Smoothing techniques for convex problems. Applications in image processing.

Pierre Weiss $^{(1)}$, Mikaël Carlavan $^{(2)}$, Laure Blanc-Féraud $^{(2)}$ and Josiane Zerubia $^{(2)}$

(1) Institute for Computational Mathematics, Kowloon Tong, Hong Kong.
(2) Projet Ariana - CNRS/INRIA/UNSA, 2004 route des Lucioles, 06902 Sophia-Antipolis, France.
(1) pierre.armand.weiss@gmail.com, (2) firstname.lastname@sophia.inria.fr

Abstract:

In this paper, we present two algorithms to solve some inverse problems coming from the field of image processing. The problems we study are convex and can be expressed simply as sums of l^p-norms ($p \in \{1, 2, \infty\}$) of affine transforms of the image. We propose 2 different techniques. They are - to the best of our knowledge - new in the domain of image processing and one of them is new in the domain of mathematical programming. Both methods converge to the set of minimizers. Additionally, we show that they converge at least as $O\left(\frac{1}{N}\right)$ (where N is the iteration counter) which is in some sense an "optimal" rate of convergence. Finally, we compare these approaches to some others on a toy problem of image super-resolution with impulse noise.

1. Introduction

Many image processing tasks like reconstruction or segmentation can be done efficiently by solving convex optimization problems. Recently these models received considerable attention and this led to some breakthrough. Among them are the new sampling theorems [5] and the impressive results obtained using sparsity or regularity assumptions in image reconstruction (see e.g. [4]).

These results motivate an important research to accelerate the convergence speed of the minimization schemes. In the last decade, many algorithms like iterative thresholding or dual approaches were reinvented by the "imaging community" (see for instance [2, 3] for old references). Recently, the "mathematical programming community" got interested in those problems and it led to some drastic improvements. As examples let us cite the papers by Y. Nesterov [9, 10] and M. Teboulle [1] which improve by one order of magnitude most first order approaches.

In this paper, we mainly follow the lines of Y. Nesterov [9]. We consider the problem of minimizing the sum of l^p-norms ($p \in \{1, 2, \infty\}$) of affine transforms of the image. The general mechanism of the algorithms we propose consists in smoothing the problem and solve it with an efficient first order scheme. Our contribution is mainly to extend the results of [9] to a more general setting and to propose a dual variant which behaves better in all problems we tested. We also give convergence rates for the proposed algorithms. We believe, this gives some insight on the important factors that influence the algorithms efficiency and helps designing solvable problems.

2. The problems considered

In this paper, we consider the following seminal model of image deterioration:

$$u^0 = Du + b \tag{1}$$

where u is an original, neat image, $D : \mathbb{R}^n \to \mathbb{R}^m$ is some known linear transform, $b \in \mathbb{R}^m$ is some additive noise and $u^0 \in \mathbb{R}^m$ is a given observed image. This simple formalism actually models many real situations. For instance, D can be an irregular sampling and a convolution. In this case recovering u from u^0 is a super-resolution problem [7]. Other applications include image inpainting, compression noise reduction, texture+cartoon decompositions, reconstruction from noisy indirect measurements...

Finding u from the observation u^0 is an inverse problem. There exists many ways to solve it. In this paper, we concentrate on two variational models. The first one consists in solving the following convex problem:

$$\min_{x \in X} \left(\underbrace{||Bx||_1 + \lambda ||Dx - u^0||_p}_{\Psi(x)} \right). \tag{2}$$

The second one consists in solving:

$$\min_{y \in Y} \left(||y||_1 + \lambda ||DB^* y - u^0||_p \right). \tag{3}$$

In both problems, $B : \mathbb{R}^n \to \mathbb{R}^o$ is a linear transform, $|| \cdot ||_p$ denotes the standard l^p-norm and X and Y are simple convex sets (like $\mathbb{R}^n$ or $[0, 1]^n$).

The interpretation of the first model is as follows: we look for an image x which minimizes $||Bx||_1$ such that Dx is close to u^0. The function $x \mapsto ||Bx||_1$ can be seen as a *regularity* a priori on the image. For instance, if B is the discrete gradient, then it corresponds to the total variation. If B is some wavelet transform, it is equivalent to a Besov semi-norm [6]. p must be chosen depending on the statistics of the additive noise. For instance, p should be equal to 2 for Gaussian noise, to 1 for impulse noise and to ∞ for uniform noise.

The interpretation of the second model is the following: we look for a decomposition y of the restored image in some dictionary B^* such that its reconstruction B^*y is close to u^0. Minimizing the l^1-norm of y is known to favor sparse structures. The underlying assumption is thus that the original image u is sparse in the dictionary B^*.

From a numerical point of view, both problems are very similar. However, the first one is slightly more general and complicated than the second. We will thus give a detailled analysis of its resolution and only provide numerical results for the second one.

The remaining of the paper is as follows. We first present an algorithm based on a regularization of the primal problem (2). Then we present a technique to regularize a dual version of (2). Finally we propose theoretical and numerical comparisons of both techniques on a problem of image super-resolution. Due to space limitations, we only provide the main ideas in this paper. We refer the reader to [12] (in French), for the proofs of the propositions.

3. Smoothing of the primal problem

In this section, we propose a method to minimize (2). Its principle is exactly the same as the method proposed by Y. Nesterov in [9]:

1. Smooth the non-differentiable terms in (2).
2. Solve the regularized problem using an accelerated gradient method.

The only difference is that we do not require the set X to be bounded, which requires a slightly different analysis. Now let us present some details of this approach.

A key observation to solve (2) is that it can be rewritten as a so called min-max problem. Let p' denote the conjugate of p $\left(\text{i.e. } \frac{1}{p'} + \frac{1}{p} = 1\right)$. We can rewrite problem (2) as follows:

$$\min_{x\in X}\left(\max_{y\in Y}\left(\langle Bx, y_1\rangle + \lambda\langle Dx - u^0, y_2\rangle\right)\right) \quad (4)$$

$$= \min_{x\in X}\left(\underbrace{\max_{y\in Y}\left(\langle Ax - h, y\rangle\right)}_{\Psi(x)}\right) \quad (5)$$

where $\langle\cdot,\cdot\rangle$ denotes the canonical scalar product,

$$A = \begin{bmatrix} B \\ \lambda D \end{bmatrix}, \quad h = \begin{bmatrix} 0 \\ \lambda u^0 \end{bmatrix} \text{ and} \quad (6)$$

$$Y = \{y = (y_1, y_2) \in \mathbb{R}^o \times \mathbb{R}^m, ||y_1||_\infty \le 1, ||y_2||_{p'} \le 1\}. \quad (7)$$

The function Ψ is a conjugate function and the set Y is bounded. It can thus be smoothed using a Moreau regularization. Let us denote:

$$\Psi_\mu(x) = \max_{y\in Y}\left(\langle Ax - h, y\rangle - \frac{\mu}{2}||y||_2^2\right). \quad (8)$$

This function can be shown to be L-Lipschitz differentiable:

$$||\nabla\Psi_\mu(x_1) - \nabla\Psi_\mu(x_2)||_2 \le L||x_1 - x_2||_2 \quad (9)$$

with $L = \frac{|||A|||^2}{\mu}$ and $|||A||| = \max_{x\in\mathbb{R}^n, ||x||_2\le 1}(||Ax||_2)$. Furthermore, it is a good uniform approximation of Ψ in the following sense:

$$0 \le \Psi(x) - \Psi_\mu(x) \le \frac{\mu}{2}D. \quad (10)$$

where $D = \left(\max_{y\in Y}\left(||y||_2^2\right)\right)$. Thus, we can make the difference between Ψ and Ψ_μ as small as desired by decreasing μ. The approximation Ψ_μ is actually very common in image processing. For instance, when $p = 1$, it corresponds to the approximation of the absolute value by a Huber function. When $p = \infty$ it is slightly more difficult, but it can still be computed in closed form.

The smoothed problem writes:

$$\min_{x\in X}\left(\Psi_\mu(x)\right). \quad (11)$$

It consists in minimizing a differentiable function over a simple set. We can thus apply projected gradient like algorithms to solve it. Unfortunately, μ has to be chosen small in order to get a good approximate solution. This requires to use small step sizes in the gradient descent and thus results in a very slow convergence rate. The main observation of Y. Nesterov in [9] is that using an accelerated version of the projected gradient methods can actually compensate the approximation error. This results in a convergence rate in $O\left(\frac{1}{N}\right)$ (where N is the iteration counter), while other first order approaches like projected subgradient descents converge as $O\left(\frac{1}{\sqrt{N}}\right)$.

Now let us write down the complete algorithm to solve (11). Let x_μ^* denote a solution of (11) (it is not unique in general). We propose the following algorithm:

Algorithm 1 (Primal)

```
Choose a number of iterations $N$.
Set a starting point $x^0$ (as close as possible to $x_\mu^*$).
Set $\mu = \mu(N) = \frac{|||A|||\cdot||x^0 - x_\mu^*||_2}{N}$.
Set $\mathcal{A} = 0$, $g = 0$ and $x = x^0$.
for $k = 0$ to $N$ do
    $a = \frac{1}{L} + \sqrt{\frac{1}{L^2} + \frac{2}{L}\mathcal{A}}$
    $v = \Pi_X\left(x^0 - g\right)$
    $y = \frac{\mathcal{A}x + av}{\mathcal{A} + a}$
    $x = \Pi_X\left(y - \frac{\nabla\Psi_\mu(y)}{L}\right)$
    $g = g + a\nabla\Psi_\mu(x)$
    $\mathcal{A} = \mathcal{A} + a$
end for
Set $x^N = x$.
```

Our main convergence results are as follows. Let x^* denote a solution of (2).

Proposition 1 *x^N converges to the set of minimizers of (2).*

Proposition 2 *The worst case convergence rate is:*

$$\Psi(x^N) - \Psi(x^*) \le \frac{2|||A|||\cdot||x^0 - x_\mu^*||_2\sqrt{D}}{N}. \quad (12)$$

Note that the distance $||x^0 - x^*_\mu||_2$ is unknown in general, so that Algorithm 1 might not seem implementable. In the case where X is a compact set, this quantity can be bounded above by the diameter of X. When X is not bounded, it actually suffices to choose μ of order $\frac{|||A|||}{N}$ to get a precision of order $O\left(\frac{1}{N}\right)$. Algorithm (1) is thus implementable and converges as $O\left(\frac{1}{N}\right)$. This convergence rate is neatly sublinear and might seem bad at first sight. Actually, it is somehow optimal. Indeed, A. Nemirovski shows in [8] that some instances of problems like (5) cannot be solved with a better rate of convergence than $O\left(\frac{1}{N}\right)$ using first order methods.

4. Smoothing of the dual problem

In this section, we propose an approach alternative to the previous one. Its flavor is similar to a proximal-method. One way to understand this scheme is that we smooth the "dual" problem instead of the primal problem. Note that the $\min$ and the $\max$ in equation (5) cannot be inverted as we do not suppose X to be compact. So we cannot use - properly speaking - the term dual problem.
Instead of solving (2), we solve:

$$\min_{x\in X}\left(||Bx||_1 + \lambda||Dx - u^0||_p + \frac{\epsilon}{2}||x - x^0||_2^2\right) \quad (13)$$

where $\epsilon \in \mathbb{R}^+_*$ and x^0 should be chosen close to the set of minimizers of (2). It can be shown that as ϵ goes to 0, the unique solution of (13) converges to the Euclidean projection of x^0 onto the set of minimizers of (2). We can rewrite (13) as a min-max problem:

$$\begin{aligned} & \min_{x\in X}\left(\max_{y\in Y}\left(\langle Ax - h, y\rangle\right) + \frac{\epsilon}{2}||x - x^0||_2^2\right) & (14)\\ = \; & \max_{y\in Y}\left(\underbrace{\min_{x\in X}\left(\langle Ax - h, y\rangle + \frac{\epsilon}{2}||x - x^0||_2^2\right)}_{\Psi_\epsilon(y)}\right) & (15)\end{aligned}$$

Note that we can invert the $\min$ and the $\max$ only because the term $\frac{\epsilon}{2}||x - x^0||_2^2$ makes the problem coercive in x. Now, the important observation is that the function Ψ_ϵ is the conjugate of a strongly convex function. It is thus concave and Lipschitz differentiable:

$$||\nabla\Psi_\epsilon(y_1) - \nabla\Psi_\epsilon(y_2)||_2 \leq L||x_1 - x_2||_2 \quad (16)$$

$\forall(y_1, y_2) \in Y \times Y$ with $L \leq \frac{|||A|||^2}{\epsilon}$. Problem (15) consists in maximizing a Lipschitz differentiable concave function over a convex set. It thus seems interesting to use a scheme similar to Algorithm 1 on this problem. Unfortunately we will get a convergence rate on the dual variable y and not on the variable of interest:

$$x(y) = \arg\min_{x\in X}\left(\langle Ax - h, y\rangle + \frac{\epsilon}{2}||x - x^0||_2^2\right). \quad (17)$$

Actually, a slight modification of Nesterov's scheme (an ergodic version) can be shown to ensure convergence of x^N with the desired convergence rate. In the following, we detail briefly our main results.

Let x^*_ϵ denote the solution of (13) and y^*_ϵ denote a solution of (15). Let X^* denote the set of minimizers of (2) and let us consider the following algorithm:

Algorithm 2 (Dual)

```
Choose a number of iterations $N$.
Set a point $x^0$ (as close as possible to $X^*$).
Set a starting point $y^0$ (as close as possible to $y^*_\epsilon$).
Set $\epsilon = \epsilon(N) = \frac{|||A|||\cdot||x^0 - x^*_\epsilon||_2}{N}$.
Set $\mathcal{A} = 0$, $g = 0$, $\bar{x} = 0$ and $y = y^0$.
for $k = 0$ to $N$ do
  $a = \frac{1}{L} + \sqrt{\frac{1}{L^2} + \frac{2}{L}\mathcal{A}}$
  $v = \Pi_Y\left(y^0 - g\right)$
  $z = \frac{\mathcal{A}y + av}{\mathcal{A} + a}$
  $y = \Pi_Y\left(z + \frac{\nabla\Psi_\epsilon(z)}{L}\right)$
  $\bar{x} = \bar{x} + ax(y)$   (cf. equation (17))
  $g = g - a\nabla\Psi_\epsilon(y)$
  $\mathcal{A} = \mathcal{A} + a$
end for
Set $\bar{x}^N = \frac{\bar{x}}{\mathcal{A}}$.
```

This algorithm can be shown to have the following properties.

Proposition 3 *$\bar{x}^N$ converges to the projection of x^0 onto the set of minimizers of (2).*

Proposition 4 *The worst case convergence rate is:*

$$\Psi(\bar{x}^N) - \Psi(x^*) \leq \frac{2|||A|||\cdot||x^0 - x^*_\epsilon||_2\sqrt{D}}{N}. \quad (18)$$

Rate (18) is actually very similar to (12). It is thus natural to wonder if there is an interest in using this dual approach. Let us present some interesting aspects of this scheme:

- In the dual approach, the solution of the regularized problem is unique. This guarantees a certain stability of the iterates.
- We can show an additional convergence rate in norm to the regularized solution. Namely, for a fixed ϵ, we have for all k:

$$||\bar{x}^k - x^*_\epsilon||_2^2 \leq \frac{D|||A|||}{\epsilon\cdot k^2} \quad (19)$$

- In practical experiments, model (13) with a small ϵ leads to slightly better SNR than model (2) for some restoration purposes in image processing.
- The practical convergence rates of the dual approach were better than those of the primal approach in all our experiments.

To conclude the theoretical part of this paper, let us precise that problem (3) can be solved with the same algorithms. However, it is preferable not to regularize the term $y \mapsto ||y||_1$ which can be minimized using accelerated soft-thresholding algorithms [1, 10, 12].

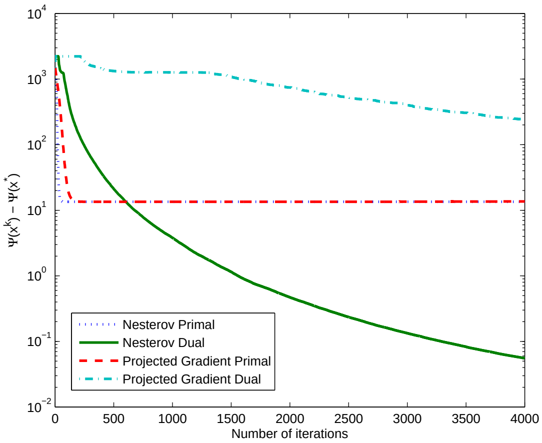
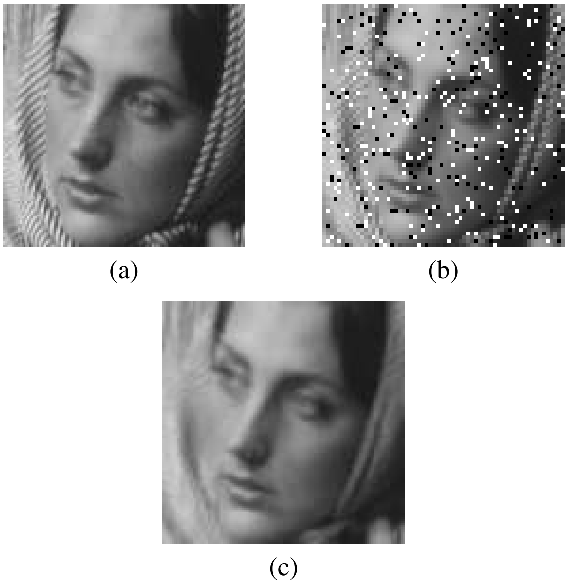


Figure 1: Cost function w.r.t. the number of iterations.



(a)

(b)

(c)

Figure 2: Restoration of a down-sampled and noised image. (a) Original image, (b) down-sampled (by a factor 2) and noised image by 10% of ”Salt & Pepper” noise and finally (c) result of the Nesterov dual approach.

5. Numerical results

In this section we present some comparisons for a problem of image zooming with impulse noise. To solve this problem, we simply set:

- D: convolution by a low-pass filter followed by a down sampling of factor d in the x and y directions.
- $p = 1$ (which is adapted to impulse noise).
- B: a redundant wavelet transform. We set B to be the Dual-Tree Complex Wavelet Tranform (DTCW) [11].

In that case $|||A|||^2$ can be computed explicitly. For the general case, let us point out that iterated power algorithms provide good approximations of $|||A^*A||| = |||A|||^2$.

Figure 1 shows the evolution of the cost function w.r.t. the number of iterations for different techniques. The primal approach is very fast (it converges in 50 iterations) but only decreases the cost function by a factor 100. The dual method decreases the cost function by a factor 10^5 in 4000 iterations, while the projected gradient method only decreases the cost function by a factor 10. We, thus, can see the major improvement of Y. Nesterov's scheme on these problems. Figure 2 shows the result of the model. The DTCW transform slightly blurs the image but allows to retrieve thin details. Unfortunately, we did not have sufficient time to give some comparisons with other approaches. This will be included in the final paper version.

References:

[1] A. Beck and M. Teboulle. Fast iterative shrinkagethresholding algorithm for linear inverse problems. *SIAM J. on Imaging Science*, to appear.

[2] A. Bermudez and C. Moreno. Duality methods for solving variational inequalities. *Comp. and Maths. with Appls.*, 7:43-58, 1981.

[3] R.J. Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *J. Math. Anal. Appl*, 61:159-164, 1977.

[4] J.F. Cai, R. Chan, Z.W. Shen, and L.X. Shen. Convergence analysis of tight framelet approach for missing data recoverys. *Advances in Computational Mathematics*, to appear.

[5] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Inf. Theory*, 2006.

[6] A. Chambolle, R. Devore, N.Y. Lee, and B.J. Lucier. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Processing*, 7:319-335, 1998.

[7] G. Facciolo, A. Almansa, J.-F. Aujol, and Vicent Caselles. Irregular to regular sampling, denoising and deconvolution. *SIAM Journal on Multiscale Modeling and Simulation*, in press.

[8] A. Nemirovski. Information-based complexity of linear operator equations. *Journal of Complexity*, 8:153-175, 1992.

[9] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127-152, 2005.

[10] Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper 2007/76*, 2007.

[11] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury. The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6), Nov. 2005.

[12] P. Weiss. *Algorithmes rapides d'optimisation convexe. Applications à la restauration d'images et à la détection de changements*. PhD thesis, Université de Nice Sophia Antipolis, Dec. 2008.