FAST GRADIENT BASED ALGORITHMS FOR CONSTRAINED TOTAL VARIATION IMAGE DENOISING AND DEBLURRING PROBLEMS

AMIR BECK AND MARC TEBOULLE

Optimization II – Final Project
Presented by Royi Avital

AGENDA

- Motivation
- The Total Variation Operator
- The Problem Model
- Previous Methods
  - Sub Gradient Descent
  - Dual Form (Chambolle’s Formalization)
- The Suggested Method
  - Innovations
- Results
- Remarks & Conclusions

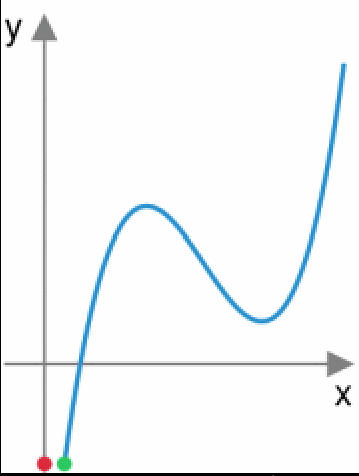
MOTIVATION

- Denoising

- Deblurring

THE TOTAL VARIATION OPERATOR

- Math Operator which operates on Function.
  Intuitively, for 1D Functions, measures the Arc Length – $\|f'(t)\|_{TV} = \int |f'(t)|dt$.



Wikipedia

- Due to its "Edge Preserving" Property, gained popularity as a Regularization Operator in the Image Processing World (Rudin, Osher & Fatemi 1992).

THE TOTAL VARIATION OPERATOR

- The Discrete Form (Isotropic)

$$\|x\|_{TV} = \sum_{i,j} \sqrt{\left(x_{i+1,j} - x_{i,j}\right)^2 + \left(x_{i,j+1} - x_{i,j}\right)^2}$$

- The Discrete Form (Anisotropic)

$$\|x\|_{TV} = \sum_{i,j} \sqrt{\left(x_{i+1,j} - x_{i,j}\right)^2} + \sqrt{\left(x_{i,j+1} - x_{i,j}\right)^2} = \sum_{i,j} \left|x_{i+1,j} - x_{i,j}\right| + \left|x_{i,j+1} - x_{i,j}\right|$$

- The Matrix Form (Isotropic)

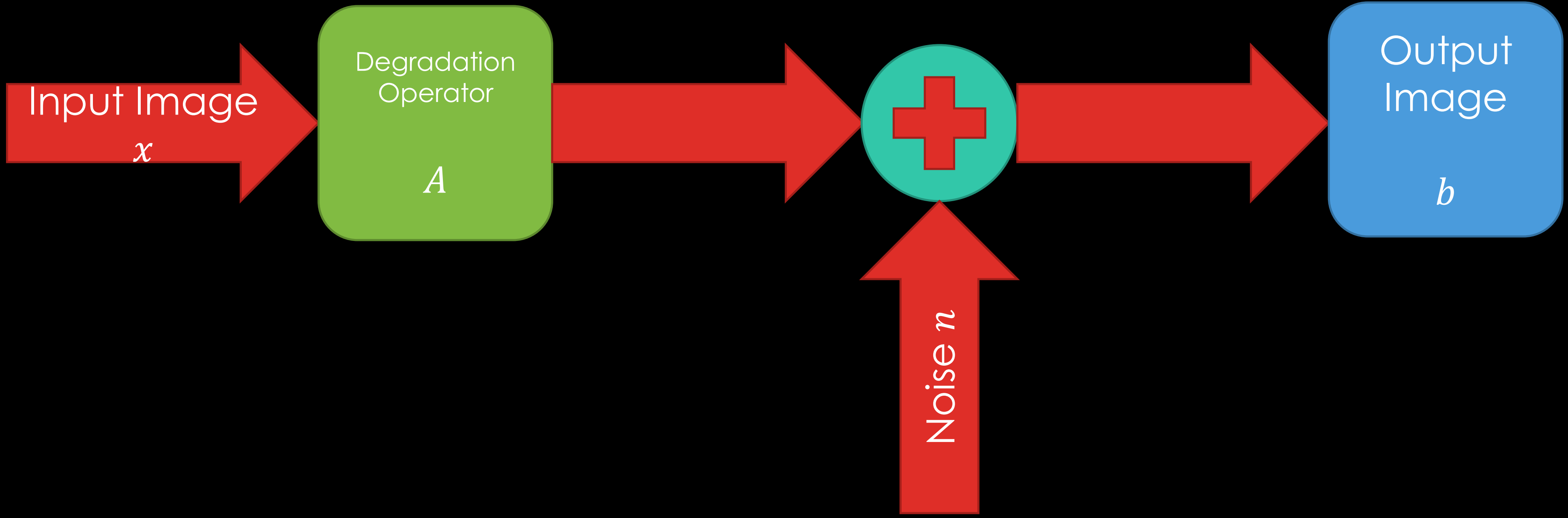
$$\|x\|_{TV} = \|\nabla x\|_{2,1}, \qquad \nabla x\colon \mathbb{R}^{m\times n} \to \mathbb{R}^{2\times mn}$$

- The Matrix Form (Anisotropic)

$$\|x\|_{TV} = \|Dx\|_1, \qquad D = \begin{bmatrix} -1 & 1 & 0 \\ & & \\ & & \end{bmatrix}$$

THE PROBLEM MODEL

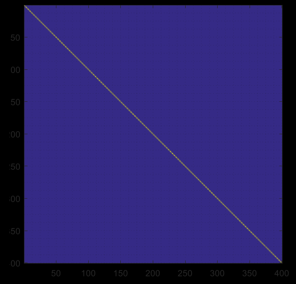
Degradation Model – $x = Ab + n.$

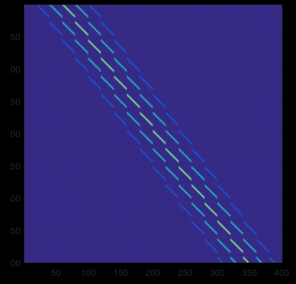


THE PROBLEM MODEL

Degradation Operators

- Denoising – $A = I$.





- Deblurring – A Block Toeplitz Matrix of LPF.

THE PROBLEM MODEL

- The Optimization Problem

$$x^* = \arg\min_x \|Ax - b\|_2^2 + \lambda g(x) = \arg\min_x f(x) + \lambda g(x)$$

- The function $g(x)$ is the Regularization Term
  - Gaussian Distribution (Tikhonov Regularization) – $g(x) = \|x\|_2$.
  - Gaussian Distribution of Derivatives – $g(x) = \|Dx\|_2$.
  - Lasso Regularization – $g(x) = \|x\|_1$.
  - Exponential Distribution of Derivatives – $g(x) = \|Dx\|_1$.
  - Sparse Model – $g(x) = \|x\|_0$.
  - Total Variation – $g(x) = \|x\|_{TV}$.
    For this presentation the Anisotropic Model is used.

PREVIOUS METHODS

- The Anisotropic TV Denoising Model in Matrix Form:

$$x^* = \arg\min_x H(x) = \arg\min_x \frac{1}{2}\|x - b\|_2^2 + \lambda\|Dx\|_1$$

- The Sub Gradient Method

$$\mathrm{x} - \mathrm{b} + \lambda \mathrm{D}^{\mathrm{T}}\mathrm{sgn}(Dx) \in \partial\left(\frac{1}{2}\|x - b\|^2 + \lambda\|Dx\|_1\right)$$

- Sub Linear Rate ($\frac{1}{\sqrt{k}}$) -> Inadequate for Image Processing.
- Could it be done using Proximal Gradient Method (For Linear Convergence Rate)? No!
  No Closed Form Solution for $Prox_{t\|D\cdot\|_1}(x)$ (Mind the Linear Operator D – Not a Tight Frame).

PREVIOUS METHODS

Antonin Chambolle – An Algorithm for Total Variation Minimization and Algorithm

- Duality Based Algorithm – The problem becomes “Smooth” in its Dual Form.
- Duality comes from a “Trick” (Dual Norm, Support Function)

$$\|\nabla x\|_1 = \max_p \{p^T \nabla x \mid \|p\|_\infty \leq 1\}$$

- In Matrix Form

$$\|Dx\|_1 = \max_p \{p^T \mathrm{D}x \mid \|p\|_\infty \leq 1\}$$

- Intuition tells that the expected p^* should be given by $p^* = sgn(Dx)$.
  Namely the solution given by Extreme Points of the Set.
  Some remarks on that later on.

PREVIOUS METHODS

Chamoblle's Dual Method

1. Problem formulation
$$\arg\min_{x}\left\{\frac{1}{2}\|x-b\|^2+\lambda\|Dx\|_1\right\}=\arg\min_{x}\max_{\|p\|_\infty\le 1}\left\{\frac{1}{2}\|x-b\|^2+\lambda p^T Dx\right\}$$
2. By the Min Max Theorem (The objective is Convex in x and Concave in p) one could switch the order of the Maximum and Minimum
$$\arg\max_{\|p\|_\infty\le 1}\min_{x}\left\{\frac{1}{2}\|x-b\|^2+\lambda p^T Dx\right\}$$
3. Given $x^*=b-\lambda D^T p$ the problem becomes
$$\arg\max_{\|p\|_\infty\le 1}\left\{\frac{1}{2}\|x^*-b\|^2+\lambda p^T Dx^*\right\}=\arg\min_{\|p\|_\infty\le 1}\left\{\frac{1}{2}\lambda^2 p^T DD^T p-\lambda b^T D^T p\right\}$$
4. Strong Duality holds for this problem hence x^* can be extracted from p^*.

PREVIOUS METHODS

Chamoblle's Dual Method

5. This is a Minimization of a Convex Function over a Convex Set. The Gradient is given by

$$\nabla\left(\frac{1}{2}\lambda^2 p^T DD^T p - \lambda b^T D^T p\right) = \lambda^2 DD^T p - \lambda Db$$

6. Solution using Projected Gradient Method

$$p^{k+1} = P_{\|p\|_\infty \leq 1}\left(p^k - t_k\left(\lambda^2 DD^T p^k - \lambda Db\right)\right)$$

7. Where the projection is given by (Per Element)

$$P_{\|\cdot\|_\infty \leq 1}(x)_i = \frac{x_i}{\max(1, x_i)}$$

PREVIOUS METHODS

Chamoblle's Dual Method – Projection onto the ℓ_∞ Ball

- By Moreau Decomposition

$$prox_f(x) + prox_{f^*}(x) = x$$

- In case of $f(x) = \|x\|$ then $f^*(x) = P_{B_{\|\cdot\|_*}}(x)$ then

$$Prox_{\|\cdot\|_1}(x) = x - P_{B_{\|\cdot\|_\infty}}(x)$$

- It's known that $Prox_{\|\cdot\|_1}(x)$ is given by Soft Thresholding

$$Prox_{\|\cdot\|_1}(x)_i = \begin{cases} x_i - 1 \ \ if \ \ x_i \geq 1 \\ x_i - x_i \ \ if \ \ |x_i| < 1 \\ x_i + 1 \ \ if \ \ x_i \leq -1 \end{cases} \Rightarrow P_{B_{\|\cdot\|_\infty}}(x)_i = \begin{cases} 1 \ \ if \ \ x_i \geq 1 \\ x_i \ \ if \ \ |x_i| < 1 \\ -1 \ \ if \ \ x_i \leq -1 \end{cases}$$

PREVIOUS METHODS

Chamoblle’s Dual Method – Step Size

- The Minimization problem had Quadratic Form

$$\mathrm{h(p)} = \frac{1}{2}\lambda^2 p^T D D^T p - \lambda b^T D^T p$$

- The Lipschitz Constant is given by $\lambda_{max}(DD^T)$.
  The article showed $\lambda_{max}(DD^T) \leq 8$.
  Could a Tighter Bound be found? Yes!

PREVIOUS METHODS

Chamoblle's Dual Method – Step Size

- The Gershgorin Circle Theorem states that all Eigen Values of a given Matrix $A \in \mathbb{R}^{n \times n}$ are within Discs defined by $D(a_{ii}, R_i)$ where $R_i = \sum_{j \neq i} |a_{ij}|$.
- All elements on the diagonal of DD^T equal to 2, Hence $a_{ii} = 2$.
- All off diagonal elements of are either 1 or -1 and there are not more than 2 on each row, Hence $R_i \leq 2$.
- Applying the Circle Theorem $\lambda_{max}(DD^T) \leq \max_i \{a_{ii} + R_i\} \leq 4$.
- The Step Size t_k must obey $t_k \leq \frac{1}{4\lambda^2}$.
- Can we get even better? Yes!

PREVIOUS METHODS

Chamoblle's Dual Method – Step Size

- Look at the Optimizing Step of p^{k+1} and examine it (Neglecting λ)

$$\mathrm{p}^{\mathrm{k}+1} \cong p^k - t_k\left(DD^T p^k - Db\right) = (I - t_k DD^T)p^k + t_k Db$$

- This iteration converges if $\lambda_{max}(I - t_k DD^T) \leq 1$.
- Defining $\lambda_{DD^T} = \lambda_{max}(DD^T)$ suggests that $\lambda_{max}(I - t_k \lambda^2 DD^T) = 1 - t_k \lambda_{DD^T}$.
- Since DD^T is PD Matrix hence $\lambda_{max}(DD^T) > 0$.
- Limiting $-1 \leq 1 - t_k \lambda_{DD^T} \leq 1$ results in $t_k \leq \frac{2}{\lambda_{DD^T}}$.
- Since it is shown that $\lambda_{DD^T} \leq 4$ and plugging in the λ factor yields $t_k \leq \frac{1}{2\lambda^2}$.

PREVIOUS METHODS

Chamoblle's Dual Method – Practical Notes

- The Derivative Operator D is applied using Convolution.
  Once for the Horizontal Derivative and Once for Vertical Derivative (The Operator and its Adjoint). Both images are vectorized into one long Vector.
- The Operator DD^T is basically the Divergence / Discrete Laplace Operator.
  Again, should be applied using Convolution.
- The term Db, The Gradient of b, can be calculated once.
- Initialization of p should be made by $p^0 = P_{\|p\|_\infty \leq 1}\left(\frac{1}{\lambda}\left(DD^T\right)^{-1} D\left(b - x^0\right)\right)$.
- One could calculate x^* only once at the end of the iterations.
- The Step Size t_k must obey $t_k \leq \frac{1}{4\lambda}$.

SUGGESTED METHOD

Innovations

- Using FISTA to accelerate the convergence.
- Adding constrain on the value of the output image.
- Using Denoising Operator for Deblurring.
- Monotonic FISTA.

SUGGESTED METHOD

Innovations – Using FISTA to accelerate the convergence

- Apply FISTA on the Projected Gradient Descent problem
  - Set $p^0 = P_{\|p\|_\infty \le 1}\left(\frac{1}{\lambda}\left(DD^T\right)^{-1} D\left(b - x^0\right)\right)$ and $y^0 = p^0$.
  - For $k = 0, 1, 2, \ldots$ do the following (Until Convergence Criterion is met)
    1. Set $p^{k+1} = P_{\|p\|_\infty \le 1}\left(y^k - t_k\left(\left(\lambda^2 DD^T p^k\right) - (\lambda Db)\right)\right)$.
    2. Set $y^{k+1} = p^{k+1} + \frac{k}{k+2}\left(p^{k+1} - p^k\right)$.
    3. Set $x^{k+1} = b - \lambda D^T p^{k+1}$.
  - Set $x^* = b - \lambda D^T p^*$.
- Where step 3 is not mandatory.

SUGGESTED METHOD

Innovations – Using FISTA to accelerate the convergence

- To match the Article one should note

$$\left(\lambda^2 D D^T p^k\right) - (\lambda D b) = -\lambda D\left(b - \lambda D^T p^k\right) = -\lambda D x^k$$

  - Set $p^0 = P_{\|p\|_\infty \le 1}\left(\frac{1}{\lambda}\left(DD^T\right)^{-1} D\left(b - x^0\right)\right)$ and $y^0 = p^0, \;\; x^0 = x_{Init}$.
  - For $k = 0, 1, 2, \ldots$ do the following (Until Convergence Criterion is met)
    1. Set $p^{k+1} = P_{\|p\|_\infty \le 1}\left(y^k + t_k \lambda D x^k\right)$.
    2. Set $y^{k+1} = p^{k+1} + \frac{k+1}{k+2}\left(p^{k+1} - p^k\right)$.
    3. Set $x^{k+1} = b - \lambda D^T p^{k+1}$.
  - Set $x^* = b - \lambda D^T p^*$.
- Where step 3 is not mandatory.

SUGGESTED METHOD

Innovations – Adding constrain on the value of the output image

- Real World images have known bounded values ([0, 1], {0, 1, …, 255}).
- This prior knowledge should be used as a constraint.
- The proper way to use it is the apply Orthogonal Projection (Basically projecting onto a box) on Step 3 (Which becomes a must!).

$$\mathrm{x}^{k+1} = P_{B_{[u,l]}}\left(b - \lambda D^T p^{k+1}\right)$$

- The Projection Operator is given by

$$P_{B_{[u,l]}}(x)_i = \max\{\min\{x_i, u\}, l\}$$

SUGGESTED METHOD

Innovations – Using Denoising Operator for Deblurring

- Given the Deblurring Model

$$x^* = \arg\min_x \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|Dx\|_1$$

- Using the same "Trick" yields

$$\arg\max_{\|p\|_\infty \le 1} \min_x \left\{\tfrac{1}{2}\|Ax - b\|^2 + \lambda p^T D x\right\} \rightarrow x^* = (A^T A)^{-1}(A^T b - \lambda D^T p)$$

- Yet the term $A^T A$ might be singular which makes this solution not viable.
- If one would be given the $Prox_{tg}(x)$ of the given problem, the PGM would be

$$x^{k+1} = Prox_{t_k g}\left(x^k - t_k A^T\left(Ax^k - b\right)\right)$$

SUGGESTED METHOD

Innovations – Using Denoising Operator for Deblurring

- If one would be given the $Prox_{tg}(x)$ of the given problem, the PGM would be

$$x^{k+1} = Prox_{t_k g}\left(x^k - t_k A^T(Ax^k - b)\right)$$

- Writing it explicitly yields

$$x^{k+1} = \arg\min_y \left\{t_k g(y) + \frac{1}{2}\left\|y - \left(x^k + t_k A^T(Ax^k - b)\right)\right\|^2\right\}$$

- Which is exactly the Denoising Problem solved earlier with $\lambda_{Den} = t_k \lambda_{\mathrm{Deb}}$ and $b_{Den} = x^k + t_k A^T(Ax^k - b_{Deb})$.

SUGGESTED METHOD

Innovations – Using Denoising Operator for Deblurring

- Practical Notes
  - Keep of the previous Denoising iteration as initialization for the next. This will reduce the needed Denoising Internal Iterations.
  - The Step Size for the Deblurring is again by Quadratic Form. Namely $t_k \leq \frac{1}{\lambda_{max}(A^T A)}$. Can we do better? Maybe! But not this time…

SUGGESTED METHOD

Innovations – Monotonic FISTA

- For the Deblurring Process "Jumps" in the minmization process might cause issues of convergence (The Denoising Solution isn't accurate). Monotonic Property of the Solver would improve results greatly.

FISTA

- Set $y_0 = x_0$ and $t_0 = 1$.
- Step $k = 0, 1, 2, \ldots$

  Set $x_{k+1} = Prox_{Lg}(y_k)$.

  Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ or $t_{k+1} = \frac{k+1}{2}$.

  Set $y_{k+1} = x_k + \left(\frac{t_{k-1}}{t_{k+1}}\right)(x_{k+1} - x_k)$.

Monotone FISTA (MFISTA)

Same Convergence Rate!

- Set $y_0 = x_0$ and $t_0 = 1$.
- Step $k = 0, 1, 2, \ldots$

  Set $z_k = Prox_{Lg}(y_k)$.

  Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ or $t_{k+1} = \frac{k+1}{2}$.

  Set $x_{k+1} = \arg\min_x \{H(x_k), H(z_k)\}$.

  Set $y_{k+1} = x_{k+1} + \left(\frac{t_k}{t_{k+1}}\right)(z_k - x_{k+1}) + \left(\frac{t_{k-1}}{t_{k+1}}\right)(x_{k+1} - x_k)$.

SUGGESTED METHOD

Innovations – Monotonic FISTA

- What happens for any case of evaluation?

Prox Is Monotonic

Monotone FISTA (MFISTA)

- Set $y_0 = x_0$.
- Step $k = 0, 1, 2, \ldots$

  Set $z_k = Prox_{Lg}(y_k)$.

  Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ or $t_{k+1} = \frac{k+1}{2}$.

  Set $x_{k+1} = x_k$.

  Set $y_{k+1} = x_k + \left(\frac{t_k}{t_{k+1}}\right)(z_k - x_k)$.

FISTA!

Monotone FISTA (MFISTA)

- Set $y_0 = x_0$.
- Step $k = 0, 1, 2, \ldots$

  Set $z_k = Prox_{Lg}(y_k)$.

  Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ or $t_{k+1} = \frac{k+1}{2}$.

  Set $x_{k+1} = z_k$.

  Set $y_{k+1} = z_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(z_k - x_k)$.

RESULTS

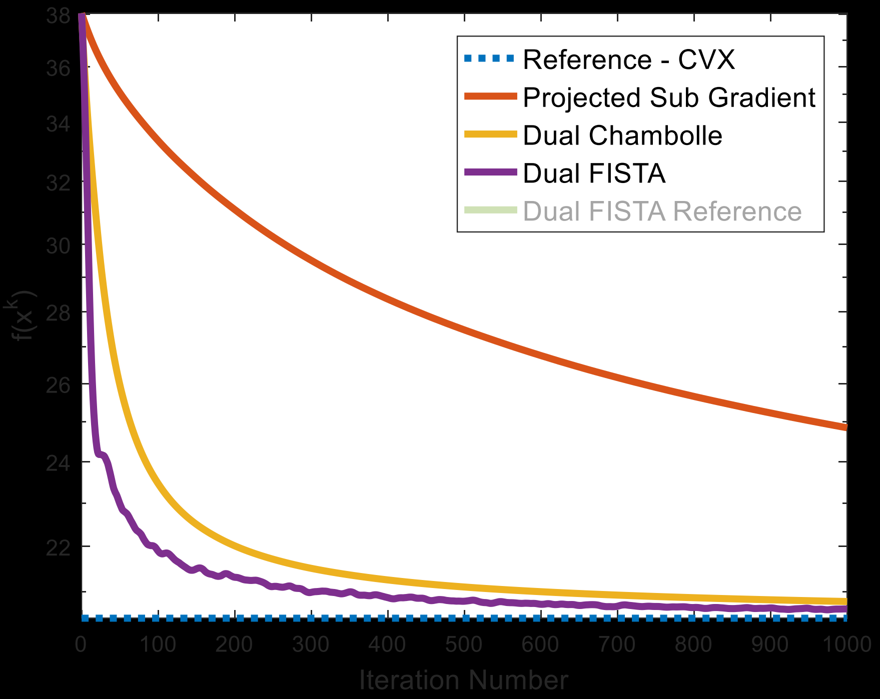
Denoising

Reference Image
Noisy Image

CVX
Sub Gradient

Dual Chambolle
Dual FISTA

Dual FISTA Ref
Reference Image



RESULTS

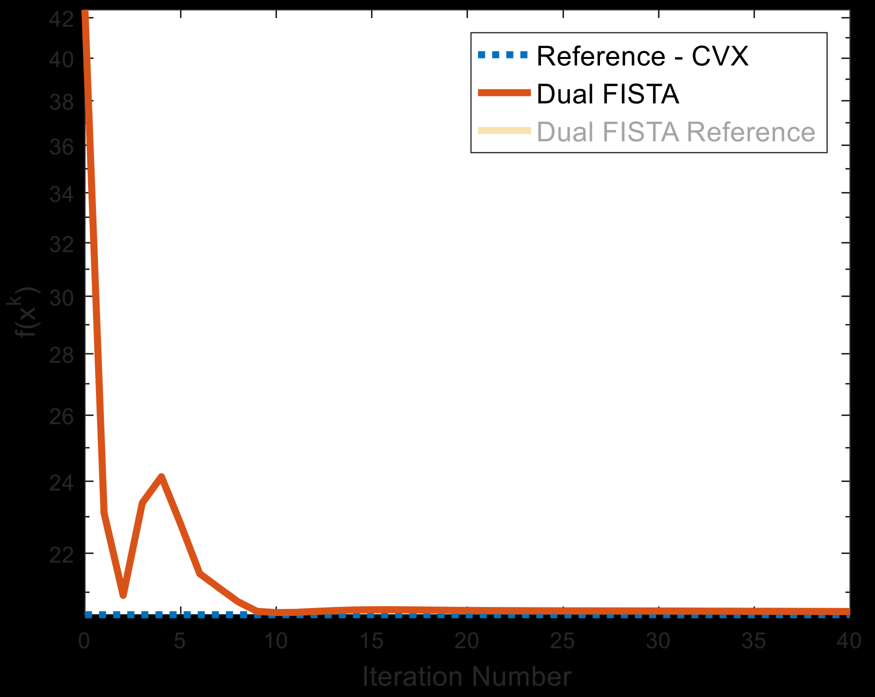
Deblurring

Reference Image

Blurred Image

CVX

FISTA By Denoising



REMARKS

Convergence Analysis

- The PGM suggest Sub Linear Convergence Rate

$$f\left(x^k\right) - f(x^*) \leq \frac{L\left\|x^0 - x^*\right\|^2}{2k}$$

- Yet, in practice the convergence rate is faster. This is due to $f(x)$ being Strongly Convex.
- Define (As in Lecture Notes)

$$H(x) = f(x) + g(x)$$

- Using Gradient Inequality

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

- Using Strongly Convex Property

$$f(y) \geq g(x) + \langle \nabla f(x), y - x \rangle + \frac{s}{2}\|y - x\|^2$$

- Combining them at $y = x - \frac{1}{L}G_L(x)$ yields

$$\frac{s}{2L^2}\|G_L(x)\|^2 \leq f\left(x - \frac{1}{L}G_L(x)\right) - f(x) + \frac{1}{L}\langle \nabla f(x), G_L(x) \rangle \leq \frac{1}{2L}\|G_L(x)\|^2$$

REMARKS

Convergence Analysis

- In similar way to lecture notes what would see that

$$H\left(x^{k+1}\right) \leq H(y) + \langle G_L(x), x - y\rangle - \frac{1}{2L}\|G_L(x)\|^2 - \frac{s}{2}\|x - y\|^2$$

- Same derivation as in Lecture Notes 007 Page 010 yields

$$H\left(x^{k+1}\right) - H(x^*) \leq \frac{1}{2L}\left(\left(1 - \frac{s}{L}\right)\left\|x^* - x^{k+1}\right\|^2 - \left\|x^* - x^k\right\|^2\right)$$

- Since one $H\left(x^{k+1}\right) \geq H(x^*)$ could see that

$$\left\|x^* - x^k\right\|^2 \leq \left(1 - \frac{s}{L}\right)\left\|x^* - x^{k+1}\right\|^2$$

- Yet this holds for k hence

$$\left\|x^* - x^k\right\|^2 \leq \left(1 - \frac{s}{L}\right)^k \|x^* - x^0\|^2$$

REMARKS

Min Max Switching

- Have a second look at

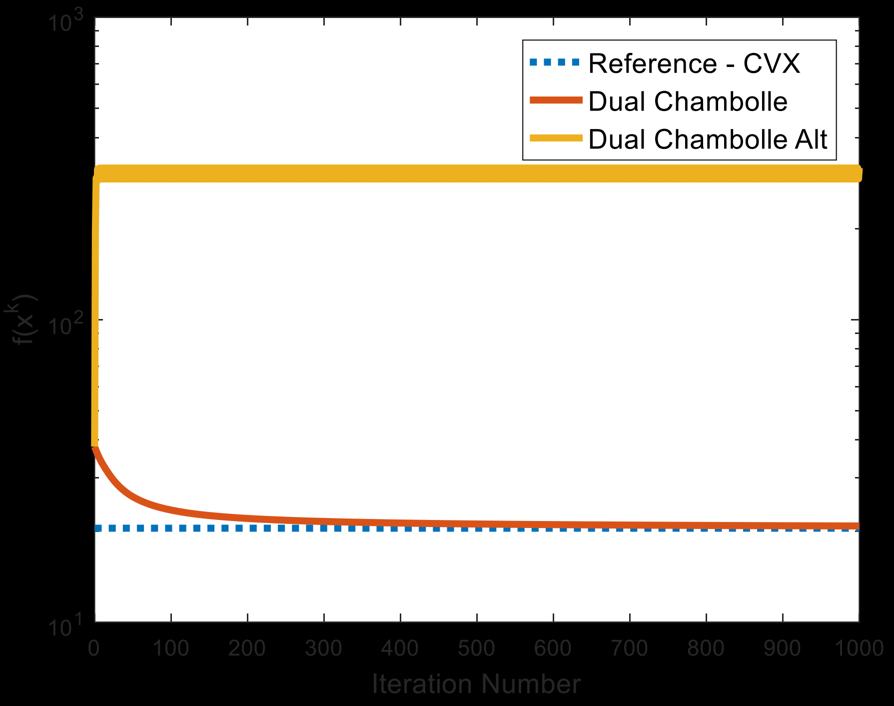
$$\arg\min_{x} \max_{||p||_\infty \le 1} \left\{ \frac{1}{2}||x-b||^2 + \lambda p^T D x \right\}$$

- It is clear that $sgn(Dx)$ maximizes the function and obeys the constraints. Yet, by switching (Min Max Theorem) the result is different.
  - Why isn't the result the same using Min Max Theorem?
    The solution isn't unique (Basically it is the Support Function $\sigma_{||p||_\infty \le 1}(Dx)$ and this is not guaranteed to be unique).
  - Can we solve the problem using this step instead?
    Better to still switch (No one wants to deal with the Sign Function).
    Then, have $p^* = sgn(Dx^*)$.

REMARKS

Min Max Switching

Let's Try It…



CONCLUSIONS

- The article added Value Constraints (Box Constraints) to the TV Denoising / Deblurring Problem which improves results with negligible computational cost (Projection into a Box).
- The article suggest a framework to accelerate the TV Denoising method without any change of the “Cost Function”.
- The article suggests a framework which enables solving the TV Deblurring problem.
- The article derived a Monotone version of FISTA which greatly improves the Deblurring results with negligible computational cost (2 Calculations of the Cost Function).
- Idea – One might be able to us the same “Deblurring” trick to solve the Inpainting Problem as Inpainting can also be described by A.