On the Tradeoff between Convergence Speed and Reconstruction Accuracy in Inverse Problems

Raja Giryes
School of EE
Tel Aviv University University
Tel Aviv, Israel 69978
raja@tauex.tau.ac.il

Yonina C. Eldar
EE Department
Technion - IIT
Haifa, Israel, 32000
yonina@ee.technion.ac.il

Alex M. Bronstein
CS Department
Technion - IIT
Haifa, Israel, 32000
bron@cs.technion.ac.il

Guillermo Sapiro
ECE Department
Duke University
Durham, NC, 27708
guillermo.sapiro@duke.edu

I. Introduction

Solving inverse problems with iterative algorithms such as the iterative shrinkage and thresholding algorithm (ISTA) [1], [2], [3], [4], [5] is a popular method, especially for large data. In numerous applications, due to time constraints, the number of iterations one may apply is usually limited, consequently limiting the accuracy achievable by traditional methods. Given a recovery error one is willing to tolerate, an important question is whether it is possible to modify the original iterations to obtain a faster convergence within the allowed error.

In this work we provide theoretical foundations elucidating the tradeoff between the allowed minimization error and the computational cost of iterative algorithms for solving inverse problems. In particular, we aim at providing an explanation for the acceleration provided for ISTA by the learned ISTA (LISTA) strategy [6], [7], [8], [9]. In order to do that, we formally show that it is possible to design algorithms with better convergence speed if we allow a certain reconstruction error in the solution. It is interesting to note that such a tradeoff is natural when working with real data, where both the data and the assumed models are already noisy or approximate; searching for the *exact* solution of an optimization problem where all the variables are affected by measurement or model noise is an unnecessary use of valuable computational resources.

II. Accelerated convergence by inaccurate projections

We focus in this work on the projected gradient descent (PGD) algorithm for recovering a signal $\mathbf{x}$ from a set of linear measurements $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$. Let $\mathbf{x} \in \mathcal{K} = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq R\}$, where $R = f(\mathbf{x})$, $\mathcal{P}_K$ is a projection onto $\mathcal{K}$, and f is a penalty function (e.g., $f(\cdot) = \|\cdot\|_1$). Then the PGD estimate at iteration t is given by

$$\mathbf{z}_t = \mathcal{P}_{\mathcal{K}}\left(\mathbf{z}_{t-1} + \mu \mathbf{M}^*(\mathbf{y} - \mathbf{M}\mathbf{z}_{t-1})\right). \quad (1)$$

It has been shown in [10] that PGD convergence depends on the Gaussian mean width of $C_f(\mathbf{x})$, the tangent cone of f at point $\mathbf{x}$.

In many scenarios the signal $\mathbf{x}$ lies in a smaller set $\hat{\mathcal{K}} \subset \mathcal{K}$ defined by a function $\hat{f}$, which has a much smaller Gaussian mean width on $C_{\hat{f}}(\mathbf{x})$ and therefore using PGD with $\mathcal{P}_{\hat{\mathcal{K}}}$ will lead to faster convergence. Yet, we may need to compromise and use PGD with $\mathcal{P}_{\mathcal{K}}$ as calculating $\mathcal{P}_{\hat{\mathcal{K}}}$ may be infeasible or computationally demanding.

In this work we introduce a tradeoff between the reconstruction error and convergence speed by approximating the projection onto $\hat{\mathcal{K}}$ by an inaccurate "projection" composed of a simple projection $p(\cdot)$ (e.g., a linear or an element-wise projection) and the standard projection onto $\mathcal{K}$, $\mathcal{P}_{\mathcal{K}}$, such that (i) p introduces only a slight distortion ϵ into $\mathbf{x}$; and (ii) the projection onto the tangent cone $\hat{C} = C_{\hat{f}}(\mathbf{x})$ of $\hat{f}$ at point $\mathbf{x}$ is well approximated by a projection using $p(\cdot)$ followed by a projection onto the tangent cone $C_f(p(\mathbf{x}))$ of f at point $p(x)$ (also incurring an ϵ error). Plugging this inaccurate projection into the PGD step results in the *inaccurate* PGD (IPGD) iteration (compare to (1))

$$\mathbf{z}_{t+1} = \mathcal{P}_{\mathcal{K}}\left(p\left(\mathbf{z}_t + \mu \mathbf{M}^*(\mathbf{y} - \mathbf{M}\mathbf{z}_t)\right)\right). \quad (2)$$

We proved that the convergence speed of IPGD is smaller than the one of PGD with the projection $\mathcal{P}_{\mathcal{K}}$ but with an additional tolerance error of order ϵ. Therefore, if ϵ is small and of the order of the allowed error (due to noise of model approximation for example), and we have a budget for only a small number of iterations, IPGD will provide a faster convergence than PGD with the original projection $\mathcal{P}_{\mathcal{K}}$.

III. Relation to Sparse Recovery and LISTA

To illustrate our theory, we generate sparse vectors with a tree structure such that the magnitude of the representation entries are smaller in the lower tree levels. We compare PGD with a projection onto $\mathcal{K}$ being the sparse vectors set (IHT) [11], IPGD with the same set $\mathcal{K}$ and p as a projection onto the first layers of the tree, and PGD with a projection onto the "ideal set" $\hat{\mathcal{K}}$ of sparse tree structures (see [12]). Note that the more levels we add in the projection p, the smaller the approximation error ϵ turns out to be. Figure 1 presents the signal reconstruction error $\|\mathbf{x} - \mathbf{z}_t\|_2$ as a function of the number of iterations for PGD and IPGD.

While in the case of the tree sparsity model it is easy to define $\hat{\mathcal{K}}$, we present next several examples for which it is hard to set $\hat{\mathcal{K}}$ accurately. Yet, we may still find a projection p that helps to approximate $\hat{\mathcal{K}}$. In these cases IPGD with this p gets better convergence compared to PGD.

In the case of spectral compressed sensing [13], one wants to recover a sparse representation in a dictionary that has high local coherence. To demonstrate how our technique can be helpful for this scenario, we use a function $p(\cdot)$ that zeros neighboring entries in a given representation with IPGD and compare it to PGD with hard thresholding. Fig. 2 demonstrates that IPGD improves the recovery.

Another possible strategy to improve the reconstruction quality is to use side information on the recovered signal [14], [15]. We demonstrate this approach, in combination with our proposed framework, for the recovery of a sparse vector under the discrete cosine transform (DCT), given information of its representation under the Haar transform. Fig. 3 presents the result.

In many scenarios, we may not know what type of simple projection causes $\mathcal{P}_{\mathcal{K}}(p(\cdot))$ to approximate $\hat{\mathcal{K}}$. In these cases it is possible to learn this projection. Such an approach is taken in the LISTA scheme (see Fig. 4), which succeeds to accelerate the ISTA iteration in this way. Though LISTA uses a proximal mapping and not a projection as in our scheme, we claim that our theory provides an explanation for its accelerated convergence. More details appear in [16].

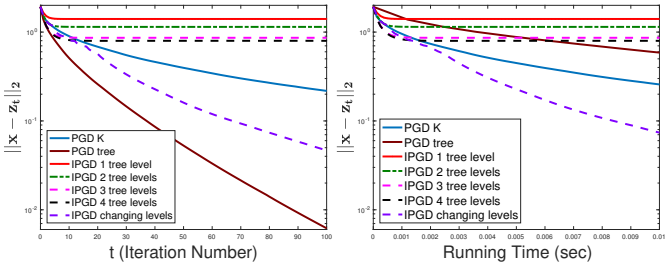

Fig. 1. Reconstruction error as a function of the iterations (left) and running time (right) for sparse recovery in a scenario of a sparse vector with tree structure. We generate the non-zero entries in $\mathbf{x}$ independently from a Gaussian distribution with zero mean and variance $\sigma^2 = 1$ if they are at the first two levels of the tree, $\sigma^2 = 0.1^2$ if they are at the second level, and $\sigma^2 = 0.01^2$ for the rest of the levels. This figure demonstrates the convergence rate of PGD with projections onto the sparse set and sparse tree set compared to IPGD with p that projects onto a certain number of levels of the tree and IPGD with changing p that projects onto an increasing number of levels as the iterations proceed. Notice that while PGD with a projection onto a tree structure converges faster than IPGD as a function of the number of iterations (left figure), it converges slower than IPGD if we take into account the actual run time of each iteration, as shown in the right figure, due to the higher complexity of the PGD projections.

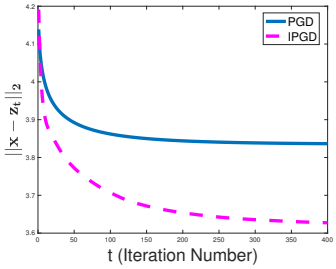

Fig. 2. Reconstruction error as a function of the iterations for sparse recovery in a scenario of a dictionary with high coherence between neighboring atoms. We generate a k-sparse vector, with $k = 20$, of dimension $d = 128$ in a four times redundant DCT dictionary such that the minimal distance between neighboring atoms in this vector is greater than 5. The value in each coefficient is generated from the normal distribution. Then we put random Gaussian values at the neighboring coefficients of each active atom in the representation, with zero mean and variances $\sigma^2 = 0.05^2$ for the neighbors at distance 1 and 2. Note the improved convergence of IPGD compared to PGD.

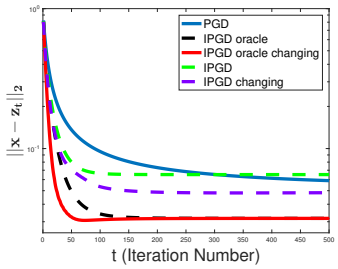

Fig. 3. Reconstruction error as a function of the iterations for sparse recovery with side information. This demonstrates the convergence rate of PGD compared to IPGD. Our sampling matrix is $\mathbf{M} = \mathbf{AD}^*$, where $\mathbf{A} \in \mathbb{R}^{700\times1024}$ is a random matrix with i.i.d. normally distributed entries, and $\mathbf{D}$ is the DCT dictionary. We use random patches of size 32×32, normalized to have unit ℓ_2 norm, from the standard *house* image. Note that these patches are not exactly sparse either in the Haar or the DCT domains. We use IPGD with $p = \mathbf{DPD}^T$, where for IPGD oracle $\mathbf{P}$ projects onto the columns of the Haar matrix that contain 95% of the energy of the signal; for IPGD oracle changing $\mathbf{P}$ projects onto an increasing number of columns from the Haar basis ordered according to their significance in representing the signal; for IPGD $\mathbf{P}$ projects onto the first 512 columns of that Haar basis; and for IPGD changing $\mathbf{P}$ projects onto an increasing number of columns from the Haar basis. Note that IPGD does not converge to zero due to the error ϵ introduced by the projection p.

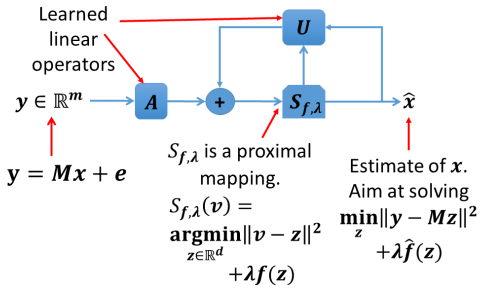

Fig. 4. The LISTA scheme.

REFERENCES

[1] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[2] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization," *Appl. Comput. Harmon. Anal.*, vol. 23, no. 3, pp. 346 – 367, 2007.

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.

[4] P. Machart, S. Anthoine, and L. Baldassarre, "Optimal computational trade-off of inexact proximal methods," in *Multi-Trade-offs in Machine Learning (NIPS workshop)*, 2012.

[5] J.-F. Aujol and C. Dossal, "Stability of over-relaxations for the forward-backward algorithm, application to FISTA," *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2408–2433, 2015.

[6] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th Annual International Conference on Machine Learning (ICML)*, 2010.

[7] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Learning efficient sparse and low rank models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1821–1833, Sept. 2015.

[8] T. Remez, O. Litany, and A. M. Bronstein, "A picture is worth a billion bits: Real-time image reconstruction from dense binary pixels," in *18th International Conference on Computational Photography (ICCP)*, 2015.

[9] J. Tompson, K. Schlachter, P. Sprechmann, and K. Perlin, "Accelerating Eulerian fluid simulation with convolutional networks," *arXiv:1607.03597*, 2016.

[10] S. Oymak, B. Recht, and M. Soltanolkotabi, "Sharp time–data tradeoffs for linear inverse problems," *arXiv:1507.04793*, 2015.

[11] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal*, vol. 27, no. 3, pp. 265 – 274, 2009.

[12] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, April 2010.

[13] M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 35, no. 1, pp. 111 – 129, 2013.

[14] A. O. Hero, R. Piramuthu, J. A. Fessler, and S. R. Titus, "Minimax emission computed tomography using high-resolution anatomical side information and b-spline models," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 920–938, Apr. 1999.

[15] L. Weizman, Y. C. Eldar, and D. B. Bashat, "Compressed sensing for longitudinal MRI: An adaptive-weighted approach," *Medical Physics journal*, vol. 42, no. 9, pp. 5195–5208, Aug. 2015.

[16] R. Giryes, Y. C. Eldar, A. M. Bronstein, and G. Sapiro, "Tradeoffs between convergence speed and reconstruction accuracy in inverse problems," *arXiv:1605.09232*, 2016.