
Fast Hard Thresholding with Nesterov’s Gradient Method

Volkan Cevher
Idiap Research Institute
Ecole Polytechnique Federale de Lausanne
volkan.cevher@epfl.ch

Sina Jafarpour
Department of Computer Science
Princeton University
sina@cs.princeton.edu

Abstract

We provide an algorithmic framework for structured sparse recovery which unifies combinatorial optimization with the non-smooth convex optimization framework by Nesterov [1, 2]. Our algorithm, dubbed Nesterov iterative hard-thresholding (NIHT), is similar to the algebraic pursuits (ALPS) in [3] in spirit: we use the gradient information in the convex data error objective to navigate over the non-convex set of structured sparse signals. While ALPS feature *a priori* approximation guarantees, we were only able to provide an *online* approximation guarantee for NIHT (e.g., the guarantees require the algorithm execution). Experiments show however that NIHT can empirically outperform ALPS and other state-of-the-art convex optimization-based algorithms in sparse recovery.

1 Introduction

We consider the following *sparse approximation* problem: given a matrix $\Phi \in \mathbb{R}^{M \times N}$ ($M < N$), a vector $u \in \mathbb{R}^M$, and $\epsilon > 0$, find a vector x satisfying the following data error objective

$$f(x) = \|u - \Phi x\|^2 \leq \epsilon,$$

whenever it exists, such that $\|Dx\|_0 = K$ (i.e., x has at most K -nonzero entries), where Dx is a sparse representation. The locations of the non-zero entries may be structured, exhibiting clustered or dispersive patterns on graphs. This particular problem permeates many learning and signal processing applications; examples include learning sparse subsets of features in classification [4], learning graphical models in statistical inference [5], and compressive sensing [6, 7].

2 Preliminaries

By sparse representations, we mean one of the following cases depending on the context. $x \in \mathbb{R}^N$ has a *synthesis*-sparse representation as $x = E\alpha$ in $E \in \mathbb{R}^{N \times N'}$ ($N' \leq N$), when $K \ll N'$ coefficients of α can well-approximate the signal x . $x \in \mathbb{R}^N$ has an *analysis*-sparse representation as $\alpha = Dx$ in $D \in \mathbb{R}^{N' \times N}$ ($N' \geq N$), when $K \ll N'$ coefficients of α can well-approximate the signal as $x = E\alpha_{|K}$, where E is the left inverse of D . Example representations include wavelets for synthesis, and overcomplete Gabor dictionary for analysis. In the sequel, we assume an orthonormal basis for synthesis representations, or a tight-frame for analysis representations; hence, $E = D^T$.

The sparse approximation problem is not only ill-posed (since the matrix Φ has a nontrivial kernel), but is also known to be *NP*-hard. To circumvent these issues, we assume that the measurement matrix provides *stable embedding* (SE) with isometry constants $\mu_{K'}$ and $L_{K'}$ ($K' = K_1 + K_2$):

$$\frac{\mu_{K'}}{2} \|x_1 - x_2\|_2^2 \leq \|\Phi(x_1 - x_2)\|_2^2 \leq \frac{L_{K'}}{2} \|x_1 - x_2\|_2^2, \quad (1)$$

for all $x_j \in \Sigma_{K_j}^M$, where $\Sigma_{K_j}^M$ is the model \mathcal{M} -restricted union-of-subspaces (RUS) spanned by K_j columns of E . For simple K -sparse signals, when E is an ortho-basis, random matrices Φ satisfy the SE in (1) with $M = \mathcal{O}(K \log(N'/K))$ for sparse signals. This is the well-known K -restricted isometry property (K -RIP) in CS [8]. Similarly, when E is an overcomplete dictionary,

recent results show that random matrices Φ also satisfy the SE in (1) with $M \gtrsim K \log(N'/K)$. This property is known as the dictionary-RIP (D-RIP) [9, 10]. In general, RUS models require $M = \mathcal{O}(\log |\Sigma_{\mathcal{M}(K)}|)$ for stable embedding, e.g., K -tree model requires $M = \mathcal{O}(K)$ [8, 11, 12].

Many RUS models are endowed with a tractable algorithm \mathcal{M}_K that projects $y \in \mathbb{R}^N$ into $\Sigma_{\mathcal{M}(K)}$:

$$\mathcal{M}_K(y) = \arg \min_{x \in \Sigma_{\mathcal{M}(K)}} \|x - y\|_2. \quad (2)$$

Examples include but not limited to (i) K -sparse signals (Σ_K in short), (ii) (K, C) -sparse signals where K -sparse coefficients live in at most C unknown contiguous blocks on a chain graph, (iii) K -tree sparse where K -sparse coefficients lie on a rooted connected subtree of an N -dimensional tree, and (iv) (K, Δ) -sparse signals where K -sparse coefficients are separated by at least Δ -zeros on a chain graph. For Σ_K , \mathcal{M}_K is simple hard thresholding (e.g., keep the largest K -coefficients in magnitude while setting the others to zero). For other RUS models, efficient combinatorial and mixed integer projection algorithms exist [13].

3 Algebraic pursuits and the NIHT algorithm

In [3], Cevher proposes two algorithms, dubbed algebraic pursuits (ALPS), that fuse Nesterov's optimal gradient methods with combinatorial model-based projection algorithms for sparse approximation. For instance, the fast Lipschitz iterative hard thresholding (FLIHT) scheme of ALPS has the following recursion ($a_{i+1} = 0.5 \left(1 + \sqrt{1 + 4a_i^2}\right)$, $a_1 = 1$, and $\theta_i = \frac{a_i - 1}{a_{i+1}}$):

$$x_{i+1} = E \times \mathcal{M}_K \left(D y_i - \frac{1}{L_{3K}} E^T \nabla f(y_i) \right), \quad y_{i+1} = x_i + \theta_i (x_i - x_{i-1}). \quad (3)$$

FLIHT features the following estimation and convergence guarantee when $L_{2K} \lesssim 2\mu_{2K}$:

Theorem 1 *The i -th iteration of FIHT satisfies $f(x_i) - f(x^*) \leq \frac{2L_{3K} \|x^* - x_0\|^2}{(i+1)^2}$, where $x^* \in \Sigma_K$ is the true vector that generates u . We also have $\|x^* - x_i\| \leq \sqrt{\frac{2L_{3K}}{\mu_{2K}}} \frac{\|x^* - x_0\|}{i+1} + \frac{2\sqrt{2}}{\sqrt{\mu_{2K}}} \|n\|$.*

In this paper, we propose a third algebraic pursuit algorithm, dubbed as NIHT for Nesterov iterative hard thresholding. The algorithm is based upon Nesterov's proximal gradient method [1]. The NIHT recursion is summarized as follows (with $\tau_i = 2/(i+3)$, $x_1 = 0$, and $z_0 = 0$):

$$y_i = x_i - \frac{1}{L} \nabla f(x_i), \quad z_i = z_{i-1} - \frac{i+1}{2L} \nabla f(x_i), \quad x_{i+1} = E \times \mathcal{M}_K(D v_{i+1}), \quad v_{i+1} = \tau_i z_i + (1 - \tau_i) y_i. \quad (4)$$

To provide the online guarantee, we also keep track of a projection error variable $\varepsilon_i = \|x_i - v_i\|_2$.

In NIHT, z_i is a gradient term that minimizes the following objective based on past estimates:

$$z_i = \arg \min_{z \in \mathbb{R}^N} \left\{ \frac{L}{\sigma} d(z) + \sum_{j=0}^i \alpha_j [f(x_j) + \langle \nabla f(x_j), z - x_j \rangle] \right\}, \quad (5)$$

where $d(x) = \frac{\sigma}{2} \|x\|^2$ is the strongly convex, *prox*-function with strong convexity parameter σ , which prevents the estimates from deviating too much from the prox-center—the origin in our case.

While NIHT does not have *a priori* convergence and estimation guarantees similar to FLIHT, it has the following online estimation guarantee:

Theorem 2 *The i -th iteration of NIHT satisfies $f(y_i) - f(x^*) \leq \frac{2L \|x^*\|^2}{(i+1)^2} + \delta_i^2$, where $x^* \in \Sigma_K$ is the true sparse vector, and $\delta_i^2 = \frac{L}{2A_i} \sum_{j=1}^i \frac{\varepsilon_j^2}{A_j \tau_{j-1}^2}$; ($A_i = (i+1)(i+2)/4$). Moreover, the signal estimates satisfy the following inequality, when $\Phi \Phi^T \approx \rho \mathbf{I}$: $\frac{\sqrt{2}(1-c_{L,\rho})}{\sqrt{\mu_{2K} c_{L,\rho}}} (2\|n\| + \delta_i) \lesssim \|x - x_i\| \lesssim \sqrt{\frac{4L}{\mu_{2K} c_{L,\rho}(i+1)}} \frac{\|x\|}{\sqrt{\mu_{2K} c_{L,\rho}}} + \frac{\sqrt{2}(3-c_{L,\rho})}{\sqrt{\mu_{2K} c_{L,\rho}}} (2\|n\| + \delta_i)$, where $c_{L,\rho} = 1 - \frac{2\rho}{L}$. For column normalized, iid Gaussian matrices, $\rho = N/M$.*

We prove Theorem 2 in the appendix. In the next section we compare NIHT with ℓ_1 -minimization methods and FLIHT in the compressive sensing to demonstrate its superiority.

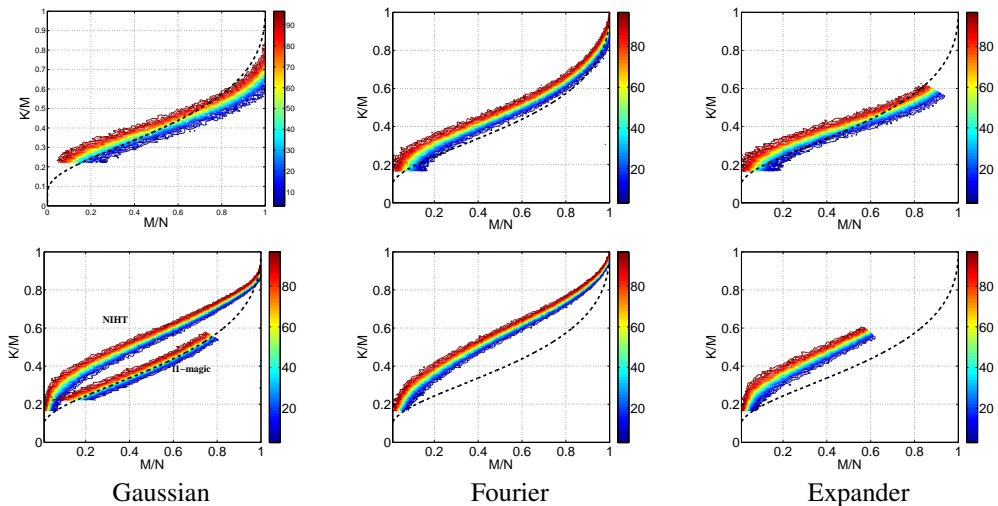


Figure 1: Phase transition curves FLIHT (top row) and NIHT (bottom row) are compared to Donoho-Tanner bound (dashed). Corresponding failure percentages are shown.

4 Experimental Results

Phase Transition. Donoho and Tanner’s combinatorial geometry based theory precisely quantifies the fundamental ℓ_1 -sparsity and compression trade-off (K vs. M) that NIHT is competing. The theory predicts the exact location in sparsity-undersampling domain where state-of-the-art algorithms exhibit phase transitions in their performance. The theory states that CS algorithms should be able to recover K -sparse signals from $M \gtrsim 2K \log(N/M)$ measurements; this threshold appears quite sharp for Gaussian, Fourier, and expander graph-based measurement matrix ensembles [14].

To see how NIHT compares to the ℓ_1 theoretical phase transitions, we perform a Monte Carlo simulations amounting to a month of CPU time. We fix the signal dimension to $N = 1000$ and sweep across K and M values (120 and 200 sample points, respectively). For each (K, M) -pair, we repeat the following 100-times: (i) generate a random sparse vector with unit norm, (ii) generate compressive measurements (no noise) using Gaussian, Fourier, and expander graph sampling matrices (incomplete), and (iii) recover the signals using NIHT and FLIHT. Both algorithms use the same number of iterations 1000, which—for the set up of our simulations— theoretically enables FLIHT and NIHT to reach an accuracy of 10^{-2} on the signal estimates (approximately 10^{-4} accuracy on the objective values). We then report the number of recoveries that obtain this accuracy or better.

Figure 1 summarizes the results, which are quite promising for NIHT. For comparison, we also provide the ℓ_1 -magic basis pursuit results (the interior point method where the Newton system is solved with conjugate gradients) [15], which match the Donoho-Tanner phase transition curve (c.f., within NIHT/Gaussian). Compared to ℓ_1 -magic, NIHT increases the number of sparse coefficients that can be recovered from the same measurements approximately by 25%. FLIHT performs on par with ℓ_1 -magic. Both FLIHT and NIHT algorithms achieve this performance at the fraction of ℓ_1 -magic’s computational cost.

Model-based recovery. Figure 2 (a) and (b) show the phase transition of NIHT with *positive* k -sparse signals, using Gaussian and Fourier sensing matrices. As before, the ℓ_1 -magic results are also provided for comparison. At the end of each iteration, the algorithm only maintains the k largest positive entries of the recovered vector. Observe that the prior positivity information, i.e. knowing that the k -sparse signal has positive values *a priori*, significantly increases the performance of the NIHT algorithm.

In this experiment, we consider a specific nested RUS model: *block sparsity*. In a block-sparse signal, the locations of the sparse coefficients cluster in blocks under a specific sorting order. Block-sparse signals have been previously studied in several different applications, including DNA microarrays and magnetoencephalography. An equivalent problem arises in signal ensembles, such as array signal processing [13], and face classification [16]. It has been shown that the block-sparse

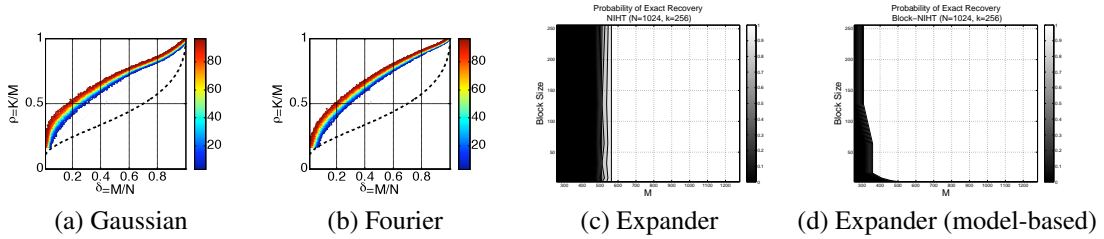


Figure 2: (a,b) Phase transition curve of the NIHT algorithm with positive sparse signals is compared to Donoho-Tanner bound (dashed). Note that the prior positivity knowledge significantly improves the reconstruction accuracy. (c,d) The impact of block-sparsity on the performance of the NIHT algorithm is significant. Exploiting the block structure in addition to signal sparsity, NIHT decreases the number of measurements significantly.

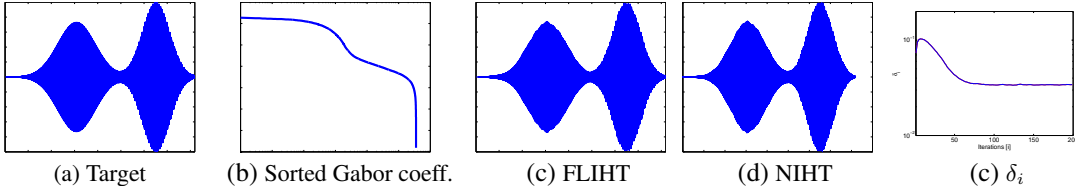


Figure 3: (a) A modulated, narrow-band signal is shown ($N=8192$). (b) Gabor analysis coefficients ($N^g=43 \times N$) of the signal are compressible. (c) FLIHT has $\|x - x_i\|/\|x\| = 0.0584$ error in CS reconstruction with the Gabor dictionary using 80 measurements (Φ is Bernoulli as in [17]). (d) NIHT outperforms FLIHT with $\|x - x_i\|/\|x\| = 0.0187$ error (approximately 10dB improvement). (e) Error variable δ_i in the online bound.

structure enables signal recovery from a reduced number of CS measurements when the recovery algorithms exploit this specific structure [13, 12].

The block sparse model approximation is quite simple: if a sparse coefficient is selected within the predefined block of size J , all the coefficients must be turned within the same block. Hence, block sparse approximation is—in a way—equivalent to unstructured sparse approximation: instead of picking the top K -coefficients by their energy, we pick the top K/J -blocks by summing up their ℓ_2 -energy. For simplicity, we consider uniform block sizes of powers of 2 on the signal vector; hence, the signal sparsity is also restricted to be a power of 2.

Figures 2 (c) and (d) investigate the advantage of incorporating the block-sparsity information on the probability of exact recovery for $N = 1024$ and $K = 256$. We vary the block sparsity level as $J = (2, 4, \dots, 128, 256)$ and also the number of expander-based measurements from $M = 256$ to 1024. Figure 2 (c) plots the probability of successful recovery while using the NIHT algorithm, whereas Figure 2 (d) plots the probability of success of the NIHT algorithm with the block-sparsity projection. We observe that the block-sparsity model significantly reduces the minimum number of measurements required for exact recovery.

Signal recovery using over-complete dictionaries. To demonstrate the great promise of the analysis sparsity model with NIHT, we recover a modulated narrow-band signal $N = 8192$ from $M = 80$ compressive measurements (c.f., Fig. 3), and compare it to FLIHT.¹ This is an unreasonably small amount of data corresponding to an under-sampling factor exceeding 100. The signal is approximately $K = 1000$ sparse in an overcomplete Gabor dictionary. FLIHT (initialized with 0) converges to a close approximation of the signal within 100 iterations. We run NIHT and monitor the error variable δ_i in the online guarantee. We provide the result at the 100-th iteration. For comparison, we also use 11-magic, using discrete cosine transform (DCT) as sparsity basis. In DCT, the signal can be closely approximated at 50-sparse. However, 11-magic (initialized with the true vector) cannot recover the signal and needs $M = 400$ for comparable performance.

Compressive Imaging. We use a real image of size 1024×1024 and generate compressive samples using a scrambled partial Fourier sensing matrix with $M = 0.33 \times N$, as described in [18]. For

¹VC thanks Michael B. Wakin for providing the code for the signal generation and the over-complete Gabor dictionary.



(a) NIHT (error: 19.81dB).

(b) ℓ_1 -minimization (error: 22.14dB).

Figure 4: Compressive imaging with NIHT and ℓ_1 optimization

sparsity basis, we pick Daubechies wavelets and judiciously choose $K = 0.15 \times N$ for sparse recovery. To recover the target image, we then run NIHT and the ℓ_1 minimization algorithm [19, 20]. Figure 4 compares the NIHT algorithm with the ℓ_1 minimization method.

5 Conclusion

NIHT algorithm creates a unifying connection between combinatorial optimization algorithms, gradient-projection methods, and the non-smooth optimization framework by Nesterov. The algorithms require two inputs: K a desired sparsity level and L the Lipschitz gradient constant, which can be overestimated by the matrix norm. NIHT effectively address the desiderata in the sparse approximation problems by (i) providing a computationally scalable algorithmic framework, (ii) having the ability to incorporate structure in sparse approximation. It further improves on the ℓ_1 -phase transition bounds in the compressed sensing problem.

Appendix: Proof of Theorem 2

Proof 1 (Theorem 2) We investigate the updates $x_i \in \Sigma_K$ and $y_i \in \mathbb{R}^N$ in (4) within the context of the following recursion \mathcal{R}_i , which trivially holds when $i = 0$ ($A_i = (i + 1)(i + 2)/4$ and $\alpha_i = (i + 1)/2$):

$$A_i f(y_i) \leq \psi_i := \min_{z \in \mathbb{R}^N} \frac{L}{\sigma} d(z) + \sum_{j=0}^i \alpha_j [f(x_j) + \langle \nabla f(x_j), z - x_j \rangle] + \frac{L}{2} \sum_{j=1}^i \frac{\varepsilon_j^2}{A_j \tau_{j-1}^2}. \quad (6)$$

Since $d(z)$ is strongly convex, the following inequality holds

$$\Rightarrow \psi_{i+1} \geq \min_{z \in \mathbb{R}^N} \psi_i + \frac{L}{2} \|z - z_i\|^2 + \alpha_{i+1} [f(x_{i+1}) + \langle \nabla f(x_{i+1}), z - x_{i+1} \rangle] + \frac{L}{2A_{i+1}\tau_i^2} \|x_{i+1} - v_{i+1}\|^2. \quad (7)$$

Consider the second term in (7): $\frac{L}{2} \|z - z_i\|^2 = \frac{L}{2\tau_i^2} \|\tau_i z - \tau_i z_i - (1 - \tau_i)y_i + (1 - \tau_i)y_i\|^2 \geq$

$$\frac{L}{2\tau_i^2} \|\tau_i z - x_{i+1} + (1 - \tau_i)y_i\|^2 - \frac{L}{2\tau_i^2} \|x_{i+1} - v_{i+1}\|^2. \quad (8)$$

Noting that $A_i + \alpha_{i+1} = A_{i+1}$, $A_{i+1}^{-1} \geq \tau_i^2$, and $1 - A_i/A_{i+1} = \tau_i$, we have

$$\begin{aligned} \psi_i + \alpha_{i+1} [f(x_{i+1}) + \langle \nabla f(x_{i+1}), z - x_{i+1} \rangle] &\geq A_i f(y_i) + \alpha_{i+1} [f(x_{i+1}) + \langle \nabla f(x_{i+1}), z - x_{i+1} \rangle] \\ &\geq A_i [f(x_{i+1}) + \langle \nabla f(x_{i+1}), y_i - x_{i+1} \rangle] + \alpha_{i+1} [f(x_{i+1}) + \langle \nabla f(x_{i+1}), z - x_{i+1} \rangle] \\ &= A_{i+1} \left\{ f(x_{i+1}) + \frac{A_i}{A_{i+1}} \langle \nabla f(x_{i+1}), y_i - x_{i+1} \rangle + \frac{\alpha_{i+1}}{A_{i+1}} [f(x_{i+1}) + \langle \nabla f(x_{i+1}), z - x_{i+1} \rangle] \right\} \\ &\geq A_{i+1} \{ f(x_{i+1}) + \langle \nabla f(x_{i+1}), \tau_i z + (1 - \tau_i)y_i - x_{i+1} \rangle \}, \text{ leading to} \end{aligned}$$

$$\begin{aligned}
\psi_{i+1} &\geq A_{i+1} \left\{ \min_{z \in \mathbb{R}^N} f(x_{i+1}) + \frac{L}{2A_{i+1}} \left\| z - \frac{1}{\tau_i} x_{i+1} + \left(\frac{1}{\tau_i} - 1 \right) y_i \right\|^2 + \langle \nabla f(x_{i+1}), \tau_i z + (1 - \tau_i) y_i - x_{i+1} \rangle \right\} \\
&\geq A_{i+1} \left\{ \min_{z \in \mathbb{R}^N} f(x_{i+1}) + \frac{L}{2} \|\tau_i z - x_{i+1} + (1 - \tau_i) y_i\|^2 + \langle \nabla f(x_{i+1}), \tau_i z + (1 - \tau_i) y_i - x_{i+1} \rangle \right\} \\
&\geq A_{i+1} \left\{ \min_{y' \in \mathbb{R}^N} f(x_{i+1}) + \frac{L}{2} \|y' - x_{i+1}\|^2 + \langle \nabla f(x_{i+1}), y' - x_{i+1} \rangle; y' = \tau_i z + (1 - \tau_i) y_i \right\} \\
&\geq A_{i+1} \left\{ f(x_{i+1}) + f(y_{i+1}) - f(x_{i+1}); y_{i+1} = x_{i+1} - \frac{1}{L} \nabla f(x_{i+1}) \right\} \geq A_{i+1} f(y_{i+1}).
\end{aligned}$$

In last step of the derivation, we exploit the following (Bregman distance) property of the objective function:

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \|\Phi(y - x)\|^2 \leq \frac{L}{2} \|x - y\|^2, \quad (9)$$

where $L = 2\|\Phi\|, \forall x, y \in \mathbb{R}^N$. Typically, the constant L from the matrix norm is a gross overestimate that could slow down the algorithm. In general, we can use a line-search approach at each iteration to determine the minimum L that satisfies (9). The guarantee below improves along with this line-search.

As the recursion holds, we have following relationship between the optimal x and y_i :

$$A_i f(y_i) \leq \psi_i \leq \frac{L}{\sigma} d(x) + A_i f(x) + A_i \delta_i^2, \quad (10)$$

where $\delta_i^2 = \frac{L}{2A_i} \sum_{j=1}^i \frac{\varepsilon_j^2}{A_j \tau_{j-1}^2}$, which implies the desired result.

To prove the estimation guarantee on the signal values, we note

$$f(y_i) - f(x) = \langle \nabla f(x), y_i - x \rangle + \|\Phi(x - y_i)\|^2. \quad (11)$$

Hence, using the guarantees on the objective values and Cauchy Schwarz inequality, we have

$$\begin{aligned}
\|\Phi(x - y_i)\|^2 &\leq \frac{L\|x - y_0\|^2}{2i} - \langle \nabla f(x), y_i - x \rangle + \delta_i^2 = \frac{L\|x - y_0\|^2}{2i} + 2 \left[\Phi^T(u - \Phi x) \right]^T (x_i - x) + \delta_i^2 \\
&= \frac{L\|x - x_0\|^2}{2i} + 2n^T \Phi(x_i - x) + \delta_i^2 \leq \frac{L\|x - x_0\|^2}{2i} + 2\|n\| \|\Phi(x_i - x)\| + \delta_i^2.
\end{aligned} \quad (12)$$

When $a^2 \leq c^2 + 2ba$, we have $a \leq c + 2b$. Moreover, if $c^2 = c_1^2 + c_2^2$, then we have $c \leq |c_1| + |c_2|$. Based on these observations, we have

$$\left\| \Phi \left(x_i - \frac{1}{L} \nabla f(x_i) - x \right) \right\| \leq \delta_i + \frac{2\sqrt{L}\|x\|}{(i+1)} + 2\|n\|, \quad (13)$$

where we used the definition of $y_i = x_i - \frac{1}{L} \nabla f(x_i)$. Now, the first term in (13) satisfies the following inequality when $\Phi\Phi^T \approx \rho\mathbf{I}$ (define $c_{L,\rho} = 1 - \frac{2\rho}{L}$):

$$\begin{aligned}
\left\| \Phi \left(x_i - \frac{1}{L} \nabla f(x_i) - x \right) \right\| &= \left\| \Phi(x_i - x) + \frac{2}{L} \Phi\Phi^T(u - \Phi x_i) \right\| \approx \left\| \Phi(x_i - x) + \frac{2\rho}{L} (\Phi x - \Phi x_i + n) \right\| \\
&= \|c_{L,\rho} \Phi(x_i - x) + (1 - c_{L,\rho})n\| \geq |c_{L,\rho}| \|\Phi(x_i - x)\| - (1 - c_{L,\rho})\|n\|.
\end{aligned} \quad (14)$$

Here, we are going to assume that the matrix Φ provides restricted isometry property (RIP), that is,

$$\frac{\mu_{K'}}{2} \|x_1 - x_2\|_2^2 \leq \|\Phi(x_1 - x_2)\|_2^2 \leq \frac{L_{K'}}{2} \|x_1 - x_2\|_2^2. \quad (15)$$

for all $x_j \in \Sigma_{K_j}$ (i.e., Σ_{K_j} is the union of all subspaces spanned by all subsets of K_j columns of Φ) with constants $\mu_{K'}$ and $L_{K'}$ ($K' = K_1 + K_2$). Combining (13) and (14) with the RIP assumption, we obtain the desired estimation guarantee.

References

- [1] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103, 2005.

- [2] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer, 2004.
- [3] V. Cevher. An ALPS view of sparse recovery. In *ICASSP (submitted)*, 2011.
- [4] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [5] P. Ravikumar and M. J. Wainwright and J. Lafferty. High-dimensional ising model selection using ℓ_1 -regularized logistic regression. *Annals of Statistics*, 28(2):1287–1319, 2010.
- [6] D. L. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, Sept. 2006.
- [7] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Info. Theory*, 52(2):489–509, 2006.
- [8] R. G. Baraniuk, V. Cevher, and M. B. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proc. of the IEEE*, 2010.
- [9] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory*, 2008.
- [10] E. J. Candes, Y. C. Eldar, and D. Needell. Compressed sensing with coherent and redundant dictionaries. 2010.
- [11] T. Blumensath and M. E. Davies. Sampling theorems for signals from the union of linear subspaces. *IEEE Trans. Info. Theory*, 2007.
- [12] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Info. Theory*, 56, 2010.
- [13] R. G. Baraniuk, V. Cevher, and M. B. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proc. of the IEEE*, 98(6), 2010.
- [14] D. L. Donoho and J. Tanner. Precise undersampling theorems. *Proc. of the IEEE*, 98(6), 2010.
- [15] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Info. Theory*, 52, 2006.
- [16] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas Huang, and Shuicheng Yan. Sparse Representation for Computer Vision and Pattern Recognition. in *Proceedings of the IEEE*, 2010.
- [17] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 2008.
- [18] R. Berinde. Advances in Sparse Signal Recovery Methods. *Master of Engineering Thesis, Massachusetts Institute of Technology*, 2009.
- [19] E. van den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.
- [20] E. van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. <http://www.cs.ubc.ca/labs/scl/spgl1>.