I. INTRODUCTION

Solving inverse problems with iterative algorithms such as the iterative shrinkage and thresholding algorithm (ISTA) \cite{Daubechies2004, Daubechies2009a, Daubechies2009b} 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 by the learned ISTA (LISTA) strategy \cite{Aharon2006, Ahmed2010a, Ahmed2010b, Lee2010}. 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 $x$ from a set of linear measurements $y = Mx + e$. Let $x \in K = \{ z \in \mathbb{R}^d : f(z) \leq R \}$, where $R = f(x)$, $P_K$ is a projection onto $K$, and $f$ is a penalty function (e.g., $f(\cdot) = \|\cdot\|_1$). Then the PGD estimate at iteration $t$ is given by

$$z_t = P_K (z_{t-1} + \mu M^* (y - Mz_{t-1})).$$

(1)

It has been shown in \cite{Candes2008} that PGD convergence depends on the Gaussian mean width of $C_f(x)$, the tangent cone of $f$ at point $x$.

In many scenarios the signal $x$ lies in a smaller set $\hat{K} \subset K$ defined by a function $f$, which has a much smaller Gaussian mean width on $C_f(x)$ and therefore using PGD with $P_{\hat{K}}$ will lead to faster convergence. Yet, we may need to compromise and use PGD with $P_K$ as calculating $P_{\hat{K}}$ 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 $K$ 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 $K$, $P_K$, such that (i) $p$ introduces only a slight distortion $\epsilon$ into $x$; and (ii) the projection onto the tangent cone $\hat{C} = C_f(x)$ of $f$ at point $x$ is well approximated by a projection using $p(\cdot)$ followed by a projection onto the tangent cone $C_f(p(x))$ of $f$ at point $p(x)$ (also incurring an $\epsilon$ error). Plugging this inaccurate projection into the PGD step results in the inaccurate PGD (IPGD) iteration (compare to (1))

$$z_{t+1} = P_K (p(z_t + \mu M^* (y - Mz_t))).$$

(2)

We proved that the convergence speed of IPGD is smaller than the one of PGD with the projection $P_K$ but with an additional tolerance error of order $\epsilon$. Therefore, if $\epsilon$ 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 $P_K$.

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 $K$ being the sparse vectors set (IHT) \cite{Bruckstein2009}, IPGD with the same set $K$ and $p$ as a projection onto the first layers of the tree, and PGD with a projection onto the “ideal set” $\hat{K}$ of sparse tree structures (see \cite{Bruckstein2010}). Note that the more levels we add in the projection $p$, the smaller the approximation error $\epsilon$ turns out to be. Figure 1 presents the signal reconstruction error $\|x - z_t\|_2$ 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 $\hat{K}$, we present next several examples for which it is hard to set $\hat{K}$ accurately. Yet, we may still find a projection $p$ that helps to approximate $\hat{K}$. In these cases IPGD with this $p$ gets better convergence compared to PGD.

In the case of spectral compressed sensing \cite{Donoho2006}, 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 \cite{Tropp2004, Roth2005}. 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 $P_K(p(\cdot))$ to approximate $\hat{K}$. 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 \cite{Giryes2011}.
higher complexity of the PGD projections.

the actual run time of each iteration, as shown in the right figure, due to the iterations (left figure), it converges slower than IPGD if we take into account levels as the iterations proceed. Notice that while PGD with a projection onto the tree and IPGD with changing $p$ set compared to IPGD with convergence rate of PGD with projections onto the sparse set and sparse tree

and

We use IPGD with $p = \text{DPD}^2$, where for IPGD oracle $P$ projects onto the columns of the Haar matrix that contain 95% of the energy of the signal; for IPGD oracle changing $P$ projects onto an increasing number of columns from the Haar basis ordered according to their significance in representing the signal; for IPGD $P$ projects onto the first 512 columns of that Haar basis; and for IPGD changing $P$ projects onto an increasing number of columns from the Haar basis. Note that IPGD does not converge to zero due to the error $\epsilon$ introduced by the projection $p$.


