Second-Order Methods for ℓ_1-Regularized Problems in Machine Learning and Sparse Signal Reconstruction

Ioannis Dassios Kimon Fountoulakis Jacek Gondzio

University of Edinburgh

Computational Linear Algebra and Optimization
for the Digital Economy

Outline

Non-smooth optimization problems

- Machine Learning (Big-Data)
- Sparse Signal Reconstruction (work in progress)

Methods and Techniques

- Newton-type methods
- Continuation Framework
- Preconditioning

Numerical Results

- Parallel Coordinate Descent Methods
- First-Order Methods
- Interior Point Methods

Non-smooth optimization problems

Formulation (Generalized Lasso)

$$\text{minimize } f_c(x) := c\|W^*x\|_1 + \frac{1}{2}\|Ax - b\|_2^2$$

- $x \in \mathbb{R}^m$
- $c \in \mathbb{R}_+$
- $W : E^l \to \mathbb{R}^m$, where $E = \mathbb{R}$ or $\mathbb{C}$
- $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$
- The image of optimal x through W^* is sparse

Difficulties

- Non-smoothness
- Large scale problems

Assumptions

Assumption (A.1)

- $\mathrm{Ker}(W^*) \cap \mathrm{Ker}(A) = \{0\} \iff$ set of minimizers is non-empty and compact

What about assumptions that guarantee a unique minimizer?

$$\mathcal{J} := \{i \in \{1, \dots, l\} \mid W_i^* \tilde{x} = 0\}$$

and

$$\tilde{x} := \arg\min_x \|W^* x\|_1, \quad \text{subject to: } Ax = \tilde{b}$$

where $b = \tilde{b} + e$ and e is a vector of noise

Assumptions (A.2) (sufficient)

- $\mathrm{Ker}(W_{\mathcal{J}}^*) \cap \mathrm{Ker}(A) = \{0\}$
- level of the error is bounded, i.e. $\|e\|_2$ is sufficiently small

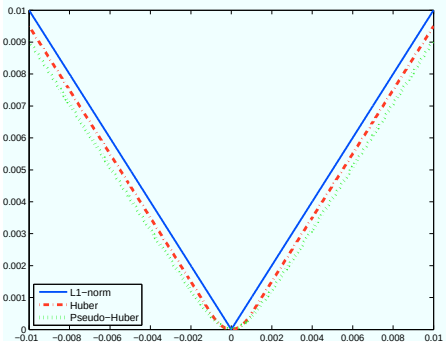
S. Vaiter, G. Peyre, C. Dossal & J. Fadili,
IEEE Transactions on Information Theory (4) 59 (2013) 2001-2016.

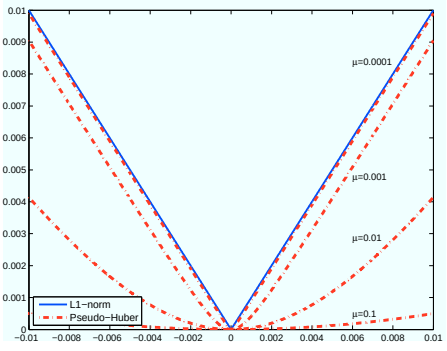
Smoothing

Options

- Huber: First-order differentiable
- **Pseudo-Huber**: *Second-order differentiable*

$$\psi_\mu(W^*x) = \sum_{i=1}^{l} \left(\sqrt{\mu^2 + (\overline{W}_i^* x)(W_i^* x)} - \mu \right)$$





Perturbation Analysis

Original solution (unique from A.2): $x(c) := \arg\min_x f_c(x)$

Perturbed solution (unique from A.1): $x(c, \mu) := \arg\min_x f_c^\mu(x)$

where $f_c^\mu(x) = c\psi_\mu(W^* x) + \frac{1}{2}\|Ax - b\|_2^2$

Theorem

There exists a $\tilde{\mu}$ such that $\forall \mu \leq \tilde{\mu}$ the difference of the two solutions is

$$\|x(c, \mu) - x(c)\|_2 = \mathcal{O}(\mu) \quad \forall c, \mu.$$

I. Dassios, K. Fountoulakis & J. Gondzio,
A second-order method for sparse signal reconstruction problems (work in progress)

Methods and Techniques

Newton-type Methods & Continuation

Methods

- Newton Conjugate Gradients (N-CG)
  - Preconditioners for N-CG: Problem dependent, to be discussed.
- Quasi-Newton, i.e. LBFGS

Continuation Framework

1: *Outer loop: For* $k = 0, 1, 2, \ldots, \vartheta$, *produce* $(c^k, \mu^k)_{k=0}^{\vartheta}$.
2: *Inner loop: Approximately solve the subproblem*

$$\text{minimize } f_{c^k}^{\mu^k}(x)$$

using a Newton-type method by initializing it with the solution of the previous subproblem.

Why continuation? Control of Spectrum

Spectrum (in practise): $\gamma I_m \preceq \nabla^2 f_c^\mu(x) + \gamma I \preceq (\mathcal{O}(\frac{c}{\mu}) + \lambda_{max}(A^\top A))I_m$

Example without continuation: $c = 10^{-2}$, $\mu = 10^{-10}$

What happens in practise?

- During all stages of Newton-type method $\kappa(\nabla^2 f_c^\mu(x)) = \mathcal{O}(\frac{c}{\gamma\mu})$

Example with continuation: $c^0 = \mu^0 = \|A^\top b\|_\infty$ and $c = 10^{-2}$, $\mu = 10^{-10}$

What happens in practise?

- early stages $\kappa(\nabla^2 f_c^\mu(x)) = \mathcal{O}(\frac{1}{\gamma})$
- late stages $\kappa(\nabla^2 f_c^\mu(x)) = \mathcal{O}(\frac{c}{\gamma\mu})$

Enable inexpensive preconditioning for small $\frac{c^k}{\mu^k} \geq 10$ (efficient only close to $x(c, \mu)$)

Warm-starting

By setting $c^0 = \mu^0$, then

$$\lim_{c^0 \to \infty} c^0 \psi_{c^0}(W^* x) = \frac{1}{2}\|W^* x\|_2^2,$$

where $c\psi_c(W^* x) = \sum_{i=1}^{l} \left(c\sqrt{c^2 + (\overline{W}_i^* x)(W_i^* x)} - c^2 \right)$.

Thus, for $c^0 \to \infty$ and $c^0 = \mu^0$, continuation is warm-started by approximately solving

$$\text{minimize}_{x \in \mathbb{R}^m} \ \frac{1}{2}\|W^* x\|_2^2 + \frac{1}{2}\|Ax - b\|_2^2,$$

which has a unique solution (remember $\text{Ker}(W^*) \cap \text{Ker}(A) = \{0\}$!) with closed form

$$x = (A^\mathsf{T} A + \text{Re}(WW^*))^{-1} A^\mathsf{T} b$$

Machine Learning (Big-Data)

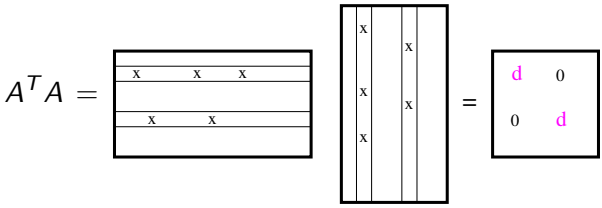
Sparse Least-Squares

$$\text{minimize } c\|x\|_1 + \frac{1}{2}\|Ax - b\|_2^2$$

where $x \in \mathbb{R}^m$, $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$ with $n \geq m$.

Properties

- m, n can be **very** large, millions or billions
- A is **very** sparse, i.e. 20 non-zeros per column



Observation

$$\min_x \ \tau\|x\|_1 + \frac{1}{2}\|Ax - b\|_2^2,$$

Quadratic Opt. with $Q = A^T A$. For overdetermined systems ($n > m$), Q is likely to be very well conditioned.

Small exercise: Regularized Steepest-descent vs Newton

Ignore ℓ_1 term and compute:

$\nabla\phi(x) = A^T(Ax - b)$ and $\nabla^2\phi(x) = A^T A$

$d_{SD} = -diag(\nabla^2\phi(x))^{-1}\nabla\phi(x)$ and $d_N = -(\nabla^2\phi(x))^{-1}\nabla\phi(x)$

If $\boldsymbol{\nabla^2\phi(x) \approx diag(\nabla^2\phi(x))}$ then $d_{SD} \approx d_N$.

Compared Solvers & Setting

- **RCDC: parallel Randomized Coordinate Descent**
  by P. Richtárik and M. Takáč
  - Exploits sparsity/separability of problem through multicore systems

- **dcNCG: doubly-continuation Newton Conjugate Gradients**
  - Continuation on both c and μ
  - PCG with $P = c\nabla^2\psi_\mu(x) + diag(A^\intercal A)$
  - Implements Ax and $A^T y$ in parallel

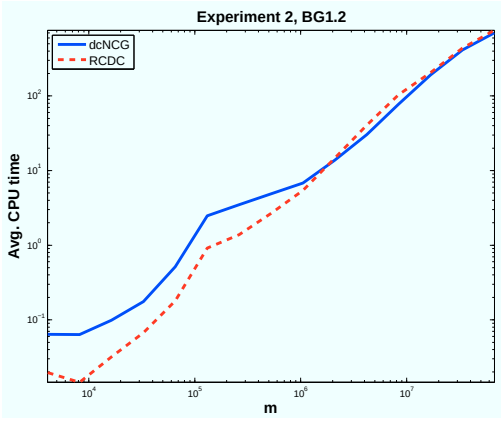
- Experiments are run on a 24 core system

- The problem generator has been first used by **Y. Nesterov**, in *Gradient Methods for Minimizing Composite Objective Function*

Sparse Least-Squares: increasing m

Info

- $\frac{\text{mass of } diag(A^TA)}{\text{mass of } A^TA} \approx 99.9\%$
- 20 non-zero elements per column of A
- $\mu = 10^{-12}$
- $f(x) - f(x^*) = \mathcal{O}(10^{-6})$

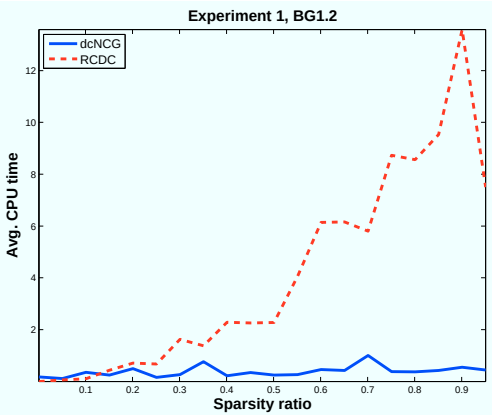


Why is RCDC so fast? Coordinate directions biased with second-order information

Sparse Least-Squares: increasing density

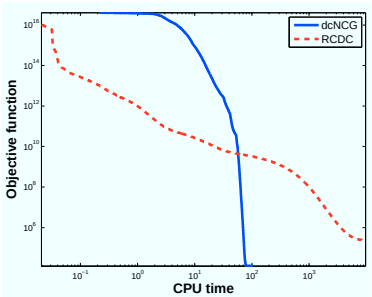
Info

- $\frac{\text{mass of } diag(A^\intercal A)}{\text{mass of } A^\intercal A} \geq 40\%$
- $cond(diag(A^\intercal A)^{-1} A^\intercal A) \leq 10^7$
- sparsity ratio: # of non-zero elements per column of A

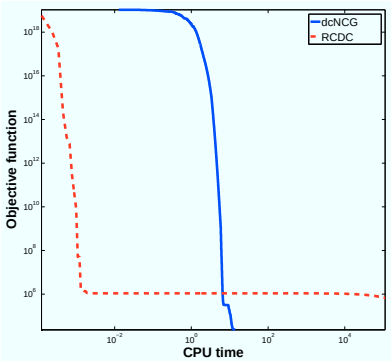


dcNCG is not affected by the sparsity ratio of A

Sparse Least-Squares: difficult cases



- The singular values of A are uniformly distributed in $(10^{-1}, 10^{6})$.



- Two clusters of singular values, 100 are 10^{6} and the rest are 10^{-1}.

PCDM was terminated after 10 million iterations $\approx 10k$ seconds and 1 billion iterations $\approx$ 31 hours.

Sparse Signal Reconstruction
(work in progress)

Sparse Signal Reconstruction Formulation

$$\text{minimize } c\|W^*x\|_1 + \frac{1}{2}\|Ax - b\|_2^2$$

where $x \in \mathbb{R}^m$, $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times m}$ with $m \geq n$.

Problems

- *Least Squares*: W is the identity
- ℓ_1-*analysis*: $W : \mathbb{R}^l \to \mathbb{R}^m$ is arbitrary
- *isotropic TV (iTV)*: $W : \mathbb{R}^{m-1} \to \mathbb{C}^m$ is tridiagonal
- *iTV* & ℓ_1-*analysis*: combination of the last two (not well studied yet)

$$\text{minimize } \sum_{i=1}^{2} c_i\|W_i^*x\|_1 + \frac{1}{2}\|Ax - b\|_2^2$$

Properties of A and W

- **rows** of A are nearly orthogonal to each other

$$\|AA^{\mathsf{T}} - I_n\|_2 \leq \delta$$

- linear combinations of any small subsets of **columns** of W satisfy the: **Restricted Isometry Property (W-RIP)**

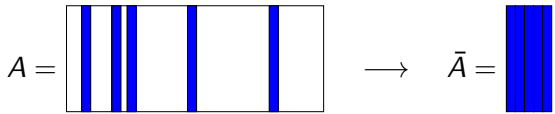
$$\|W_{\mathcal{B}}^* A^T A W_{\mathcal{B}} - \rho W_{\mathcal{B}}^* W_{\mathcal{B}}\|_2 \leq \delta_k \in (0, 1).$$

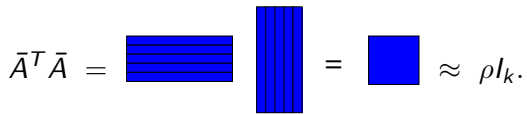
where $W_{\mathcal{B}}$ is a subset of columns of matrix W

Candès, Eldar, Needell & Randall,
Appl. Comp. Harmonic. Anal. 31 (2011) 59-73.

Identity-RIP (illustration)

Matrix $\bar{A} \in \mathcal{R}^{n \times k}$ ($k \ll m$) is built of a subset of columns of $A \in \mathcal{R}^{n \times m}$.





Preconditioner for N-CG

Approximate: $H = c\nabla^2\psi_\mu(W^*x) + A^\intercal A + \gamma I_m,\ \gamma > 0$

with: $P = c\nabla^2\psi_\mu(W^*x) + (\frac{n}{m} + \gamma)I_m,$

(A is usually built of a subset of n rows of an orthonormal matrix $U \in \mathcal{R}^{m\times m}$)
$\frac{n}{m} := \arg\min_\xi \|A^\intercal A - \xi I_m\|_F$

Based on three facts:

- For small $\mu \leq 10^{-4}$ and close to $x(c,\mu) := \arg\min_x f_c^\mu(x)$
$$\|\nabla^2 f_c^\mu(W^*x)\|_2 \approx \|c\nabla^2\psi_\mu(W^*x)\|_2$$
- $\|AA^\intercal - I_n\|_2 \leq \delta \Longrightarrow \lambda_{max}(A^\intercal A) \leq 1 + \delta$
- W-RIP

Preconditioner (Example W=identity)

Prior information for $\mathbf{x(c)}$

- $k \ll m$ components are non-zero
- the majority $m - k$ components are zero

Information for $\mathbf{x(c, \mu)}$

For sufficiently small μ & $\|x(c, \mu) - x(c)\| = \mathcal{O}(\mu)$, we expect that

- k components of $x(c, \mu)$ are non-zero
- the majority $m - k$ are $\mathcal{O}(\mu)$

Information for $\boldsymbol{\nabla^2 \psi_\mu(\mathbf{x(c, \mu)})}$ (diagonal matrix)

- k components are very small
- the majority $m - k$ are $\mathcal{O}(\frac{1}{\mu})$ (large!!)

Given that $\lambda_{max}(A^T A) \leq 1 + \delta$, for moderate c we expect that at optimality

$$\|\nabla^2 f_c^\mu(W^* x(c, \mu))\|_2 = \|c\nabla^2 \psi_\mu(W^* x(c, \mu)) + A^T A\|_2 \approx \|c\nabla^2 \psi_\mu(W^* x(c, \mu))\|_2$$

Spectral properties

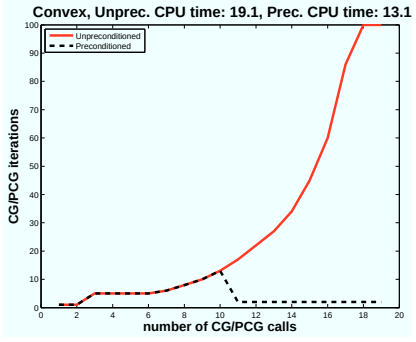
Preconditioner performance for LS (W *is the identity*)

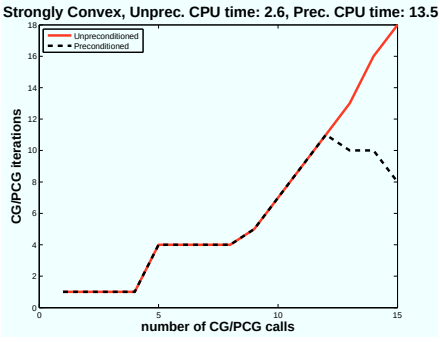
Theorem (Brief description)

Close to the solution $x(c,\mu) := \arg\min_x f_c^\mu(x)$ *the following holds*

- $|\lambda(P^{-1}H) - 1| \leq \delta_k + \mathcal{O}(\frac{\mu}{c})$

Preconditioner performance for iTV





Solving systems with P

- **For Least Squares:** P is diagonal (inexpensive)

- **For isotropic Total-Variation:**
  P is a **5-diagonal** matrix (inexpensive)

- **For ℓ_1-analysis where the matrix W is arbitrary:**
  P does not have a structure (expensive)
  Solution: control of spectrum through continuation on both c and μ

Compared Solvers & Setting

Least Squares

- *SPGL1*: Spectral Projection Gradient method
  by E. van den Berg and M. P. Friedlander
- ℓ_1-ℓ_s: interior point method
  by S. Boyd et. al.

ℓ_1-analysis and isotropic Total Variation

- *TFOCS: Templates for First-Order Conic Solvers*
  by S. R. Becker, E. J. Candés and M. C. Grant
  - Auslender and Teboulle's single-projection method
- *TwIST: Two-step Iterative Shrinkage/Thresholding*
  by J. Bioucas-Dias, M. A. T. Figueiredo

- **We reproduce experiments as given by papers/demos in existing state-of-the-art solver packages.**

SPARCO collection of LS problems

Comparison on 18 out of 26 classes of problems
(all but 6 complex and 2 installation-dependent ones).

Results

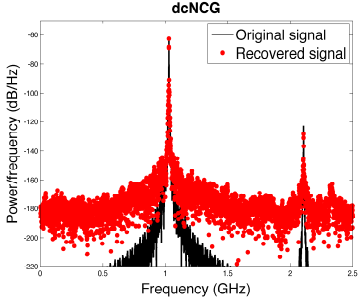
On 36 runs (noisy and noiseless problems), **dcNCG**:

- is the fastest on 6,
- is the second best on 21 (*with large gap from the first*), and
- overall is robust.

ℓ_1-Analysis (recovery of radar pulses)

Info

- W is a Gabor frame
- A is a block diagonal matrix, with ± 1 for entries
- Subsampling: $m = 12n$
- Noise is added so that the small pulse has SNR 0.1 dB
- **dcNCG** ($\mu = 10^{-5}$): time=0.3 min., rel. err.=1.56e-3
- **TFOCS**: time=1.0 min., rel. err.=1.82e-3

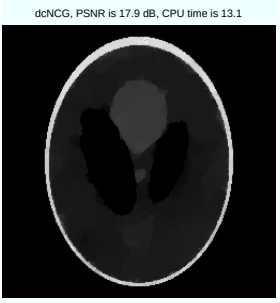


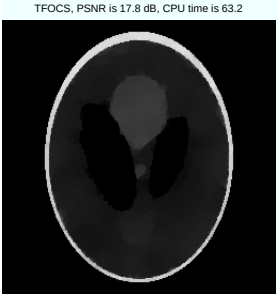
Isotropic Total-Variation

Info

- A is a partial Fourier matrix
- Sub-sampling is: $m = 4n$
- SNR 10 dB
- **dcNCG** ($\mu = 10^{-4}$): time=13.1 sec., PSNR.=17.9 dB
- **TFOCS**: time=63.2 sec., PSNR=17.8 dB

$$PSNR = 20 \log_{10} \left(\frac{\sqrt{n_1 n_2}}{\|x - \bar{x}\|_F} \right)$$





Isotropic Total-Variation and ℓ_1-Analysis (denoising)

Info

- Reg.: $c_1\|W_1^* x\|_1 + c_2\|W_2^{\mathsf{T}} x\|_1$
- W is 9/7 bi-orthogonal wavelet transform
- A is the identity
- SNR 10 dB
- **dcNCG** ($\mu = 10^{-4}$): time=6.6 sec., PSNR=27.6 dB
- **TFOCS**: time=12.7 sec., PSNR=27.5 dB

dcNCG, PSNR 27.6 dB, CPU time is 6.6

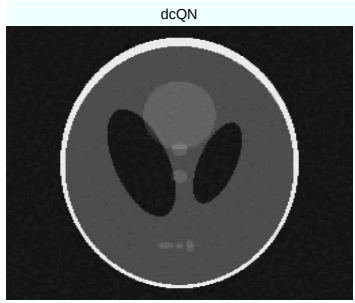
TFOCS, PSNR 27.5 dB, CPU time is 12.7

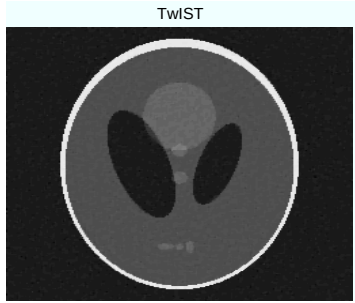
Isotropic Total-Variation, deconvolution problem

Info

- A is a deconvolution operator
- blur uniform 9*9,
  SNR 40 dB,
- **dcNCG** failed!
- **dcQN** ($\mu = 10^{-15}$): time=10.91 sec.,
  ISNR.=17.4 dB
- **TwIST**: time=9.48 sec.,
  ISNR=17.3 dB

$$ISNR = 10 \log_{10} \left(\frac{\|vec(x - x_n)\|_2}{\|vec(x - \bar{x})\|_2} \right)$$





Conclusion: First-order methods have dominated the field of ℓ_1-regularized problems, but **second-order information should not be ignored**!

Thank you!


I. Dassios, K. Fountoulakis, and J. Gondzio.
A second-order method for sparse signal reconstruction problems.
In preparation.


K. Fountoulakis and J. Gondzio.
A second-order method for strongly convex ℓ_1-regularization problems.
Technical Report ERGO-13-011.

Why $\rho = \frac{n}{m}$?

A is usually built of a subset of n rows of an orthonormal matrix $U \in \mathcal{R}^{m \times m}$. Thus,

$$diag(U_{\mathcal{B}}^{\mathsf{T}} U_{\mathcal{B}}) \approx \frac{n}{m} I$$

where $\mathcal{B}$ is the set of indices of the chosen n rows from U. Hence,

$$\rho := \arg\min_{\xi} \|A^{\mathsf{T}} A - \xi I_m\|_F = \|U_{\mathcal{B}}^{\mathsf{T}} U_{\mathcal{B}} - \xi I_m\|_F$$

and $\rho = \frac{n}{m}$.

Details for Prec. Performance of iTV

Figures 1 and 2

- Problem size: reconstruction of 256×256 images
- Tolerance of PCG: 10^{-2}
- Both versions required same number of iterations
- $\mu = 10^{-4}$ (Figure 1), $\mu = 10^{-3}$ (Figure 2)

Derivatives of Pseudo-Huber

For $W : \mathbb{C}^l \to \mathbb{R}^m$

Gradient at x: $\frac{1}{\mu}\mathrm{Re}(WG(x)W^*x)$, where $G_{ii}(x) = \frac{1}{\sqrt{1 + \frac{[Wx]_i[\bar{Wx}]_i}{\mu^2}}} \quad \forall i$

Hessian at x: $\frac{1}{2\mu}(\mathrm{Re}(WY(x)W^*) - \mathrm{Re}(W\tilde{Y}(x)W^{\mathsf{T}}))$,

where $Y_{ii}(x) = G_{ii}(x) + G_{ii}^3(x)$ and $\tilde{Y}_{ii}(x) = \frac{1}{\mu^2}[Wx]_i^2 G_{ii}^3(x) \quad \forall i$

iTV Tridiagonal Matrix

$$\|x\|_{iTV} = \sum_{i,j} \sqrt{|x_{i+1,j} - x_{i,j}|^2 + |x_{i,j+1} - x_{i,j}|^2}$$

where x is an $m \times m$ image.

Notice, that this norm calculates for all pixels the horizontal and vertical differences of a given pixel. We can rewrite is as

$$\|x\|_{iTV} = \|W^* y\|_1$$

where $y = x(:) \in \mathbb{R}^{m^2}$, and

$$[W^* y]_{(i-1)m+j} = (y_{im+j} - y_{(i-1)m+j}) + \sqrt{-1}(y_{(i-1)m+j+1} - y_{(i-1)m+j}), \quad 1 \leq i, j \leq m$$

The matrix W^* places horizontal and vertical differences around a given pixel at position i, j of an image into the real and imaginary elements of the output.

What close to the path $x(c, \mu)$ means?

For sufficiently small μ the optimal solution $x(c, \mu)$ $\forall c$ satisfies:

- $k \ll m$ components of $W^* x(c, \mu)$ tend to non-zero values,
- the rest $m - k$ components are of $\mathcal{O}(\mu)$.

In words, the term $W^* x(c, \mu)$ exhibits a separability behaviour. All points near $x(c, \mu)$ with similar separable behaviour as $x(c, \mu)$, we consider them as points close to $x(c, \mu)$. This result is quantified in the theorem, but for presentation purposes the details have been omitted.