Fast Iterative Shrinkage for Signal Declipping and Dequantization

Lucas Rencker[1], Francis Bach[2], Wenwu Wang[1], and Mark D. Plumbley[1].

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

[2]SIERRA-project team, INRIA, Paris, France

Abstract— **We address the problem of recovering a sparse signal from clipped or quantized measurements. We show how these two problems can be formulated as minimizing the distance to a convex feasibility set, which provides a convex and differentiable cost function. We then propose a fast iterative shrinkage/thresholding algorithm that minimizes the proposed cost, which provides a fast and efficient algorithm to recover sparse signals from clipped and quantized measurements.**

1 Introduction

Clipping and quantization are common distortions in digital signal processing. In this paper we address the problem of recovering a signal from clipped or quantized measurements. We consider a distorted signal $\mathbf{y} = f(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^N$ is the original clean signal and $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ is a nonlinear and non-invertible clipping or quantization function. We further assume that $\mathbf{x}$ is sparse with respect to a known *overcomplete* dictionary $\mathbf{D} \in \mathbb{R}^{N \times M}$ ($N < M$), i.e., $\mathbf{x} = \mathbf{D}\,\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}$ sparse.

Recovering a sparse signal from clipped or quantized measurements is often formulated as the following constrained sparse coding problem [1, 2]:

$$\min_{\boldsymbol{\alpha}} \Psi(\boldsymbol{\alpha}) \quad \text{s.t.} \quad \mathbf{D}\,\boldsymbol{\alpha} \in f^{-1}(\mathbf{y}) \tag{1}$$

where $\Psi(\cdot)$ is a sparsity-inducing norm or pseudo-norm, and $f^{-1}(\mathbf{y})$ is the *pre-image* of the observed signal $\mathbf{y}$ through f. The set $f^{-1}(\mathbf{y})$ can be seen as the *feasibility set* associated with the measurement $\mathbf{y}$, i.e., the set of possible input signals that could have generated $\mathbf{y}$. In the case of declipping, the feasibility set can be explicitly formulated as $f^{-1}(\mathbf{y}) = \{\mathbf{x} \mid \mathbf{M}^{\mathrm{r}}\mathbf{x} = \mathbf{M}^{\mathrm{r}}\mathbf{y}, \mathbf{M}^{\mathrm{u}}\mathbf{x} \succeq \theta^{\mathrm{u}}\mathbf{1}, \mathbf{M}^{\mathrm{l}}\mathbf{x} \preceq \theta^{\mathrm{l}}\mathbf{1}\}$ [1] where $\mathbf{M}^{\mathrm{r}}$, $\mathbf{M}^{\mathrm{u}}$ and $\mathbf{M}^{\mathrm{l}}$ are diagonal binary matrices indicating the reliable, upper and lower clipped samples respectively, and $\theta^{\mathrm{u}} > \theta^{\mathrm{l}}$ are upper and lower clipping thresholds respectively. In the case of quantization, $f^{-1}(\mathbf{y}) = [l_1, u_1] \times ... \times [l_N, u_N]$, where $[l_i, u_i]$ is the quantization region associated with each sample y_i.

Eqn. (1) is a constrained, non-smooth and possibly non-convex optimization problem which can be difficult to solve. Recently, algorithms based on Alternating Direction Method of Multipliers (ADMM) [3] or the related Douglas-Rachford algorithm [4] have been proposed to solve (1), see [1] for declipping or the implementation of [2] in [5] for dequantization. However, these algorithms involve computing proximal operators of the type

$$\operatorname*{argmin}_{\boldsymbol{\alpha}} \|\mathbf{u} - \boldsymbol{\alpha}\|_2^2 + \mathbb{1}_{f^{-1}(\mathbf{y})}(\mathbf{D}\,\boldsymbol{\alpha}) \tag{2}$$

at each iteration for some $\mathbf{u} \in \mathbb{R}^M$ (see e.g. [1]), where $\mathbb{1}_{f^{-1}(\mathbf{y})}(\cdot)$ is the indicator function of the set $f^{-1}(\mathbf{y})$. When the dictionary $\mathbf{D}$ is orthogonal or a tight frame ($\mathbf{D}^T\mathbf{D} = \mathbf{I}$), (2) can be computed efficiently in closed form. However for general overcomplete dictionaries, (2) is a non-orthogonal projection which has to be computed iteratively, using (e.g.) another nested ADMM algorithm at each iteration. This leads to a heavy computational cost, which can be prohibitive for large-scale applications.

2 Proposed problem formulation

We propose to relax the constrained problem (1) as the following unconstrained problem (already proposed in [6] in the context of declipping):

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \mathrm{d}(\mathbf{D}\,\boldsymbol{\alpha}, f^{-1}(\mathbf{y}))^2 + \lambda\Psi(\boldsymbol{\alpha}) \tag{3}$$

where $\mathrm{d}(\mathbf{x}, \mathcal{C})^2$ is the squared Euclidean distance between $\mathbf{x}$ and the set $\mathcal{C}$, defined as:

$$\mathrm{d}(\mathbf{x}, \mathcal{C})^2 = \min_{\mathbf{z} \in \mathcal{C}} \|\mathbf{x} - \mathbf{z}\|_2^2. \tag{4}$$

The proposed formulation thus enforces the estimated signal $\mathbf{D}\,\boldsymbol{\alpha}$ to be close to its feasibility set $f^{-1}(\mathbf{y})$, where the parameter λ controls a trade-off between data fidelity and sparsity. Note that when $\lambda \to 0^+$, (3) is equivalent to the constrained problem (1). However since $f^{-1}(\mathbf{y})$ is convex, the proposed problem formulation provides convenient properties which we recall here:

- The data-fidelity term $\mathrm{d}(\mathbf{x}, f^{-1}(\mathbf{y}))^2$ is convex, as a minimum of a family of convex functions $\|\cdot\|_2^2$ over a non-empty and convex set [7, Section 3.2.5]

- $\mathrm{d}(\cdot, f^{-1}(\mathbf{y}))^2$ is differentiable with gradient [8, eqn. (1.1)]:

$$\nabla_{\mathbf{x}} \frac{1}{2} \mathrm{d}(\mathbf{x}, f^{-1}(\mathbf{y}))^2 = \mathbf{x} - \Pi_{f^{-1}(\mathbf{y})}(\mathbf{x}), \tag{5}$$

  where $\Pi_{f^{-1}(\mathbf{y})}(\mathbf{x})$ is the orthogonal projection of $\mathbf{x}$ onto $f^{-1}(\mathbf{y})$.

- The gradient (5) is 1-Lipschitz. This stems from the contraction property of projection operators onto convex sets, see e.g. [9, Prop. B.11].

The proposed formulation is thus a problem of minimizing a convex and differentiable data-fidelity term, along with a sparsity-inducing regularizer, which is similar to classical sparse recovery methods such as Basis Pursuit Denoising (BPDN) [10]. Moreover when the feasibility set is a singleton $f^{-1}(\mathbf{y}) = \{\mathbf{x}\}$ (i.e., the signal is unclipped/unquantized), then (3) simplifies to a classical sparse recovery problem such as BPDN:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{D}\,\boldsymbol{\alpha} - \mathbf{x}\|_2^2 + \lambda\Psi(\boldsymbol{\alpha}). \tag{6}$$

3 Proposed algorithm

In the rest of this paper we focus on the convex ℓ_1 case, i.e. solving (3) with $\Psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$. In this case (3) becomes a problem of minimizing the sum of a convex and smooth cost data-fidelity term, along with a convex and non-smooth regularizer. This can be classically solved using Iterative Shrinkage/Thresholding Algorithms (ISTA) [11]. ISTA is an attractive class of algorithm since they only involve gradient computations and simple element-wise thresholding, making them simple and adequate for large-scale problems. More precisely, ISTA applied to (3) iterates (after an initial guess $\boldsymbol{\alpha}_0$):

$$\boldsymbol{\alpha}_{k+1} = S_{\mu\lambda}\big(\boldsymbol{\alpha}_k - \mu\mathbf{D}^T(\mathbf{D}\,\boldsymbol{\alpha}_k - \Pi_{f^{-1}(\mathbf{y})}(\mathbf{D}\,\boldsymbol{\alpha}_k))\big) \quad (7)$$

where $S_\rho(.)$ is the soft-thresholding operator:

$$S_\rho(\boldsymbol{\alpha})_i = \max(|\alpha_i| - \rho, 0)\,\mathrm{sign}(\alpha_i), \quad (8)$$

and μ is a step size that can be typically set as $1/L$, where $L = \|\mathbf{D}^T\mathbf{D}\|_2$ is the Lipschitz constant of the gradient [11]. Note that $\Pi_{f^{-1}(\mathbf{y})}(\cdot)$ here is a simple orthogonal projection that can be computed using element-wise maxima. The ISTA type algorithm (7) thus provides a simple and efficient way to solve the relaxed problem (3), which is computationally simpler than ADMM based algorithms to solve the constrained problem (1). However, for badly conditioned matrices $\mathbf{D}^T\mathbf{D}$, ISTA is also known to converge quite slowly. Several algorithms have been proposed to speed up ISTA, such as the celebrated *Fast Iterative/Shrinkage Thresholding* algorithm (FISTA) [11]. FISTA applied to the proposed problem is presented in Algorithm 1.

Algorithm 1 FISTA for declipping/dequantization

Require: $f^{-1}(\mathbf{y}), \mathbf{D}, \lambda, \boldsymbol{\alpha}_0$
  Initialize: $\mathbf{u}_1 = \boldsymbol{\alpha}_0, t_1 = 1, k = 1$
  Iterate until convergence:

$$\boldsymbol{\alpha}_k = S_{\mu\lambda}\big(\mathbf{u}_k - \mu\mathbf{D}^T(\mathbf{D}\mathbf{u}_k - \Pi_{f^{-1}(\mathbf{y})}(\mathbf{D}\mathbf{u}_k))\big) \quad (9)$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \quad (10)$$

$$\mathbf{u}_{k+1} = \boldsymbol{\alpha}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k-1}) \quad (11)$$

$$k = k + 1 \quad (12)$$

  return $\hat{\boldsymbol{\alpha}}$

FISTA thus simply computes the thresholded gradient descent step on a linear combination of the two previous estimates $\boldsymbol{\alpha}_k$ and $\boldsymbol{\alpha}_{k-1}$. Note that the cost of computing (10) and (11) is negligible compared to that of (9), so FISTA does not incur extra computational cost per iteration compared to ISTA. However, the convergence rate of FISTA can be shown to be of $\mathcal{O}(1/k^2)$, instead of $\mathcal{O}(1/k)$ for ISTA [11].

4 Numerical results

We generate a random dictionary $\mathbf{D} \in \mathbb{R}^{256\times512}$ with Gaussian i.i.d entries and 100 16-sparse vectors $\boldsymbol{\alpha} \in \mathbb{R}^{512}$, and normalize the resulting signals $\mathbf{x} = \mathbf{D}\,\boldsymbol{\alpha}$ to unit ℓ_∞ norm. We then clip or quantize each vector as $\mathbf{y} = f(\mathbf{x})$, using clipping at different clipping level θ, and a uniform midriser quantizer with bin width $\Delta = 2^{1-N_b}$, where N_b is the number of bits. We compare the constrained formulation (1) solved using ADMM (see

e.g. [1]), and the proposed relaxed approach (3) solved using (F)ISTA. All algorithms are computed using the ℓ_1-norm, and we fix $\lambda = 10^{-2}$ for (F)ISTA. The algorithms are evaluated in terms of average SNR of the reconstructed signals $\hat{\mathbf{x}} = \mathbf{D}\hat{\boldsymbol{\alpha}}$. The results are presented in Figure 1 and 2.
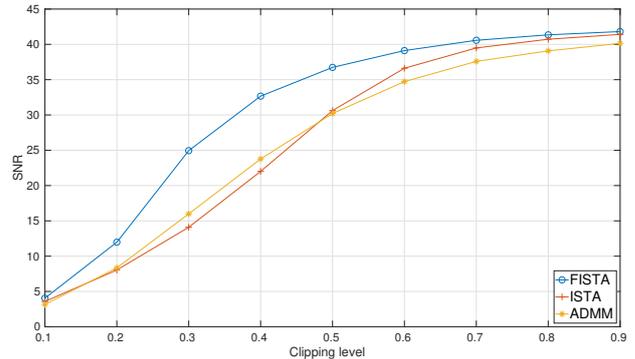

Figure 1: Declipping performance for different clipping levels
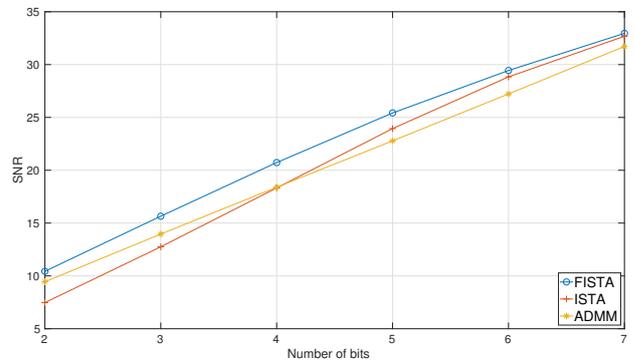

Figure 2: Dequantization performance for different quantization levels

All algorithms are computed with a maximum of 400 iterations, or until the algorithm has converged. Experiments show that the proposed formulation (3) solved with ISTA leads to comparable results to ADMM. However due to its slow convergence rate, ISTA might still be far from the optimum even after 400 iterations. FISTA on the other hand often reaches the optimum in under 150 iterations, and leads to a significant performance increase in terms of signal reconstruction. The average computational time for each algorithm is reported in Table 1. As expected, solving (3) using (F)ISTA is significantly faster than solving (1) using ADMM, since (F)ISTA only involves computing gradients and element-wise computations, while ADMM involves computing non-orthogonal projections at each iteration.

cpu time (s)	ADMM	ISTA	FISTA
declipping	968.4	3.79	**1.56**
dequantization	943.52	3.88	**1.81**

Table 1: Average computational time of each algorithm

5 Conclusion

We showed that relaxing the constrained problem (1) leads to a simple optimization problem, which can be solved using a fast iterative shrinkage algorithm. The proposed algorithm leads to increased performance and a significant reduction in computational time.

References

[1] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: a flexible non-convex approach," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Liberec, Czech Republic, 2015, pp. 243–250.

[2] A. Moshtaghpour, L. Jacques, V. Cambareri, K. Degraux, and C. De Vleeschouwer, "Consistent basis pursuit for signal and matrix estimates in quantized compressed sensing," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 25–29, 2016.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[4] P. L. Combettes and J.-C. Pesquet, "A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 564–574, 2007.

[5] N. Perraudin, V. Kalofolias, D. Shuman, and P. Vandergheynst, "UNLocBoX: A MATLAB convex optimization toolbox for proximal-splitting methods," *arXiv preprint arXiv:1402.0779*, 2014.

[6] L. Rencker, F. Bach, W. Wang, and M. D. Plumbley, "Consistent dictionary learning for signal declipping," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Guildford, UK, 2018, pp. 446–455.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[8] R. B. Holmes, "Smoothness of certain metric projections on Hilbert space," *Transactions of the American Mathematical Society*, vol. 184, pp. 87–100, 1973.

[9] D. P. Bertsekas, *Nonlinear programming*, 2nd ed. Athena Scientific Belmont, 1999.

[10] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[11] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.