frac{1}{2}\|Ax-b\|_2^2 \le \sigma$	$\min_x \frac{1}{2}\|Ax-b\|_2^2$ s.t. $\|x\|_1 \le \tau$	$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$

- BPDN: often most natural and transparent.
  (physical considerations guide σ)
- 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 surragate (Candes-Tao-Donaho)

BPDN	LASSO	Lagrangian (Penalty)
$\min_x \|x\|_1$ s.t. $\frac{1}{2}\|Ax-b\|_2^2 \le \sigma$	$\min_x \frac{1}{2}\|Ax-b\|_2^2$ s.t. $\|x\|_1 \le \tau$	$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$

- BPDN: often most natural and transparent. (physical considerations guide σ)
- Lagrangian: ubiquitous in practice. ("no constraints")

All three are (essentially) equivalent computationally!

Basis for SPGL1 (van den Berg-Friedlander '08)

Optimal Value or Level Set Framework

Problem class: Solve

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \phi(x) \\ \text{s.t.} \quad & \rho(Ax - b) \leq \sigma \end{aligned} \qquad \mathcal{P}(\sigma)$$

Optimal Value or Level Set Framework

Problem class: Solve

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \phi(x) \\ \text{s.t.} \quad & \rho(Ax - b) \le \sigma \end{aligned} \qquad \mathcal{P}(\sigma)$$

Strategy: Consider the “flipped” problem

$$\begin{aligned} v(\tau) := \min_{x \in \mathcal{X}} \quad & \rho(Ax - b) \\ \text{s.t.} \quad & \phi(x) \le \tau \end{aligned} \qquad \mathcal{Q}(\tau)$$

Optimal Value or Level Set Framework

Problem class: Solve

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \phi(x) \\ \text{s.t.} \quad & \rho(Ax - b) \le \sigma \end{aligned} \qquad \mathcal{P}(\sigma)$$

Strategy: Consider the "flipped" problem

$$\begin{aligned} v(\tau) := \min_{x \in \mathcal{X}} \quad & \rho(Ax - b) \\ \text{s.t.} \quad & \phi(x) \le \tau \end{aligned} \qquad \mathcal{Q}(\tau)$$

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

$$\boxed{v(\tau) = \sigma}$$

Queen Dido's Problem

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**.

The Role of Convexity

Convex Sets

Let $C \subset \mathbb{R}^n$. We say that C is convex if

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

Convex Sets

Let $C \subset \mathbb{R}^n$. We say that C is convex if

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

Convex Functions

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

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

is a convex set.

The Role of Convexity

Convex Sets

Let $C \subset \mathbb{R}^n$. We say that C is convex if

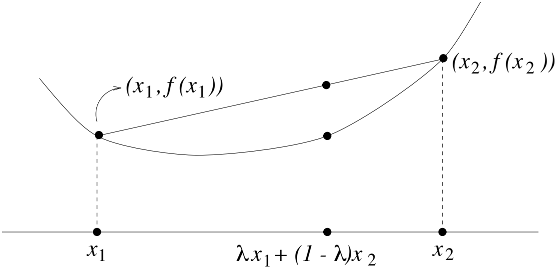
$$(1-\lambda)x + \lambda y \in C \text{ whenever } x, y \in C \text{ and } 0 \leq \lambda \leq 1.$$

Convex Functions

Let $f : \mathbb{R}^n \to \bar{R} := \mathbf{R} \cup \{+\infty\}$. 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 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 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 $\lambda_i \geq 0, i = 1, \ldots, k$

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

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 $f : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbf{R}}$ is convex, then so is

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

since

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

Convexity of v

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

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

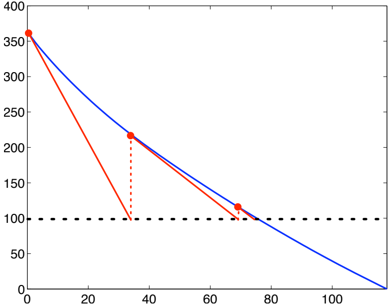
$$= \min_{x} \quad \rho(Ax - b) + \delta_{\text{epi}\,(\phi)}(x, \tau) + \delta_{\mathcal{X}}(x)$$

Newton and Secant Methods

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

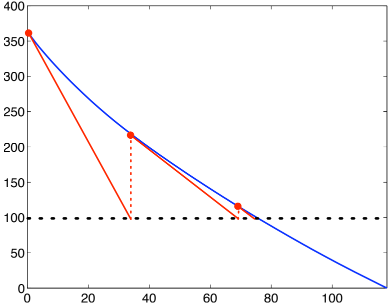
Newton and Secant Methods

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



Newton and Secant Methods

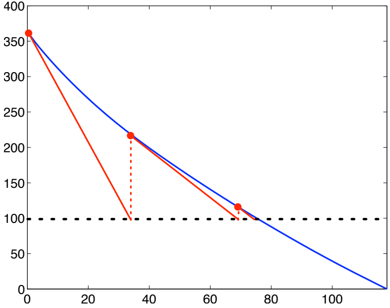
For f convex and non-increasing, solve $f(\tau) = 0$.



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

$$\partial f(x) := \{\, z \,:\, f(y) \geq f(x) + z^T(y - x) \quad \forall\ y \in \mathbb{R}^n \,\}$$

Superlinear Convergence

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

Superlinear Convergence

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

Initialization: $\tau_{-1} < \tau_0 < \tau_*$

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

and

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

Superlinear Convergence

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

Initialization: $\tau_{-1} < \tau_0 < \tau_*$

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

and

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

If either sequence terminates finitely at some τ_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.

Inexact Root Finding

- 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

Inexact Root Finding

- 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)

Inexact Root Finding

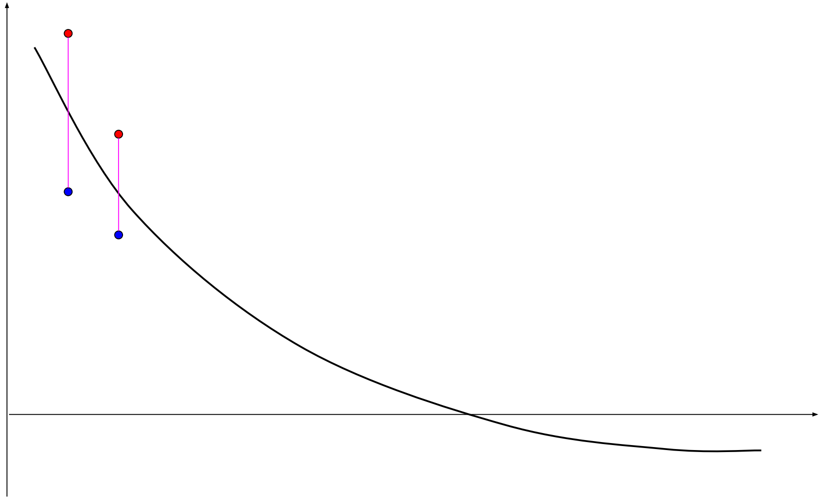
- Problem: Find root of the inexactly known convex function

$$v(\cdot) - \sigma.$$

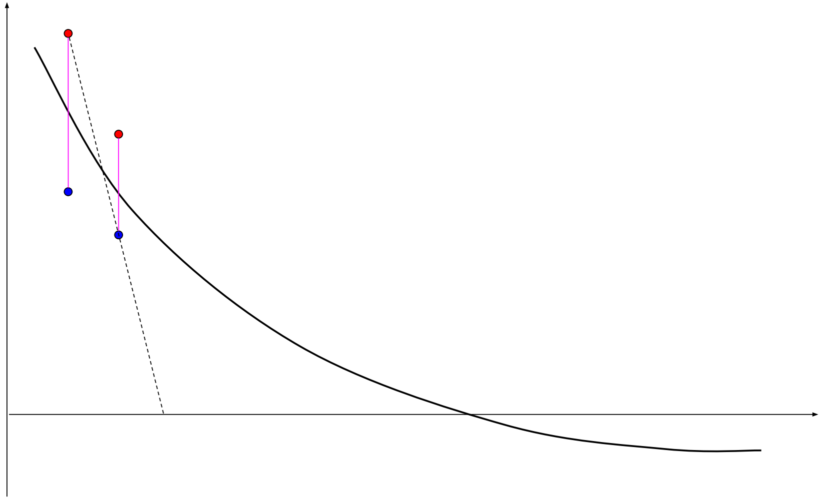
- Bisection is one approach
  - nonmonotone iterates (bad for warmstarts)
  - at best linear convergence (with perfect information)

- Solution:
  - modified secant
  - approximate Newton methods

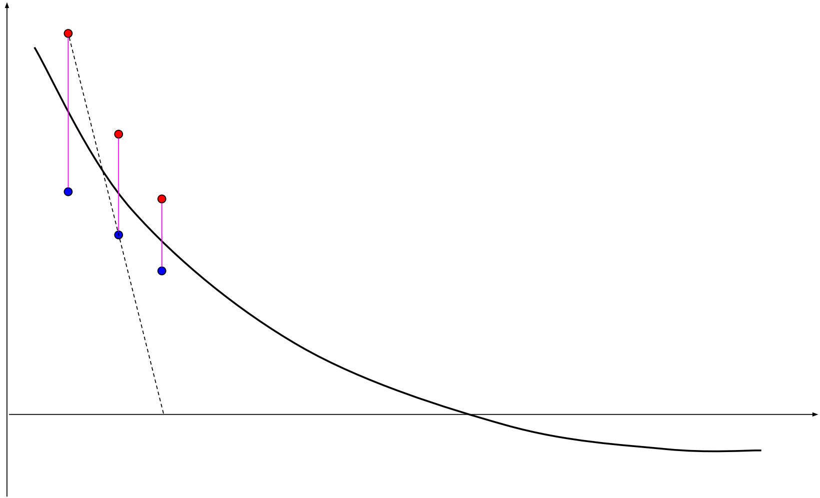
Inexact Root Finding: Secant



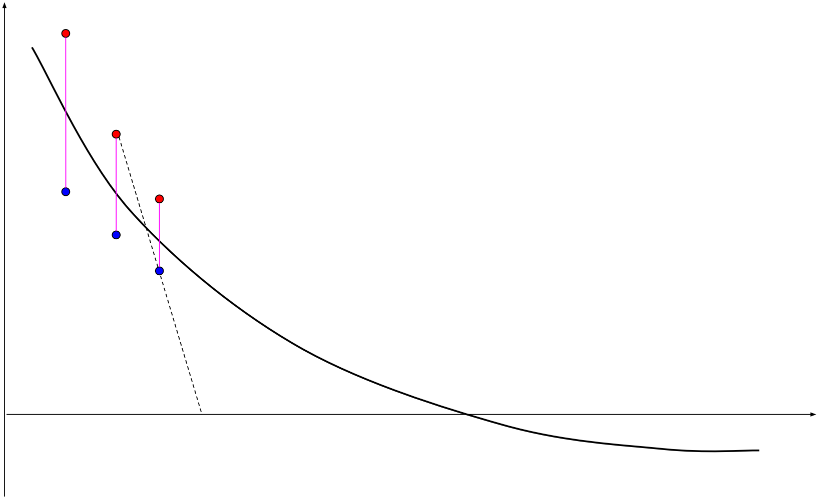
Inexact Root Finding: Secant



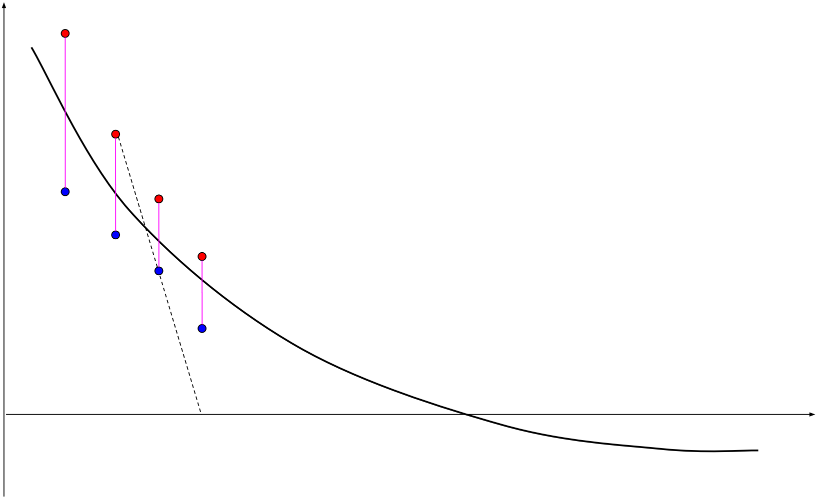
Inexact Root Finding: Secant



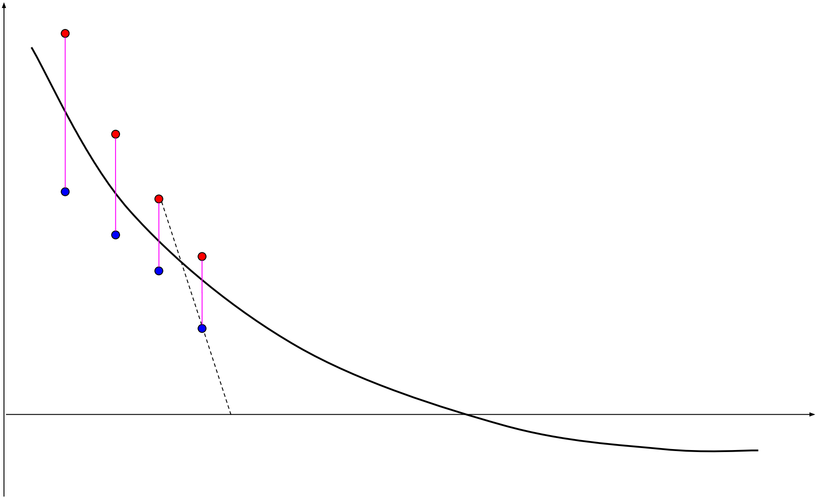
Inexact Root Finding: Secant



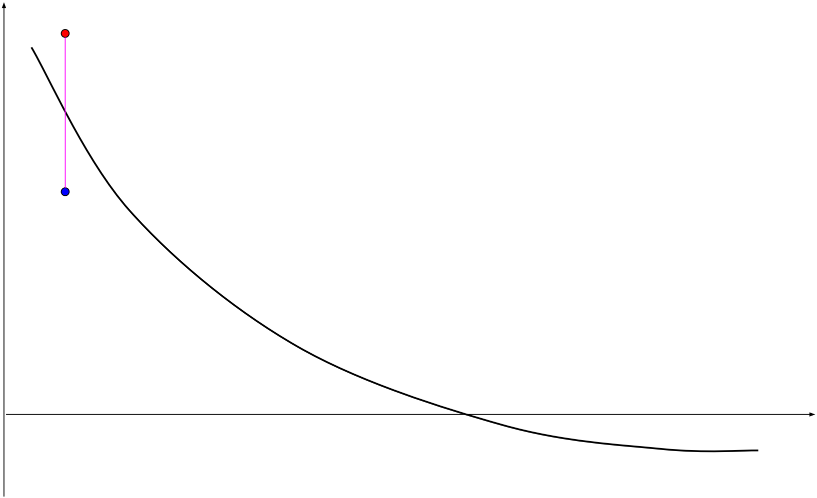
Inexact Root Finding: Secant



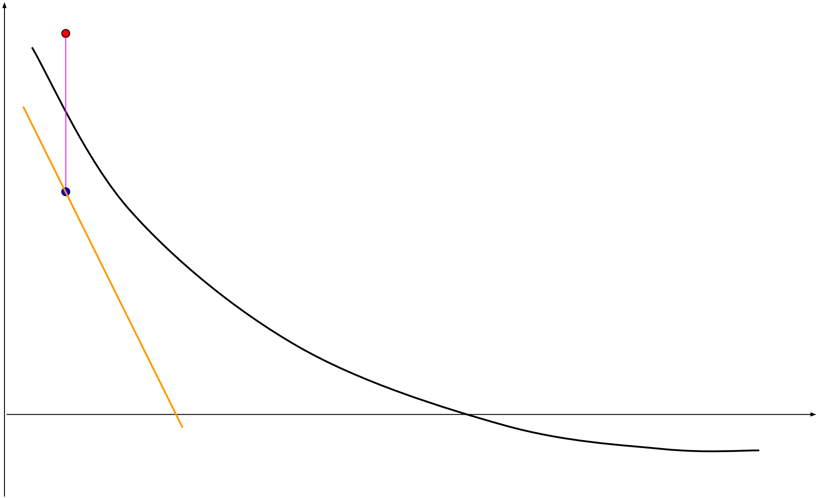
Inexact Root Finding: Secant



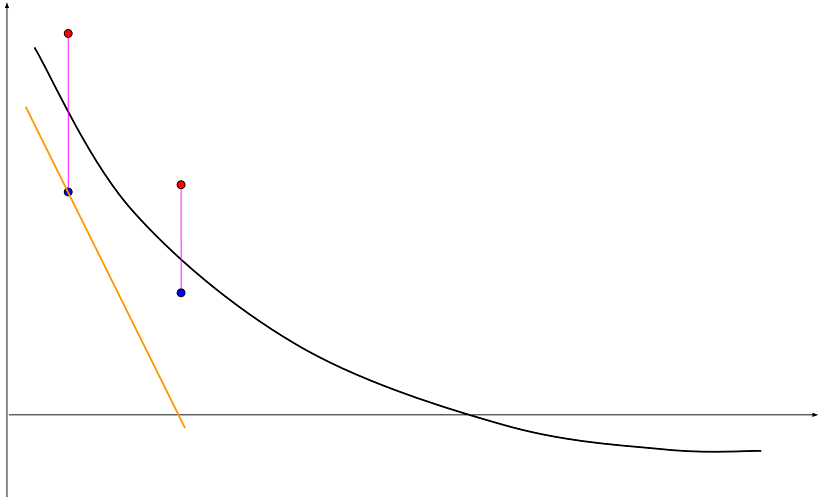
Inexact Root Finding: Newton



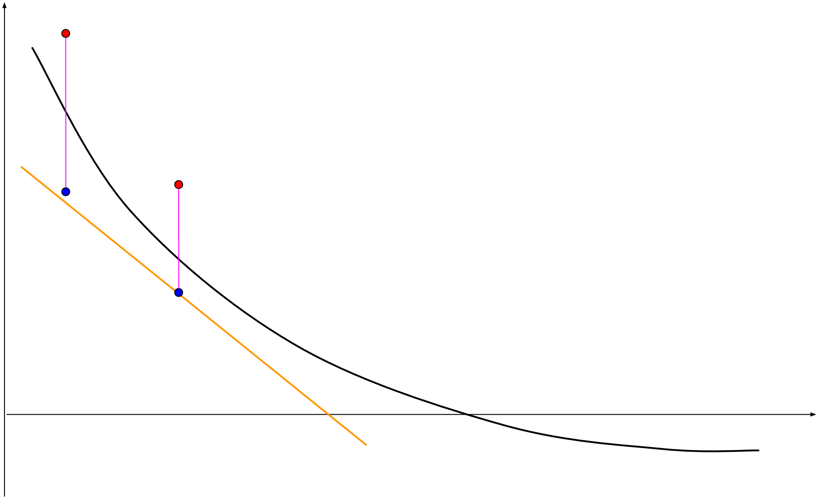
Inexact Root Finding: Newton



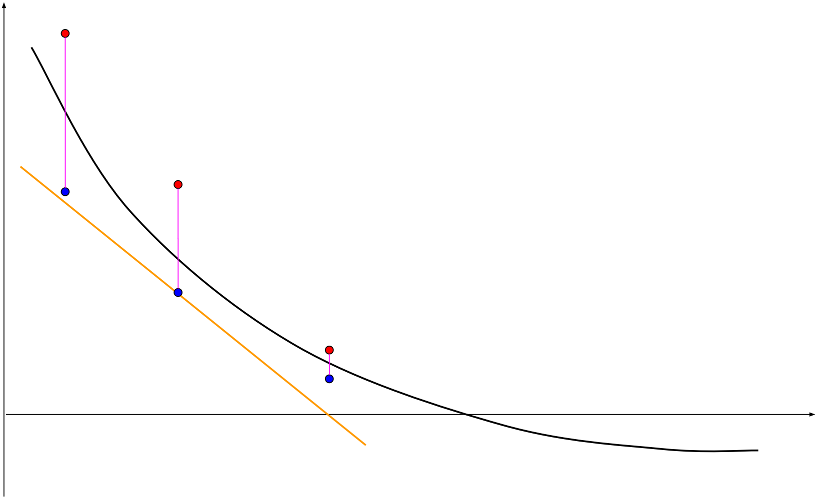
Inexact Root Finding: Newton



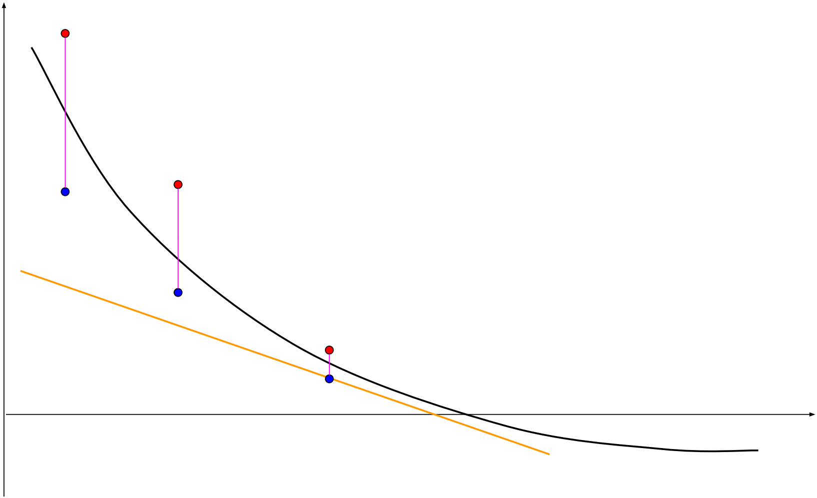
Inexact Root Finding: Newton



Inexact Root Finding: Newton

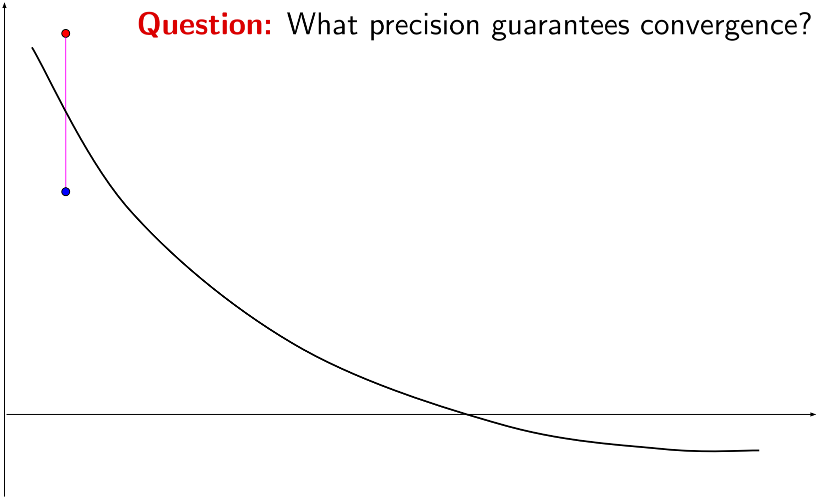


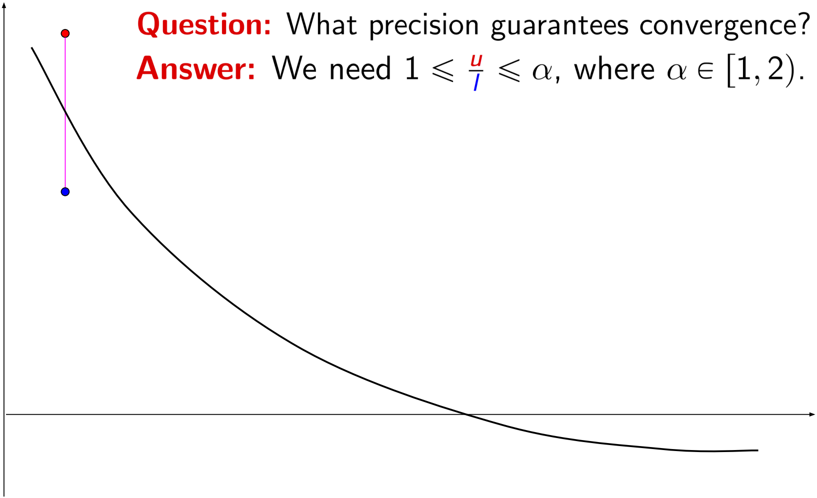
Inexact Root Finding: Newton



Inexact Root Finding: Convergence

Question: What precision guarantees convergence?

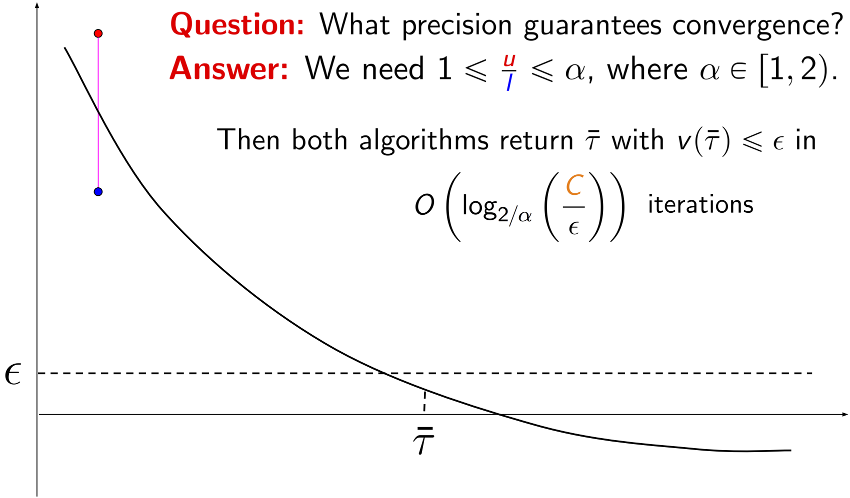




Question: What precision guarantees convergence?

Answer: We need $1 \leqslant \frac{u}{l} \leqslant \alpha$, where $\alpha \in [1, 2)$.

Inexact Root Finding: Convergence



Question: What precision guarantees convergence?

Answer: We need $1 \leqslant \frac{u}{l} \leqslant \alpha$, where $\alpha \in [1, 2)$.

Then both algorithms return $\bar{\tau}$ with $v(\bar{\tau}) \leqslant \epsilon$ in

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

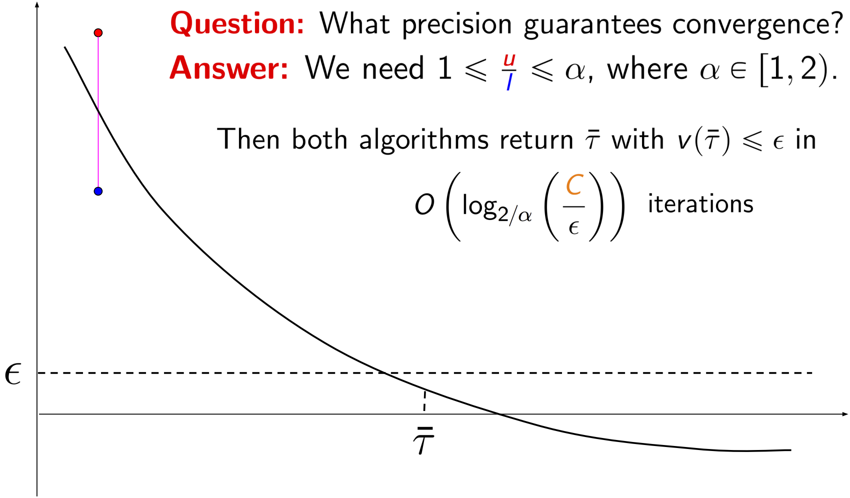
Inexact Root Finding: Convergence

Question: What precision guarantees convergence?

Answer: We need $1 \leqslant \frac{u}{l} \leqslant \alpha$, where $\alpha \in [1, 2)$.

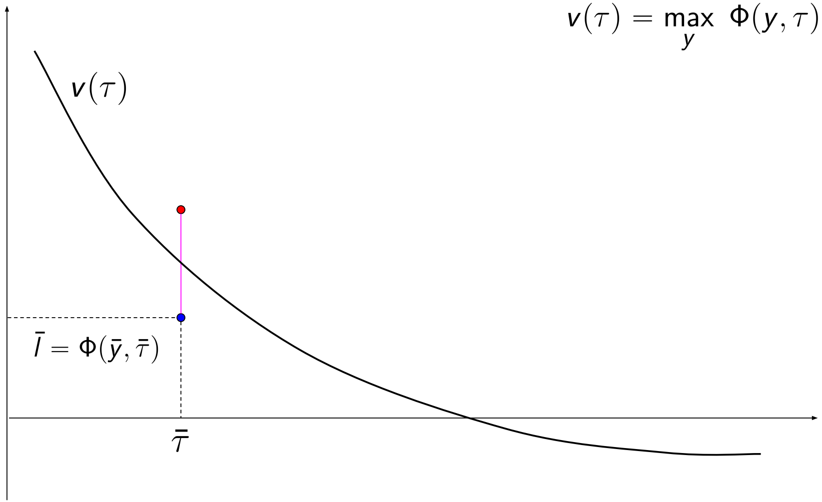
Then both algorithms return $\bar{\tau}$ with $v(\bar{\tau}) \leqslant \epsilon$ in

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



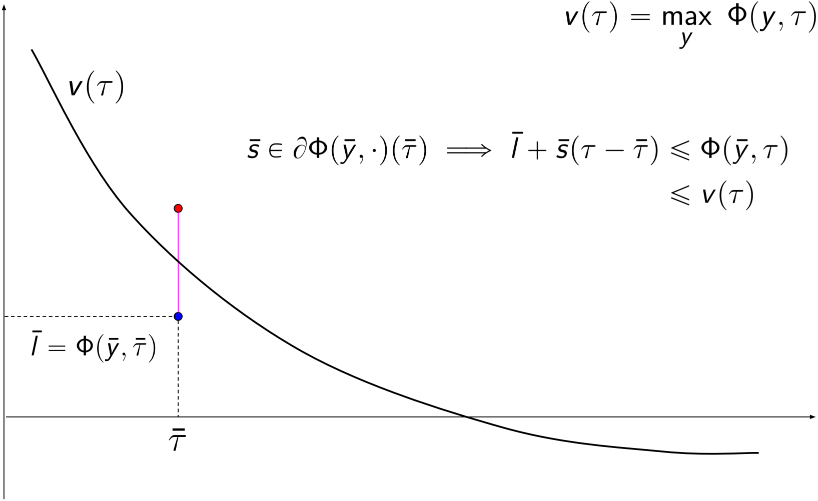
Key observation: $C = C(\tau_0)$ is independent of $v'(\tau^*)$.

Minorants from Duality



Minorants from Duality

$$v(\tau) = \max_y \; \Phi(y, \tau)$$

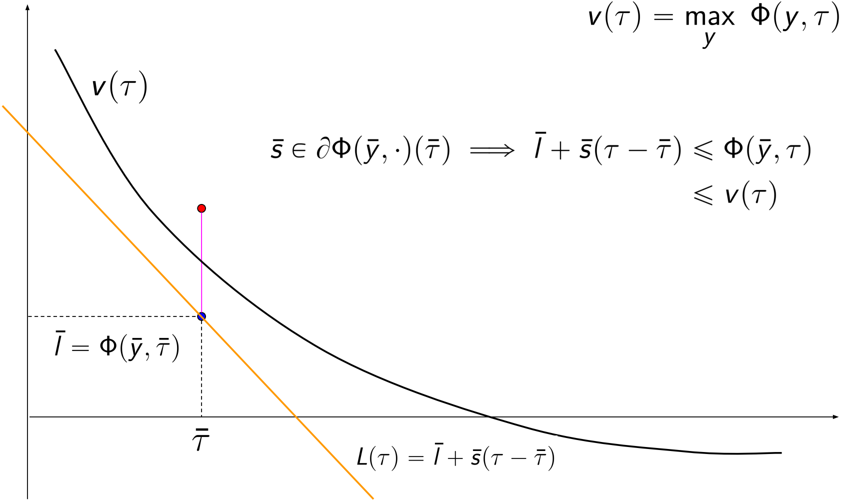


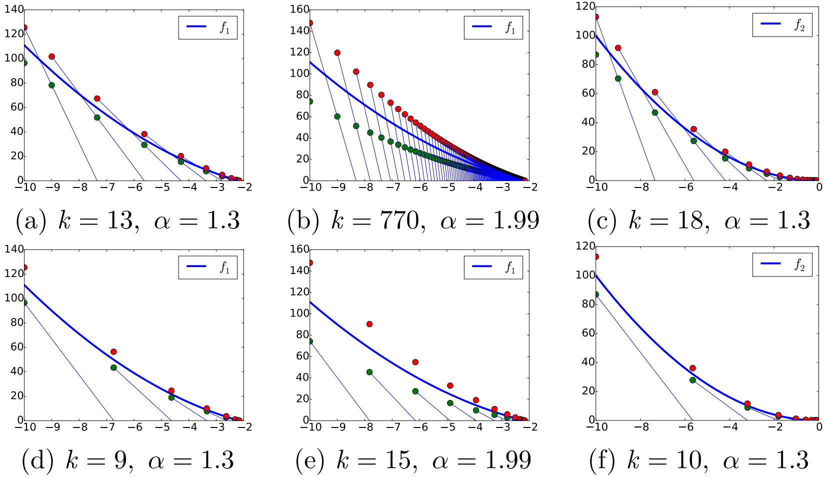
$$\bar{s} \in \partial\Phi(\bar{y}, \cdot)(\bar{\tau}) \implies \bar{l} + \bar{s}(\tau - \bar{\tau}) \leqslant \Phi(\bar{y}, \tau) \leqslant v(\tau)$$

Minorants from Duality

$$v(\tau) = \max_{y} \; \Phi(y, \tau)$$

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



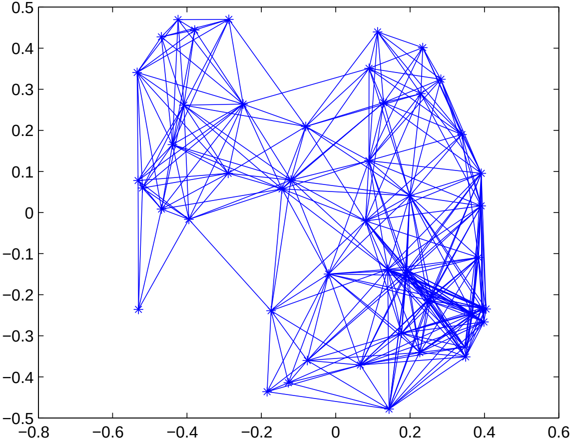


(a) $k = 13,\ \alpha = 1.3$ (b) $k = 770,\ \alpha = 1.99$ (c) $k = 18,\ \alpha = 1.3$

(d) $k = 9,\ \alpha = 1.3$ (e) $k = 15,\ \alpha = 1.99$ (f) $k = 10,\ \alpha = 1.3$

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, α is the oracle accuracy, and k is the number of iterations needed to converge, i.e., to reach $f_i(\tau_k) \leq \epsilon = 10^{-2}$.

Sensor Network Localization (SNL)



Given a weighted graph $G = (V, E, d)$ find a **realization**:

$$p_1, \ldots, p_n \in \mathbf{R}^2 \quad \text{with} \quad d_{ij} = \|p_i - p_j\|^2 \quad \text{for all } ij \in E.$$

Sensor Network Localization (SNL)

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

$$\begin{aligned} \max \quad & \operatorname{tr}(X) \\ \text{s.t.} \quad & \|\mathcal{P}_E \mathcal{K}(X) - d\|_2^2 \le \sigma \\ & Xe = 0, \quad X \succeq 0 \end{aligned}$$

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):

$$\begin{aligned} \max \quad & \mathrm{tr}\,(X) \\ \text{s.t.} \quad & \|\mathcal{P}_E \mathcal{K}(X) - d\|_2^2 \le \sigma \\ & Xe = 0, \quad X \succeq 0 \end{aligned}$$

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

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

Sensor Network Localization (SNL)

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

$$\begin{aligned} \max \quad & \operatorname{tr}(X) \\ \text{s.t.} \quad & \|\mathcal{P}_E \mathcal{K}(X) - d\|_2^2 \le \sigma \\ & Xe = 0, \quad X \succeq 0 \end{aligned}$$

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

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

Flipped problem:

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

Sensor Network Localization (SNL)

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

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

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

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

Flipped problem:

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

- Perfectly adapted for the Frank-Wolfe method.

Sensor Network Localization (SNL)

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

$$\begin{aligned} \max \quad & \operatorname{tr}(X) \\ \text{s.t.} \quad & \|\mathcal{P}_E \mathcal{K}(X) - d\|_2^2 \le \sigma \\ & Xe = 0, \quad X \succeq 0 \end{aligned}$$

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

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

Flipped problem:

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

- Perfectly adapted for the Frank-Wolfe method.

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

Approximate Newton

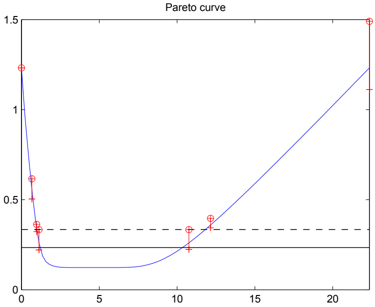


Figure : $\sigma = 0.25$

Approximate Newton

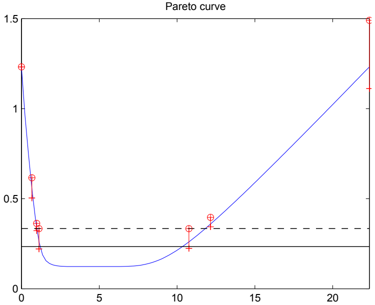


Figure : $\sigma = 0.25$

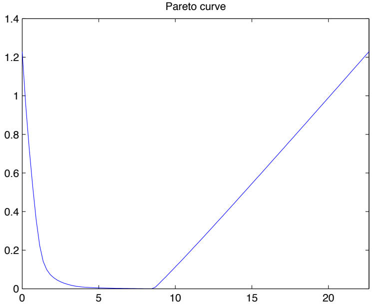
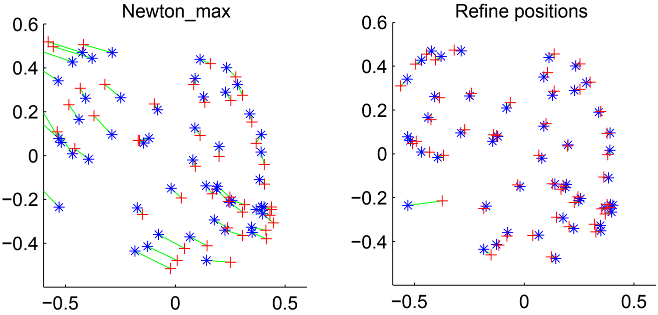
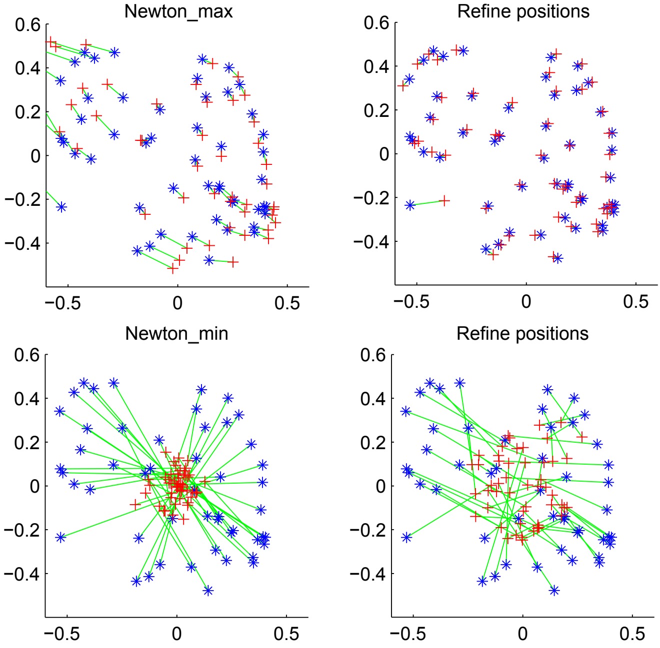


Figure : $\sigma = 0$

Max-trace



Max-trace


Newton_max
Refine positions
Newton_min
Refine positions
0.6
0.4
0.2
0
−0.2
−0.4
−0.5
0
0.5

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)
    - ...

Convex Indicator

For any convex set C, the convex indicator function for C is

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

Convex Indicator

For any convex set C, the convex indicator function for C is

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

Support Functionals

For any set C, the support functional for C is

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

Convex Indicator

For any convex set C, the convex indicator function for C is

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

Support Functionals

For any set C, the support functional for C is

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

Convex Conjugates

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

$$g^*(y) \ := \ \delta^*\left((y, -1) \mid \operatorname{epi}(g)\right) \ = \ \sup_x [\langle x, y \rangle - g(x)] \ .$$

Conjugate's and the Subdifferential

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

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

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 $\operatorname{epi}(g)$ is closed and $\operatorname{dom}(g) \neq \emptyset$, 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

$$z \in \partial g(x) \quad \text{ and } \quad x \in \partial g^*(z).$$

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

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

$$z \in \partial g(x) \quad \text{and} \quad x \in \partial g^*(z).$$

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 ∂g is a maximal monotone operator with $\partial g^{-1} = \partial g^*$.

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

The Bi-Conjugate Theorem
If $\mathrm{epi}\,(g)$ is closed and $\mathrm{dom}\,(g) \neq \emptyset$, 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

$$z \in \partial g(x) \quad \text{and} \quad x \in \partial g^*(z).$$

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

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

Note: *The lsc hull of g is* $\mathrm{cl}\, g := g^{**}$.

The perspective function

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

The perspective function

$$\mathrm{epi}\,(g^\pi) := \mathrm{cl}\,\mathrm{cone}\,(\mathrm{epi}\,(g)) = \mathrm{cl}\,\left(\bigcup_{\lambda>0} \lambda \mathrm{epi}\,(g)\right)$$

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

where g^∞ is the *horizon* function of g:

$$g^\infty(z) := \sup_{x \in \mathrm{dom}\, g} \left[g(x+z) - g(x) \right].$$

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

Then

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

and

$$\delta^*\left(y\,|\,[g \le \tau]\right) = \text{cl}\inf_{\mu \ge 0}\,[\tau\mu + (g^*)^\pi(y,\mu)],$$

where

$$\text{epi}\,(g) := \{(x,\mu)\,|\,g(x) \le \mu\}$$

$$[g \le \tau] := \{x\,|\,g(x) \le \tau\}$$

$$\delta^*\left(z\,|\,C\right) := \sup_{w \in C}\langle z, w\rangle$$

The perturbation function

$$f(x, b, \tau) := \rho(b - Ax) + \delta\left((x, \tau) \,|\, \mathrm{epi}\,(\phi)\right)$$

Its conjugate

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

The perturbation function

$$f(x, b, \tau) := \rho(b - Ax) + \delta\left((x, \tau) \,|\, \mathrm{epi}\,(\phi)\right)$$

Its conjugate

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

The Primal Problem infimal projection in x

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

The perturbation function

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

Its conjugate

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

The Primal Problem

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

The Dual Problem

$$\begin{aligned} \mathcal{D}(b, \tau): \qquad \hat{v}(b, \tau) &:= \sup_{u,\mu} \langle b, u\rangle + \tau\mu - f^*(0, u, \mu) \\ \text{(reduced dual)} \qquad &= \sup_u \langle b, u\rangle - \rho^*(u) - \delta^*\left(A^T u \mid [\phi \le \tau]\right). \end{aligned}$$

The perturbation function

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

Its conjugate

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

The Primal Problem

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

The Dual Problem

$$\begin{aligned} \mathcal{D}(b,\tau): \qquad \hat{v}(b,\tau) &:= \sup_{u,\mu} \langle b,u\rangle + \tau\mu - f^*(0,u,\mu) \\ \text{(reduced dual)} \qquad &= \sup_u \langle b,u\rangle - \rho^*(u) - \delta^*\left(A^T u \mid [\phi \le \tau]\right). \end{aligned}$$

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

$$\emptyset \neq \partial v(b,\tau) = \operatorname*{argmax}_{u,\mu}\ \mathcal{D}(b,\tau)$$

Piecewise Linear-Quadratic Penalties

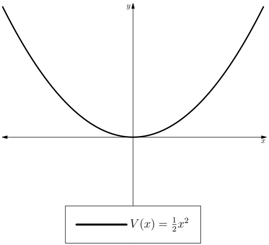
$$\phi(x) := \sup_{u \in U} [\langle x, u \rangle - \frac{1}{2} u^T B u]$$

$U \subset \mathbb{R}^n$ is nonempty, closed and convex with $0 \in U$ (not nec. poly.)

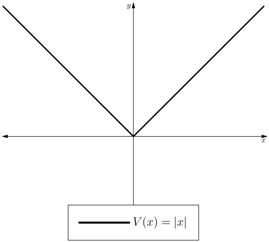
$B \in \mathbb{R}^{n \times n}$ is symmetric positive semi-definite.

Examples:

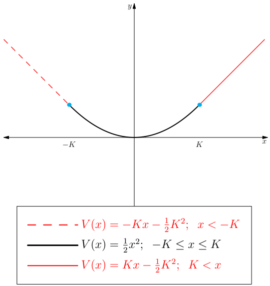
1. Support functionals: $B = 0$
2. Gauge functionals: $\gamma(\cdot \mid U^\circ) = \delta^*(\cdot \mid U)$
3. Norms: $\mathbb{B}$ = closed unit ball, $\|\cdot\| = \gamma(\cdot \mid \mathbb{B})$
4. Least-squares: $U = \mathbb{R}^n$, $B = I$
5. Huber: $U = [-\epsilon, \epsilon]^n$, $B = I$



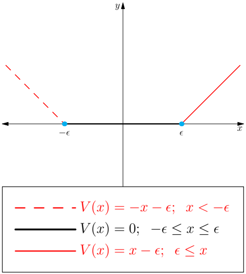
Gauss



ℓ_1



Huber



Vapnik

Computing v' for PLQ Penalties ϕ

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

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

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

A Few Special Cases

$$v(\tau) := \min \tfrac{1}{2}\|b - Ax\|_2^2 \quad \text{st } \phi(x) \le \tau$$

Optimal Solution: $\bar{x}$ Optimal Residual: $\bar{r} := A\bar{x} - b$

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

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

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

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

5. **Vapnik:** $\phi(x) = \|(x-\epsilon)_+\|_1 + \|(-x-\epsilon)_+\|_1 \implies$
   $v'(\tau) = -(\|A^T\bar{r}\|_\infty + \epsilon\|A^T\bar{r}\|_2)$

Basis Pursuit with Outliers

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

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

Basis Pursuit with Outliers

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

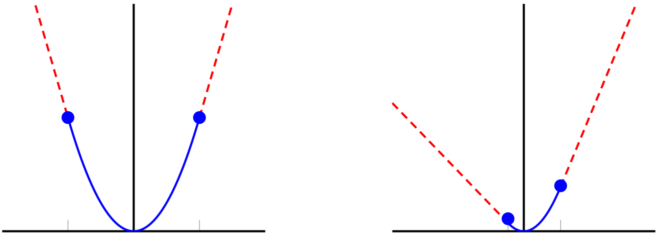
Standard least-squares: $\rho(z) = \|z\|_2$ or $\rho(z) = \|z\|_2^2$.

Quantile Huber:

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

Standard Huber when $\tau = 0.5$.

Huber



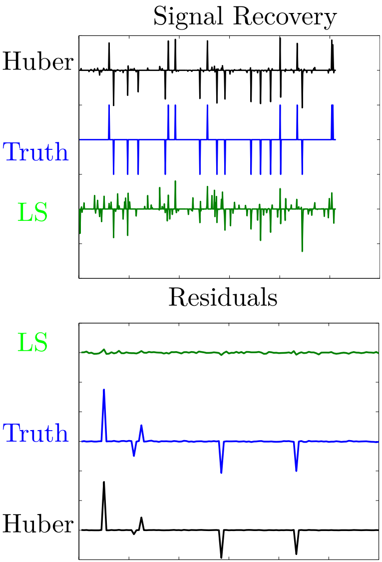
Sparse and Robust Formulation

$$\text{HBP}_\sigma: \quad \min \quad \|x\|_1 \quad \text{st} \quad \rho(b - Ax) \le \sigma$$

Problem Specification

x	20-sparse spike train in $\mathbb{R}^{512}$
b	measurements in $\mathbb{R}^{120}$
A	Measurement matrix satisfying RIP
ρ	Huber function
σ	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

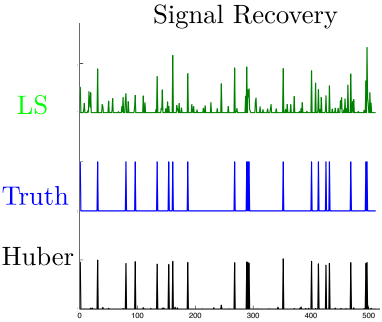
$$\text{HBP}_\sigma: \quad \min_{0 \le x} \quad \|x\|_1 \quad \text{st} \quad \rho(b - Ax) \le \sigma$$

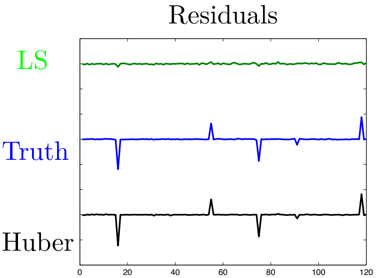
Problem Specification

- x 20-sparse spike train in $\mathbb{R}^{512}_{+}$
- b measurements in $\mathbb{R}^{120}$
- A Measurement matrix satisfying RIP
- ρ Huber function
- σ 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
  SIAM J. Sci. Comput. **31**(2008), 890–912.

- "Sparse optimization with least-squares constraints"
  van der Berg - Friedlander
  SIOPT **21**(2011), 1201–1229.

- "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