On the convergence of the IRLS algorithm in Non-Local Patch Regression

Kunal N. Chaudhury*

Abstract

Recently, it was demonstrated in [4], [10] that the robustness of the classical Non-Local Means (NLM) algorithm [1] can be improved by incorporating $\ell^p (0 < p \leq 2)$ regression into the NLM framework. This general optimization framework, called Non-Local Patch Regression (NLPR), contains NLM as a special case. Denoising results on synthetic and natural images show that NLPR consistently performs better than NLM beyond a moderate noise level, and significantly so when p is close to zero. An iteratively reweighted least-squares (IRLS) algorithm was proposed for solving the regression problem in NLPR, where the NLM output was used to initialize the iterations. Based on exhaustive numerical experiments, we observe that the IRLS algorithm is globally convergent (for arbitrary initialization) in the convex regime $1 \leq p \leq 2$, and locally convergent (fails very rarely using NLM initialization) in the non-convex regime $0 < p < 1$. In this letter, we adapt the "majorize-minimize" framework introduced in [8] to explain these observations.

Index Terms

Non-local means, non-local patch regression, ℓ^p minimization, non-convex optimization, iteratively reweighted least-squares, majorize-minimize, stationary point, relaxation sequence, linear convergence.

I. Introduction

In the last decade, some very effective frameworks for image restoration have been proposed that (a) exploit long-distance correlations in natural images, and (b) use patches instead of pixels to robustly compare photometric similarities. This includes the classical Non-Local Means (NLM) algorithm [1], and the more sophisticated BM3D algorithm [2]. The latter combines the NLM framework with other classical algorithms, and is widely considered as the

*K. N. Chaudhury is with the Program in Applied and Computational Mathematics (PACM), Princeton University, Princeton, NJ 08544, USA (mail: kchaudhu@math.princeton.edu).

state-of-the-art in image denoising. We refer the reader to [3] for a comprehensive review of patch-based algorithms.

Let $u = (u_i)$ be some linear indexing of the input noisy image. In NLM, the restored image $\hat{u} = (\hat{u}_i)$ is computed using the simple formula

$$\hat{u}_i = \frac{\sum_{j \in S(i)} w_{ij} u_j}{\sum_{j \in S(i)} w_{ij}}. \tag{1}$$

Here, w_{ij} is some weight (affinity) assigned to pixels i and j, and $S(i)$ is some non-local (sufficiently large) neighborhood of pixel i over which the averaging is performed [1]. In particular, for a given pixel i, let $\mathbf{P}_i$ denote the restriction of u to a square window around i. Letting k be the length of this window, this associates every pixel i with a point $\mathbf{P}_i$ in $\mathbf{R}^{k^2}$ (the patch space). The weights in NLM are set to be $w_{ij} = \exp(-\|\mathbf{P}_i - \mathbf{P}_j\|^2/h^2)$, where $\|\mathbf{P}_i - \mathbf{P}_j\|$ is the Euclidean distance between $\mathbf{P}_i$ and $\mathbf{P}_j$, and h is a smoothing parameter.

Recently, it was demonstrated in [4], [10] that the robustness of the Non-Local Means (NLM) algorithm [1] can be improved by incorporating ℓ^p regression into the NLM framework. The idea was to fix some $0 < p \leq 2$, and consider the following unconstrained optimization on the patch space:

$$\hat{\mathbf{P}}_i = \arg \min_{\mathbf{P}} \sum_{j \in S(i)} w_{ij} \ \|\mathbf{P} - \mathbf{P}_j\|^p, \tag{2}$$

where w_{ij} are the weights used in NLM (one could also use other weights, e.g., see [9]). The denoised pixel $\hat{u}_i$ was then set to be the center pixel in $\hat{\mathbf{P}}_i$. Note that this reduces to the simple formula in (1) when $p = 2$. In this case, the optimization is performed pixelwise. For any other value of p, the optimization in (2) becomes a generic optimization on the patch space – the regression needs to be performed on patches and not just pixels. In particular, when $0 < p \leq 1$, the resulting estimator turns out to be more robust to "outliers" in the patch space (compared to standard NLM), and this leads to significant improvement in the denoising quality. We refer the reader to [10] for an intuitive understanding of the robustness in NLPR, and for denoising results on synthetic and natural images.

Note that we can generally write the optimization problem in (2) as

$$\min_{\boldsymbol{x} \in \mathbf{R}^d} \sum_{j=1}^{n} w_j \|\boldsymbol{x} - \boldsymbol{a}_j\|^p, \tag{3}$$

where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are given points in $\mathbf{R}^d$, and $w_1, \ldots, w_n$ are some positive weights. Motivated by the work on algorithms for ℓ^p minimization in [5], [6], the authors in [10] proposed to optimize (3) using iteratively reweighted least-squares (IRLS). Fixing some small $\varepsilon > 0$, and

initializing $\boldsymbol{x}^{(0)}$ as the NLM output, the update rule was set to be

$$\boldsymbol{x}^{(t+1)} = \frac{\sum_j \mu_j^{(t)} \boldsymbol{a}_j}{\sum_j \mu_j^{(t)}} \qquad (t \geq 0), \tag{4}$$

where

$$\mu_j^{(t)} = w_j(||\boldsymbol{x}^{(t)} - \boldsymbol{a}_j||^2 + \varepsilon)^{p/2-1}. \tag{5}$$

We refer the reader to [10] for the heuristics behind the IRLS update in (4). Extensive numerical experiments show us the that the algorithm is globally convergent (for arbitrary $\boldsymbol{x}^{(0)}$) in the convex regime $1 \leq p \leq 2$, and locally convergent (fails very rarely compared to, say, the gradient or Newton method) in the non-convex regime $0 < p < 1$, provided that $x^{(0)}$ is set as the NLM estimate. The former observation can be explained using the existing literature on the Wieszfeld algorithm [7], [8], which is perhaps less well-known in the signal processing community. In this letter, we adapt the "majorize-minimize" framework introduced in [8] to specifically analyze (4) when $0 < p < 1$. The analysis automatically covers the case $1 \leq p \leq 2$. This is the content of Section II. In particular, we will show that the algorithm forces the cost to be non-increasing as the iteration progresses, and that it exhibits linear convergence both in the convex and non-convex (when convergence does occur) regime. We will also discuss how these results relate to the experimental findings. We note that in [6], the authors have done an analysis of a related (but more complex) IRLS algorithm in the non-convex regime, but the analysis is quite involved. The present analysis is much more simple, and is based on elementary results from smooth optimization.

II. Convergence Analysis

The key question is what is the cost associated with the IRLS iterations in (4)? This must be resolved even before we ask questions about convergence. It turns out that the iterations corresponds to a regularized version of the original cost (3). This is given by

$$\Phi_\varepsilon(x) = \sum_{j=1}^{n} w_j |\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^p, \tag{6}$$

where $|\boldsymbol{x}|_\varepsilon$ is the regularized version of the Euclidean norm,

$$|\boldsymbol{x}|_\varepsilon^2 = \|\boldsymbol{x}\|^2 + \varepsilon \qquad (\varepsilon > 0).$$

Note that Φ_ε corresponds to the original cost (3) when $\varepsilon = 0$. It can be argued that the minimizers of Φ_ε converge to the minimizer of the original problem as ε tends to zero. In other words, for sufficiently small ε, the minimizer of Φ_ε is close to that of the original problem. Henceforth, to simplify notation, we fix some small $\varepsilon > 0$ and denote Φ_ε by Φ. We note that

in [10], the authors proposed to start with, say, $\varepsilon = 1$, and then gradually shrink it to zero as the iteration progresses. While this does tend to speed up the convergence, the associated analysis becomes quite complicated.

The advantage we get by considering the regularized problem is that the function $\Phi(x)$ is smooth (infinitely differentiable). This allows us to use the powerful tools of smooth optimization. Moreover, Φ inherits the convex nature of the original problem, namely, that it is strictly convex for $1 \leq p \leq 2$. Since Φ is smooth, it suffices to show that its Hessian is positive definite. In fact, the gradient (denoted by Φ') and the Hessian (denoted by Φ'') are given by

$$\Phi'(\boldsymbol{x}) = p \sum_j w_j |\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^{p-2} (\boldsymbol{x} - \boldsymbol{a}_j),$$

and

$$\Phi''(\boldsymbol{x}) = p \sum_j w_j |\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^{p-4} \Big[|\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^2 \mathbf{I}_d - (2-p)(\boldsymbol{x} - \boldsymbol{a}_j)(\boldsymbol{x} - \boldsymbol{a}_j)^T \Big].$$

Here, $\mathbf{I}_d$ is the identity matrix of size $d \times d$. For any non-zero $\boldsymbol{u} \in \mathbf{R}^d$,

$$\boldsymbol{u}^T \Phi''(x) \boldsymbol{u} = p \sum_j w_j |\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^{p-4} \Big[|\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^2 \|\boldsymbol{u}\|^2 - (2-p)(\boldsymbol{u}^T(\boldsymbol{x} - \boldsymbol{a}_j))^2 \Big].$$

Since $\boldsymbol{u}^T(\boldsymbol{x} - \boldsymbol{a}_j) \leq \|\boldsymbol{u}\| \cdot \|\boldsymbol{x} - \boldsymbol{a}_j\| < \|\boldsymbol{u}\| \cdot |\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon$, the quadratic form is strictly larger than

$$p(p-1)\|\boldsymbol{u}\|^2 \sum_j w_j |\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^{p-2}.$$

This is non-negative if and only if $1 \leq p \leq 2$. Therefore, Φ is strictly convex in this case, and has a unique global minimizer $\boldsymbol{x}^\star$ for which $\Phi'(\boldsymbol{x}^\star) = 0$. On the other hand, it is not difficult to see that Φ need not be convex when $p < 1$. The best we can hope for in this case is that the iterates in (4) converge to some local stationary point. In fact, we can show that

Theorem 1: The IRLS update in (4) guarantees the following:

1) For $0 \leq p \leq 2$, the sequence $(\Phi(\boldsymbol{x}^{(t)}))$ is strictly monotonic, $\Phi(\boldsymbol{x}^{(t+1)}) < \Phi(\boldsymbol{x}^{(t)})$ for all t.
2) When $1 \leq p \leq 2$, $(\boldsymbol{x}^{(t)})$ converges linearly to the unique global minimizer of Φ.
3) For $0 < p < 1$, under some additional assumptions, the convergence is again linear and the limit of $(\boldsymbol{x}^{(t)})$ is a stationary point of Φ.

By linear convergence, we mean that the convergence happens at an exponential rate. The relaxation property is particularly important for the non-convex setting, providing the guarantee that the cost at the end of the iterations is less than that obtained from the initial estimate $\boldsymbol{x}^{(0)}$. In optimization literature, one calls $(\boldsymbol{x}^{(t)})$ a *relaxation sequence* if $\Phi(\boldsymbol{x}^{(t+1)}) \leq \Phi(\boldsymbol{x}^{(t)})$. Since Φ is trivially bounded below, this implies convergence of the sequence $(\Phi(\boldsymbol{x}^{(t)}))$. As we will see, the iterates in (4) *unconditionally* generate a relaxation sequence in both the convex

and non-convex regime. This property turns out to be a central ingredient in the guarantees provided by Theorem 1. We note that the relaxation property was recently observed in [9] for the special case $p = 1$. However, we remark that the fact that the convergence of $(\Phi(\boldsymbol{x}^{(t)}))$ is not sufficient to guarantee convergence of $(\boldsymbol{x}^{(t)})$, as claimed in [9]. For example, it is possible that $(\boldsymbol{x}^{(t)})$ keeps oscillating, or escapes to infinity, while ensuring that $\Phi(\boldsymbol{x}^{(t+1)}) \leq \Phi(\boldsymbol{x}^{(t)})$. One of the cases can however be ruled out immediately:

Proposition 2: For $0 < p \leq 2$, the IRLS iterates $(\boldsymbol{x}^{(t)})$ are bounded (does not escape to infinity).

This is a simple consequence of the observation that $\Phi(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$. Indeed, if $(\boldsymbol{x}^{(t)})$ does escape to infinity, then we would have a contradiction since we just showed that $(\Phi(\boldsymbol{x}^{(t)}))$ is bounded.

In the convex regime, we will show that oscillations can also be ruled out. To do so in the non-convex regime, we will need additional assumptions.

A. Majorize-Minimize interpretation

To establish Theorem 1, we will use the majorize-minimize (MM) framework from [8]. In this framework, the key idea is to globally approximate Φ from above using a sequence of quadratic functions. More precisely, after having found $\boldsymbol{x}^{(t)}$, we construct a function $\Psi_t(\boldsymbol{x}) = \Psi(\boldsymbol{x}; \boldsymbol{x}^{(t)})$ such that

- $\Phi(\boldsymbol{x}) \leq \Psi_t(\boldsymbol{x})$ for all x.
- $\Phi(\boldsymbol{x}^{(t)}) = \Psi_t(\boldsymbol{x}^{(t)})$.
- $\Psi_t(\boldsymbol{x})$ has $\boldsymbol{x}^{(t+1)}$ in (4) as its unique global minimizer.

Once we have Ψ_t with the above properties, we immediately see that

$$\Phi(\boldsymbol{x}^{(t+1)}) \leq \Psi_t(\boldsymbol{x}^{(t+1)}) < \Psi_t(\boldsymbol{x}^{(t)}) = \Phi(\boldsymbol{x}^{(t)}).$$

That is, we are guaranteed that $(\boldsymbol{x}^{(t)})$ is a relaxation sequence. We now need to specify $\Psi_t(\boldsymbol{x})$.

Proposition 3: The following choice will suffice:

$$\Psi_t(\boldsymbol{x}) = \Phi(\boldsymbol{x}^{(t)}) + (\boldsymbol{x} - \boldsymbol{x}^{(t)})^T \Phi'(\boldsymbol{x}^{(t)}) + \frac{p}{2} \sum_j \mu_j^{(t)} \|\boldsymbol{x} - \boldsymbol{x}^{(t)}\|^2, \tag{7}$$

where $\mu_j^{(t)}$ is as defined in (5).

Note that the linear part of $\Psi_t(\boldsymbol{x})$ is simply the linear approximation of Φ at $\boldsymbol{x}^{(t)}$, and the quadratic form is derived from the dominant part of $\Phi''(\boldsymbol{x}^{(t)})$. By construction, $\Psi_t(\boldsymbol{x}^{(t)}) =$

$\Phi(\boldsymbol{x}^{(t)})$. Moreover, it is clear that $\Psi_t(x)$ is strictly convex (for all p), and has a global unique minimizer. Setting $x^{(t+1)}$ to be this minimizer, we have

$$\Psi_t'(\boldsymbol{x}^{(t+1)}) = \Phi'(\boldsymbol{x}^{(t)}) + p\sum_j \mu_j^{(t)}(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}) = 0. \tag{8}$$

Substituting for Φ', we get the update rule in (4).

To complete the proof, we need to show that $\Phi(\boldsymbol{x}) \leq \Psi_t(\boldsymbol{x})$. Note that we can write $\Phi(\boldsymbol{x}) - \Psi_t(\boldsymbol{x})$ as

$$\begin{aligned}\sum_{j=1}^n w_j\Big(|\boldsymbol{x}^{(t)} - \boldsymbol{a}_j|_\varepsilon^p - |\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^p \;&+\; p\sum_j w_j|\boldsymbol{x}^{(t)} - \boldsymbol{a}_j|_\varepsilon^{p-2}(\boldsymbol{x} - \boldsymbol{x}^{(t)})^T(\boldsymbol{x}^{(t)} - \boldsymbol{a}_j)\\ &+\; \frac{p}{2}\sum_j |\boldsymbol{x}^{(t)} - \boldsymbol{a}_j|_\varepsilon^{p-2}\|\boldsymbol{x} - \boldsymbol{x}^{(t)}\|^2\Big).\end{aligned}$$

We substitute the following above:

$$(\boldsymbol{x} - \boldsymbol{x}^{(t)})^T(\boldsymbol{x}^{(t)} - \boldsymbol{a}_j) = (\boldsymbol{x} - \boldsymbol{a}_j)^T(\boldsymbol{x}^{(t)} - \boldsymbol{a}_j) - \|\boldsymbol{x}^{(t)} - \boldsymbol{a}_j\|^2,$$

and

$$\|\boldsymbol{x} - \boldsymbol{x}^{(t)}\|^2 = \|\boldsymbol{x} - \boldsymbol{a}_j\|^2 + \|\boldsymbol{x}^{(t)} - \boldsymbol{a}_j\|^2 - 2(\boldsymbol{x} - \boldsymbol{a}_j)^T(\boldsymbol{x}^{(t)} - \boldsymbol{a}_j).$$

This allows us to simplify the expression to

$$\sum_{j=1}^n w_j\Big(\alpha_j^{p/2} - \beta_j^{p/2} + \frac{p}{2}\alpha_j^{p/2-1}(\beta_j - \alpha_j)\Big).$$

where $\alpha_j = |\boldsymbol{x}^{(t)} - \boldsymbol{a}_j|_\varepsilon^2$ and $\beta_j = |\boldsymbol{x} - \boldsymbol{a}_j|_\varepsilon^2$. It can be verified that each term in the sum is non-negative for any $0 \leq p \leq 2$ (for $p = 0, 1,$ and 2 this is obvious). This shows that $\Phi(x) \leq \Psi_t(x)$, concluding the proof of (7).

B. Global and local convergence

Since the sequence $(\Phi(x^{(t)}))$ is monotonic and bounded below, it is convergent. In particular, $\Phi(x^{(t)}) - \Phi(x^{(t+1)}) \to 0$ as $t \to \infty$. So, what can we say about the sequence $(\boldsymbol{x}^{(t)})$?

Proposition 4: We claim that $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}\| \to 0$ as t gets large.

To do so, we use (8),

$$\Phi'(\boldsymbol{x}^{(t)}) = -p\sum_j \mu_j^{(t)}(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}),$$

and the majorizing property,

$$\Phi(\boldsymbol{x}^{(t+1)}) < \Psi(\boldsymbol{x}^{(t+1)}) = \Phi(\boldsymbol{x}^{(t)}) + (\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)})^T\Phi'(\boldsymbol{x}^{(t)}) + \frac{p}{2}\sum_j \mu_j^{(t)}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|^2.$$

Combining these, we see that

$$\Phi(\boldsymbol{x}^{(t+1)}) < \Phi(\boldsymbol{x}^{(t)}) - \frac{p}{2}\sum_j \mu_j^{(t)} \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|^2.$$

Now, from (5) we have the trivial bound $\mu_j^{(t)} > w_j \varepsilon^{p/2-1}$. We can then write

$$\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|^2 \leq \gamma\big[\Phi(\boldsymbol{x}^{(t)}) - \Phi(\boldsymbol{x}^{(t+1)})\big],$$

where

$$\gamma = \frac{2\varepsilon^{1-p/2}}{p(\sum_j w_j)}.$$

Since $(\Phi(x^{(t)}))$ is convergent, we arrive at our claim.

Note that cannot directly conclude from Proposition 4 that $(\boldsymbol{x}^{(t)})$ is convergent. However, since $(\boldsymbol{x}^{(t)})$ is bounded, it has convergent subsequences (by compactness). In the convex regime, we can say something more:

Proposition 5: For $1 \leq p \leq 2$, every convergent subsequence has the same limit, and this limit is the unique stationary point of Φ. In particular, it is necessary that $(\boldsymbol{x}^{(t)})$ converges to $\boldsymbol{x}^\star$

We now establish the above claim. Let $(\boldsymbol{x}^{(m)})$ be a subsequences that converges to $\boldsymbol{x}^\star$. We know that $\|\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m+1)}\| \to 0$, so that the limit of both $(\boldsymbol{x}^{(m)})$ and $(\boldsymbol{x}^{(m+1)})$ must be $\boldsymbol{x}^\star$. Note that

$$0 = \Psi'(\boldsymbol{x}^{(m+1)}) = \Phi'(\boldsymbol{x}^{(m)}) + p\sum_j \mu_j^{(m)}(\boldsymbol{x}^{(m+1)} - \boldsymbol{x}^{(m)}).$$

Since both $\Phi'(\boldsymbol{x})$ and $\mu_j = \mu_j(\boldsymbol{x})$ are smooth, letting $m \to \infty$, we have

$$0 = \Phi'(x^\star) + p\sum_j w_j |\boldsymbol{x}^\star - \boldsymbol{a}_j|^{p-2}(\boldsymbol{x}^\star - \boldsymbol{x}^\star) = \Phi'(\boldsymbol{x}^\star).$$

That is, $\boldsymbol{x}^\star$ is a stationary point of Φ. In the convex regime $1 \leq p \leq 2$, we know that Φ has a unique stationary point $\boldsymbol{x}^\star$. Therefore, every convergent subsequence of $(\boldsymbol{x}^{(t)})$ must have the same limit $\boldsymbol{x}^\star$.

Proposition 6: For $0 < p < 1$, the above claim is true only under certain local assumptions.

The problem in this case is that there can be multiple stationary points of Φ, and the previous argument breaks down (as is typical with non-convex problems). All we know in this case is that $(\boldsymbol{x}^{(t)})$ is bounded, and that the entire $(\boldsymbol{x}^{(t)})$ can be restricted to a ball $B_r(\boldsymbol{x}^\star)$ of radius r around $\boldsymbol{x}^\star$. Suppose we assume that that the initialization $\boldsymbol{x}^{(0)}$ is "good", in that it is situated sufficiently close to a local (probably global) minimizer $\boldsymbol{x}^\star$. It is then possible that r is small enough and $B_r(\boldsymbol{x}^\star)$ contains no other stationary points of Φ. In this case, we are guaranteed that every convergent subsequence, and hence the whole sequence $(\boldsymbol{x}^{(t)})$, converges to $\boldsymbol{x}^\star$.

C. Convergence rate

Plots of $\log(\Phi(\boldsymbol{x}^{(t+1)}) - \Phi(\boldsymbol{x}^\star))$ versus t for (4) suggests a linear trend both for the convex and non-convex cases (assuming convergence for the latter). For the convex regime, we can indeed guarantee that

Proposition 7: For $1 \leq p \leq 2$,

$$\Phi(\boldsymbol{x}^{(t+1)}) - \Phi(\boldsymbol{x}^\star) < \nu\big(\Phi(\boldsymbol{x}^{(t)}) - \Phi(\boldsymbol{x}^\star)\big) \qquad (\text{large } t,\ \nu < 1).$$

In other words, the convergence is exponential, where the convergence rate is controlled by ν. To establish our claim, we being by comparing Φ at the points $\boldsymbol{x}^{(t+1)}$ and the linear combination $\theta_t \boldsymbol{x}^{(t)} + (1-\theta_t)\boldsymbol{x}^\star$ (we will define θ_t later),

$$\Phi(\boldsymbol{x}^{(t+1)}) < \Psi_t(\boldsymbol{x}^{(t+1)}) \leq \Psi_t(\theta_t \boldsymbol{x}^{(t)} + (1-\theta_t)\boldsymbol{x}^\star).$$

The first inequality follows from majorization, and the second from the optimality of $\boldsymbol{x}^{(t+1)}$. From (7), we can write $\Psi_t(\theta_t \boldsymbol{x}^{(t)} + (1-\theta_t)\boldsymbol{x}^\star)$ as

$$\Phi(\boldsymbol{x}^{(t)}) + (1-\theta_t)(\boldsymbol{x}^\star - \boldsymbol{x}^{(t)})^T \Phi'(\boldsymbol{x}^{(t)}) + \frac{p(1-\theta_t)^2}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}^{(t)}\|^2 \sum_j \mu_j^{(t)}.$$

On the other hand,

$$\Psi_t(\boldsymbol{x}^\star) = \Phi(\boldsymbol{x}^{(t)}) + (\boldsymbol{x}^\star - \boldsymbol{x}^{(t)})^T \Phi'(\boldsymbol{x}^{(t)}) + \frac{p}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}^{(t)}\|^2 \sum_j \mu_j^{(t)}. \tag{9}$$

Using this to eliminate the term containing $\Phi'(\boldsymbol{x}^{(t)})$, we can write $\Psi_t(\theta_t \boldsymbol{x}^{(t)} + (1-\theta_t)\boldsymbol{x}^\star)$ as

$$\theta_t \Phi(\boldsymbol{x}^{(t)}) + (1-\theta_t)\Big(\Psi(\boldsymbol{x}^\star) - \frac{p\theta_t}{2}\sum_j \mu_j^{(t)}\|\boldsymbol{x}^\star - \boldsymbol{x}^{(t)}\|^2\Big).$$

Now, set

$$\theta_t = \frac{2(\Psi(\boldsymbol{x}^\star) - \Phi(\boldsymbol{x}^\star))}{p\|\boldsymbol{x}^\star - \boldsymbol{x}^{(t)}\|^2 \sum_j \mu_j^{(t)}}. \tag{10}$$

Then $\Phi(\boldsymbol{x}^{(t+1)}) < \Psi_t(\boldsymbol{x}^{(t+1)}) = \theta_t \Phi(\boldsymbol{x}^{(t)}) + (1-\theta_t)\Psi(\boldsymbol{x}^\star)$, and hence

$$\Phi(\boldsymbol{x}^{(t+1)}) - \Phi(\boldsymbol{x}^\star) < \theta_t(\Phi(\boldsymbol{x}^{(t)}) - \Phi(\boldsymbol{x}^\star)).$$

By construction, $\theta_t \geq 0$ for all t. We are done if we can show that $\theta_t \leq \nu < 1$ for some constant ν. By Taylor's theorem,

$$\Phi(\boldsymbol{x}^\star) = \Phi(\boldsymbol{x}^{(t)}) + (\boldsymbol{x}^\star - \boldsymbol{x}^{(t)})^T \Phi'(\boldsymbol{x}^{(t)}) + \frac{1}{2}(\boldsymbol{x}^\star - \boldsymbol{x}^{(t)})^T \Phi''(\boldsymbol{y}^{(t)})(\boldsymbol{x}^\star - \boldsymbol{x}^{(t)}), \tag{11}$$

where $\boldsymbol{y}^{(t)}$ is some point on the segment joining $\boldsymbol{x}^\star$ and $\boldsymbol{x}^{(t)}$. Plugging (9) and (11) in (10),

$$\theta_t = 1 - \frac{(\boldsymbol{x}^\star - \boldsymbol{x}^{(t)})^T \Phi''(\boldsymbol{y}^{(t)})(\boldsymbol{x}^\star - \boldsymbol{x}^{(t)})}{p\|\boldsymbol{x}^\star - \boldsymbol{x}^{(t)}\|^2 \sum_j \mu_j^{(t)}} \leq 1 - \Big(p\sum_j \mu_j^{(t)}\Big)^{-1} \lambda_{\min}(\Phi''(\boldsymbol{y}^{(t)})),$$

where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of the matrix A. Now, since Φ'' is continuous, $\lambda_{\min}(\Phi''(\boldsymbol{y}^{(t)}))$ approaches $\lambda_{\min}(\Phi''(\boldsymbol{x}^\star))$ as $t \to \infty$. Moreover, $\boldsymbol{x}^\star$ is a (local) minimizer. Hence, $\Phi''(\boldsymbol{x}^\star)$ is necessarily positive semidefinite, that is, $\lambda_{\min}(\Phi''(\boldsymbol{x}^\star)) \geq 0$. This is true for $0 < p \leq 2$. Therefore, for all sufficiently large t, $0 \leq \theta_t \leq \nu \leq 1$, where

$$\nu = 1 - \Big(p \sum_j \mu_j^{(\infty)}\Big)^{-1} \lambda_{\min}(\Phi''(\boldsymbol{x}^\star)),$$

and where $\mu_j^{(\infty)} = w_j(||\boldsymbol{x}^\star - \boldsymbol{a}_j||^2 + \varepsilon)^{p/2-1} > 0$.

For the convex regime $1 \leq p \leq 2$, $\Phi''(\boldsymbol{x}^\star)$ is guaranteed to be positive definite, so that $\nu < 1$. This completes the proof of Proposition 7.

For the non-convex setting, the above argument holds under additional assumptions:

Proposition 8: For $0 < p < 1$, assume that $(\boldsymbol{x}^{(t)})$ converges to the local minimizer $\boldsymbol{x}^\star$, and that $\Phi''(\boldsymbol{x}^\star)$ is positive definite. Then $(\Phi(\boldsymbol{x}^{(t)}))$ converges linearly to $\Phi(\boldsymbol{x}^\star)$.

III. Discussion

The linear convergence of IRLS is typical of first-order methods. We have also tried second-order Newton methods for optimizing (6), which are guaranteed to exhibit quadratic convergence (locally). Indeed, Newton methods typically require less than half the number of iterations needed by the updates in (4) to reach a given accuracy. However, the cost of a single Newton step (computation of Φ'' and its inversion) is significantly more than that of the simple update in (4). As a result, the total execution time of IRLS turns out to be smaller than that of Newton methods. The other point that we noticed from the numerical simulations is that IRLS is much more stable than Newton (or gradient descent) methods in the non-convex regime. In particular, the iterates in the Newton method often diverge to infinity if the initialization is not "close" to a local minimum. However, the IRLS iterates never escape to infinity (this is clear from Theorem 1), and almost always converge to the global optimum when we initialize using the NLM output. In rare cases when (4) gets "stuck" in a local minimum (say, due to bad initialization), Newton methods are found to have the same problem. It would be interesting to see if one could give a more accurate analysis of the IRLS algorithm in the non-convex regime.

IV. Acknowledgments

The author would like to thank Prof. Gilad Lerman and Prof. Amit Singer for interesting discussions.

REFERENCES

[1] A. Buades, B. Coll, J. M. Morel, "A review of image denoising algorithms, with a new one," Multiscale Modeling and Simulation, vol. 4, pp. 490-530, 2005.

[2] K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," IEEE Transactions on Image Processing, vol. 16, pp. 2080-2095, 2007.

[3] P. Milanfar, "A tour of modern image filtering", IEEE Signal Processing Magazine, vol. 30, no. 1, 2013.

[4] K. N. Chaudhury, A. Singer, "Non-local Euclidean medians," IEEE Signal Processing Letters, vol. 19, no. 11, pp. 745 - 748, 2012.

[5] R. Chartrand, "Iteratively reweighted algorithms for compressive sensing," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3869-3872, 2008.

[6] I. Daubechies, R. Devore, M. Fornasier, C. S. Gunturk "Iteratively reweighted least squares minimization for sparse recovery," Communications on Pure and Applied Mathematics, vol. 63, pp. 1-38, 2009.

[7] E. Weiszfeld, "Sur le point par lequel le somme des distances de n points donnes est minimum," Tohoku Mathematical Journal, vol. 43, pp. 355-386, 1937.

[8] H. Voss, U. Eckhardt, "Linear convergence of generalized Weiszfeld's method," Computing, vol. 25(3), pp. 243-251, 1980.

[9] Z. Sun, S. Chen, "Analysis of non-local Euclidean medians and its improvement", IEEE Signal Processing Letters, vol. 20, no. 4, pp. 303-306, 2013.

[10] K. N. Chaudhury, A. Singer, "Non-local patch regression: Robust image denoising in patch space," accepted to IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.