Nonconvex splitting algorithms and video decomposition

Rick Chartrand

Los Alamos National Laboratory

July 26, 2011

In theory, there's no difference between theory and practice. In practice, there is.

–Yogi Berra

Outline

Motivating example

A convex splitting

Nonconvex splitting

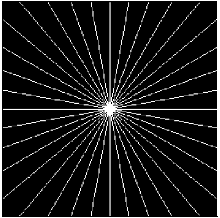
Examples

Nonconvexity is better

We reconstruct an image from samples of its Fourier transform:

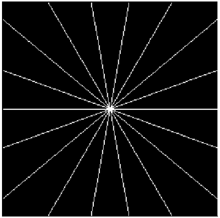
$$\min_{\boldsymbol{x}} F(\nabla x), \text{ subject to } \hat{x}|_\Omega = \hat{s}|_\Omega.$$

test image $\boldsymbol{s}$



18 lines/7% sampled

recon., $\boldsymbol{F} = \|\cdot\|_1$



9 lines/3.5% sampled

recon., $\boldsymbol{F} = \|\cdot\|_1$

recon., nonconvex $\boldsymbol{F}$

Outline

Splitting algorithm

Consider the following, suitable for denoising, deblurring, compressive sensing, etc.:

$$\min_{x} \|\nabla x\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2.$$

Splitting algorithm

Consider the following, suitable for denoising, deblurring, compressive sensing, etc.:

$$\min_{x} \|\nabla x\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2.$$

Now we introduce an auxiliary variable[1]:

$$\min_{w,x} \|w\|_1 + \frac{1}{2\lambda}\|w - \nabla x\|_2^2 + \frac{\mu}{2}\|Ax - b\|_2^2.$$

[1] J. Yang, W. Yin, Y. Zhang, Y. Wang, *SIAM J. Imaging Sci.*, 2009

Splitting algorithm

Consider the following, suitable for denoising, deblurring, compressive sensing, etc.:

$$\min_{x} \|\nabla x\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2.$$

Now we introduce an auxiliary variable[1]:

$$\min_{w,x} \|w\|_1 + \frac{1}{2\lambda}\|w - \nabla x\|_2^2 + \frac{\mu}{2}\|Ax - b\|_2^2.$$

This Douglas-Rachford splitting[2] decouples the objective function from the operators.

[1] J. Yang, W. Yin, Y. Zhang, Y. Wang, *SIAM J. Imaging Sci.*, 2009
[2] S. Setzer, *Int. J. Comput. Vision*, 2011

Moreau envelope

Minimization proceeds by alternation, iterating:

$$\min_{x}\left(\min_{w}\left[\|w\|_1 + \frac{1}{2\lambda}\|w - \nabla x\|_2^2\right] + \frac{\mu}{2}\|Ax - b\|_2^2\right).$$

Moreau envelope

Minimization proceeds by alternation, iterating:

$$\min_x \left(\min_w \left[\|w\|_1 + \frac{1}{2\lambda} \|w - \nabla x\|_2^2 \right] + \frac{\mu}{2} \|Ax - b\|_2^2 \right).$$

The Moreau envelope e_λ of $\| \cdot \|_1$ tells us what the objective function is:

$$\min_w \|w\|_1 + \frac{1}{2\lambda} \|w - \nabla x\|_2^2 = e_\lambda \| \cdot \|_1 (\nabla x) = H_{\lambda,1}(\nabla x),$$

Moreau envelope

Minimization proceeds by alternation, iterating:

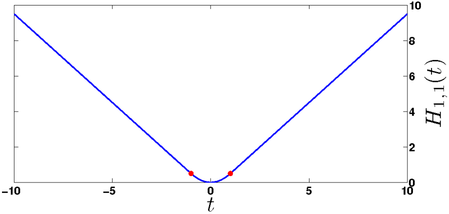
$$\min_{x}\left(\min_{w}\left[\|w\|_1 + \frac{1}{2\lambda}\|w - \nabla x\|_2^2\right] + \frac{\mu}{2}\|Ax - b\|_2^2\right).$$

The Moreau envelope e_λ of $\|\cdot\|_1$ tells us what the objective function is:

$$\min_{w}\|w\|_1 + \frac{1}{2\lambda}\|w - \nabla x\|_2^2 = e_\lambda\|\cdot\|_1(\nabla x) = H_{\lambda,1}(\nabla x),$$

where $H_{\lambda,1}(\vec{t})_k = h_{\lambda,1}(t_k)$ is the (componentwise) Huber function:

$$h_{\lambda,1}(t) = \begin{cases} |t|^2/2\lambda, & \text{if } |t| \le \lambda, \\ |t| - \lambda/2, & \text{if } |t| \ge \lambda. \end{cases}$$



Simple iterations

Solving for $\boldsymbol{w}$ is separable and easy:

$$\boldsymbol{w = P_\lambda \|\cdot\|_1 (\nabla x) = \min\{0, |\nabla x| - \lambda\} \frac{\nabla x}{|\nabla x|}.}$$

This is shrinkage or soft thresholding.

Simple iterations

Solving for w is separable and easy:

$$\boldsymbol{w = P_\lambda \| \cdot \|_1 (\nabla x) = \min\{0, |\nabla x| - \lambda\} \frac{\nabla x}{|\nabla x|}.}$$

This is shrinkage or soft thresholding.

Solving for x leads to a linear equation:

$$\boldsymbol{\left(\frac{1}{\lambda}\nabla^* \nabla + \mu A^* A\right) x = \frac{1}{\lambda}\nabla^* w + \mu A^* b.}$$

In many cases, this can be solved in the Fourier domain, at the cost of two FFTs.

Outline

Nonconvexification

In the objective function $H_{\lambda,1} = e_{\lambda}\|\cdot\|_1$, we can generalize in two different ways:

Nonconvexification

In the objective function $H_{\lambda,1} = e_\lambda \| \cdot \|_1$, we can generalize in two different ways:

- Use $e_\lambda \| \cdot \|_p^p$, with $0 < p < 1$[3]:

$$\min_{w,x} \|w\|_p^p + \frac{1}{2\lambda}\|w - \nabla x\|_2^2 + \frac{\mu}{2}\|Ax - b\|_2^2.$$

[3] Krishnan and Fergus, *Neural Information Processing Systems*, 2009

Nonconvexification

In the objective function $H_{\lambda,1} = e_\lambda \|\cdot\|_1$, we can generalize in two different ways:

- Use $e_\lambda \|\cdot\|_p^p$, with $0 < p < 1$[3]:

$$\min_{w,x} \|w\|_p^p + \frac{1}{2\lambda}\|w - \nabla x\|_2^2 + \frac{\mu}{2}\|Ax - b\|_2^2.$$

  Solving for w is no longer a shrinkage, and can only be done analytically for special values of p.

[3] Krishnan and Fergus, *Neural Information Processing Systems*, 2009

Nonconvexification

In the objective function $H_{\lambda,1} = e_\lambda \|\cdot\|_1$, we can generalize in two different ways:

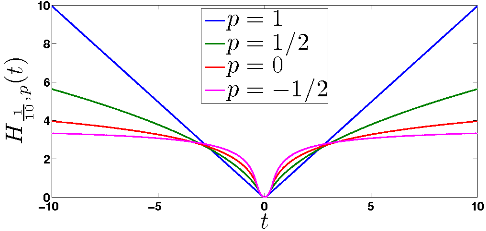
- Use $e_\lambda \|\cdot\|_p^p$, with $0 < p < 1$[3]:

$$\min_{w,x} \|w\|_p^p + \frac{1}{2\lambda}\|w - \nabla x\|_2^2 + \frac{\mu}{2}\|Ax - b\|_2^2.$$

  Solving for w is no longer a shrinkage, and can only be done analytically for special values of p.

- Use $H_{\lambda,p}$, for any $p < 1$[4]:

$$h_{\lambda,p}(t) = \begin{cases} |t|^2/2\lambda, & \text{if } |t| \le \lambda^{\frac{1}{2-p}}, \\ \frac{1}{p}|t|^p - \delta, & \text{if } |t| \ge \lambda^{\frac{1}{2-p}}. \end{cases}$$



[3] Krishnan and Fergus, *Neural Information Processing Systems*, 2009

[4] C., *International Symposium on Biomedical Imaging*, 2009

(Non)convex analysis

Is there some $G_{\lambda,p}$ such that $H_{\lambda,p} = e_\lambda G_{\lambda,p}$?

(Non)convex analysis

Is there some $G_{\lambda,p}$ such that $H_{\lambda,p} = e_\lambda G_{\lambda,p}$?

$$h(t) = \min_s g(s) + \frac{1}{2\lambda}|s - t|^2$$

is equivalent to

$$\frac{|t|^2}{2} - \lambda h(t) = \left(\frac{|\cdot|^2}{2} + \lambda g\right)^*(t). \tag{☺}$$

(Non)convex analysis

Is there some $G_{\lambda,p}$ such that $H_{\lambda,p} = e_\lambda G_{\lambda,p}$?

$$h(t) = \min_s g(s) + \frac{1}{2\lambda}|s-t|^2$$

is equivalent to

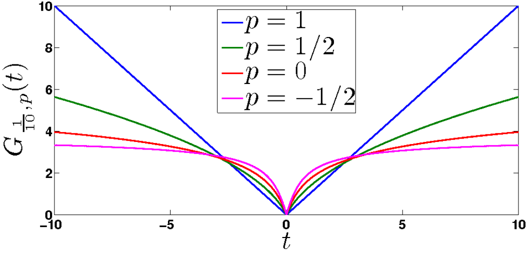
$$\frac{|t|^2}{2} - \lambda h(t) = \left(\frac{|\cdot|^2}{2} + \lambda g\right)^*(t). \qquad (☺)$$

Since $|t|^2/2 - \lambda h_{\lambda,p}(t)$ is convex by construction, we can define $g_{\lambda,p}$ by

$$\frac{|s|^2}{2} + \lambda g_{\lambda,p}(s) = \left(\frac{|\cdot|^2}{2} - \lambda h_{\lambda,p}\right)^*(s),$$

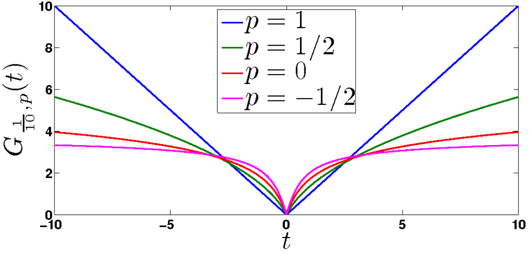
and (☺) follows.

Proximal analog of $\| \cdot \|_1$



$$\min_{x}\left(\min_{w}\left[G_{\lambda,p}(w) + \frac{1}{2\lambda}\|w - \nabla x\|_2^2\right] + \frac{\mu}{2}\|Ax - b\|_2^2\right).$$

Proximal analog of $\|\cdot\|_1$



$$\min_{x}\left(\min_{w}\left[G_{\lambda,p}(w) + \frac{1}{2\lambda}\|w - \nabla x\|_2^2\right] + \frac{\mu}{2}\|Ax - b\|_2^2\right).$$

The solution for w is a p-shrinkage:

$$w = P_\lambda G_{\lambda,p}(\nabla x) = \min\{0, |\nabla x| - \lambda|\nabla x|^{p-1}\}\frac{\nabla x}{|\nabla x|}.$$

Method of multipliers

We can enforce equality with the method of multipliers (also called split Bregman[5] in this context):

$$\min_{w,x} G_{\lambda,p}(w) + \frac{1}{2\lambda}\|w - \nabla x - \Lambda_1\|_2^2 + \frac{\mu}{2}\|Ax - b - \Lambda_2\|_2^2,$$

where at each iteration we update

$$\Lambda_1^{n+1} = \Lambda_1^n + c_n(w^n - \nabla x^n),$$

$$\Lambda_2^{n+1} = \Lambda_2^n + d_n(b - Ax^n).$$

[5] T. Goldstein and S. Osher, *SIAM J. Imaging Sci.*, 2009

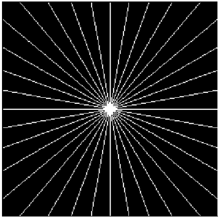
Outline

Nonconvexity is better

We reconstruct an image from samples of its Fourier transform:

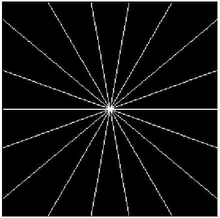
$$\min_{x} F(\nabla x), \text{ subject to } \hat{x}|_{\Omega} = \hat{s}|_{\Omega}.$$

test image s



18 lines/7% sampled

recon., $F = \|\cdot\|_1$



9 lines/3.5% sampled

recon., $F = \|\cdot\|_1$

$F = H_{\lambda,-1/2}$

Low rank + sparse decomposition

We seek to decompose a matrix $\boldsymbol{D}$ of high-dimensional data into low-rank and sparse components:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \operatorname{rank}(\boldsymbol{L}) + \boldsymbol{\mu}\|\boldsymbol{S}\|_{\boldsymbol{0}}, \text{ subject to } \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{D}.$$

Low rank + sparse decomposition

We seek to decompose a matrix $\boldsymbol{D}$ of high-dimensional data into low-rank and sparse components:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \mathbf{rank}(\boldsymbol{L}) + \boldsymbol{\mu}\|\boldsymbol{S}\|_{\mathbf{0}}, \text{ subject to } \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{D}.$$

- $\boldsymbol{L}$ represents a low-dimensional model of the data, as in PCA.

Low rank + sparse decomposition

We seek to decompose a matrix D of high-dimensional data into low-rank and sparse components:

$$\min_{L,S} \operatorname{rank}(L) + \mu\|S\|_0, \text{ subject to } L + S = D.$$

- L represents a low-dimensional model of the data, as in PCA.
- S consists of possibly-large discrepancies from the data. S can contain useful information, while allowing the model L to be more robust[6].

[6] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, *Neural Information Processing Systems*, 2009

Low rank + sparse decomposition

We seek to decompose a matrix D of high-dimensional data into low-rank and sparse components:

$$\min_{L,S} \operatorname{rank}(L) + \mu\|S\|_0, \text{ subject to } L + S = D.$$

- L represents a low-dimensional model of the data, as in PCA.
- S consists of possibly-large discrepancies from the data. S can contain useful information, while allowing the model L to be more robust[6].

A tractable approximation is to solve[7]:

$$\min_{L,S} \sum_k g_{\lambda,p}(\sigma_k(L)) + \mu G_{\lambda,p}(S) + \frac{\mu}{2\lambda}\|D - L - S - \Lambda\|_F^2.$$

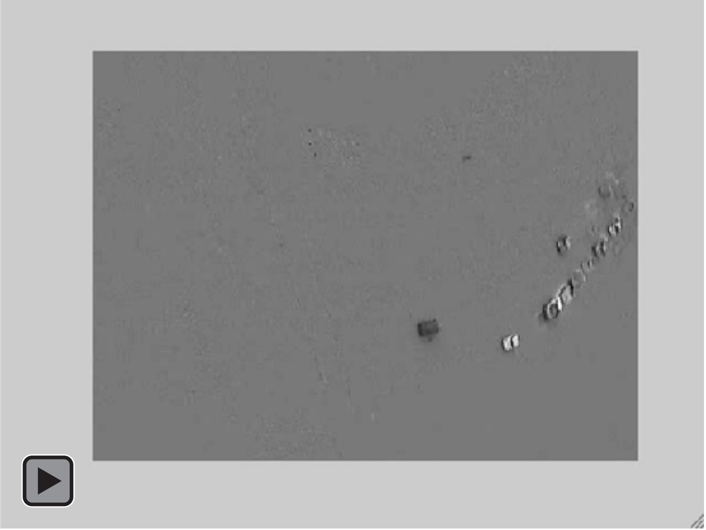
[6] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, *Neural Information Processing Systems*, 2009

[7] Z. Lin, M. Chen, Y. Ma, preprint, 2010 ($p = 1$ case)

Video example

video D, 240×320 pixels, 288 frames

Video example



sparse component S

Video example

low-rank component L

Noisy data

Now we penalize the 3D total (q-)variation of S, with $q \leq 1$:

$$\min_{L,S,V,W} \sum_k g_{\lambda,p}(\sigma_k(L)) + \alpha G_{\mu,p}(V) + \beta G_{\nu,q}(W)$$
$$+ \frac{1}{2\lambda}\|D - L - S\|_F^2 + \frac{\alpha}{2\mu}\|V - S - \Lambda_1\|_F^2 + \frac{\beta}{2\nu}\|W - \nabla S - \Lambda_2\|_F^2.$$

Noisy data

Now we penalize the 3D total (q-)variation of S, with $q \leq 1$:

$$\min_{L,S,V,W} \sum_k g_{\lambda,p}(\sigma_k(L)) + \alpha G_{\mu,p}(V) + \beta G_{\nu,q}(W)$$
$$+ \frac{1}{2\lambda}\|D - L - S\|_F^2 + \frac{\alpha}{2\mu}\|V - S - \Lambda_1\|_F^2 + \frac{\beta}{2\nu}\|W - \nabla S - \Lambda_2\|_F^2.$$

Solving for L, V, and W is done by shrinkage, and the quadratic problem for S can be solved using two FFTs.

Preserving noisy shapes

Why $q < 1$?

Preserving noisy shapes

Why $q < 1$?

$\int |\nabla u|$ penalizes the length of edges in 2D, or the surface area of boundaries in 3D. This can distort shapes.

Preserving noisy shapes

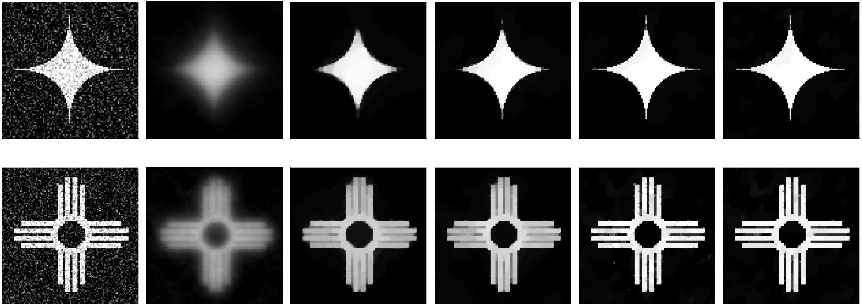
Why $q < 1$?

$\int |\nabla u|$ penalizes the length of edges in 2D, or the surface area of boundaries in 3D. This can distort shapes. $\int |\nabla u|^q$, in contrast, penalizes a $2 - q$ dimensional analog of length instead[8], or a $3 - q$ dimensional analog of area, thus preserving more complex shapes.

[8] C., *International Conference on Image Processing*, 2007

Preserving noisy shapes

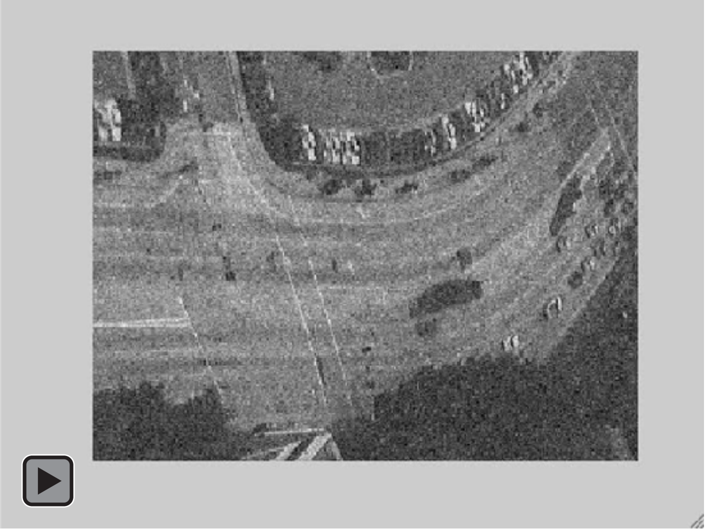
Why $q < 1$?
$\int |\nabla u|$ penalizes the length of edges in 2D, or the surface area of boundaries in 3D. This can distort shapes. $\int |\nabla u|^q$, in contrast, penalizes a $2 - q$ dimensional analog of length instead[8], or a $3 - q$ dimensional analog of area, thus preserving more complex shapes.



noisy image | $q = 2$ | $q = 1$ | $q = 3/4$ | $q = 1/2$ | $q = 1/4$

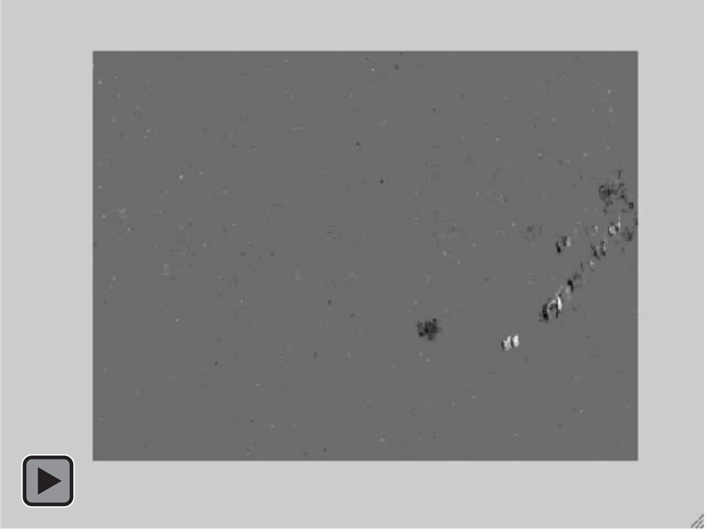
[8] C., *International Conference on Image Processing*, 2007

Noisy video example



noisy video D

Noisy video example

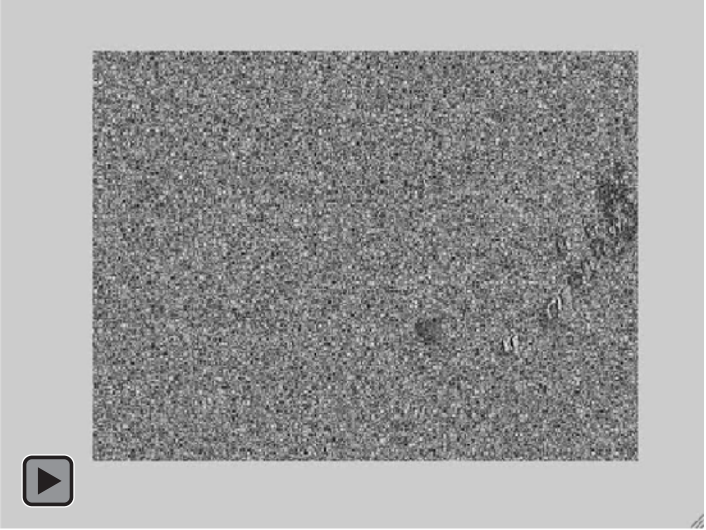


sparse component S, $q = 1/2$

Noisy video example

low-rank component L

Noisy video example



residual; SNR of $L + S$ is 16.9 dB for $q = 1/2$, 15.7 dB for $q = 1$

Summary

- Nonconvex optimization gives better results for sparse recovery (in practice).
- State-of-the-art convex optimization methods can be extended to the nonconvex case, giving excellent computational efficiency.

`math.lanl.gov/~rick`