# Gradient Hard Thresholding Pursuit for Sparsity-Constrained Optimization

**Xiao-Tong Yuan** XTYUAN1980@GMAIL.COM
Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA
Dept. of Statistics & Biostatistics, Dept. of Computer Science, Rutgers University, Piscataway, NJ 08854, USA

**Ping Li** PINGLI@STAT.RUTGERS.EDU
Dept. of Statistics & Biostatistics, Dept. of Computer Science, Rutgers University, Piscataway, NJ 08854, USA

**Tong Zhang** TZHANG@STAT.RUTGERS.EDU
Dept. of Statistics & Biostatistics, Rutgers University, Piscataway, NJ 08854, USA

## Abstract

Hard Thresholding Pursuit (HTP) is an iterative greedy selection procedure for finding sparse solutions of underdetermined linear systems. This method has been shown to have strong theoretical guarantees and impressive numerical performance. In this paper, we generalize HTP from compressed sensing to a generic problem setup of sparsity-constrained convex optimization. The proposed algorithm iterates between a standard gradient descent step and a hard truncation step with or without debiasing. We prove that our method enjoys the strong guarantees analogous to HTP in terms of rate of convergence and parameter estimation accuracy. Numerical evidences show that our method is superior to the state-of-the-art greedy selection methods when applied to learning tasks of sparse logistic regression and sparse support vector machines.

## 1. Introduction

In this paper, we focus on the following generic sparsity-constrained optimization problem

$$\min_{x\in\mathbb{R}^p} f(x), \quad \text{s.t. } \|x\|_0 \leq k, \tag{1}$$

where $f : \mathbb{R}^p \mapsto \mathbb{R}$ is a smooth and convex cost function. Among others, several examples falling into this model include: (i) Sparsity-constrained linear regression model (Tropp & Gilbert, 2007) where the residual error is used to measure data reconstruction error; (ii) Sparsity-constrained logistic regression model (Bahmani et al., 2013) where the sigmoid loss is used to measure prediction error; (iii) Sparsity-constrained graphical models learning (Jalali et al., 2011) where the likelihood of samples drawn from an underlying probabilistic model is used to measure data fidelity.

However, due to the non-convex cardinality constraint, the problem (1) is generally NP-hard even for quadratic cost functions (Natarajan, 1995). Thus, one must instead seek approximate solutions. In particular, the special case of (1) in least square regression models has gained significant attention in the area of compressed sensing (Donoho, 2006). A vast body of greedy selection algorithms for compressing sensing have been proposed including matching pursuit (Mallat & Zhang, 1993), orthogonal matching pursuit (Pati et al., 1993), compressive sampling matching pursuit (Needell & Tropp, 2009), hard thresholding pursuit (Foucart, 2011), iterative hard thresholding (Blumensath & Davies., 2009) and subspace pursuit (Dai & Milenkovic, 2009). Those methods successively select the locations of nonzero entries and estimate their values via exploring the residual error from the previous iteration. Comparing to first-order convex optimization methods developed for $\ell_1$-regularized sparse learning (Beck & Teboulle, 2009; Langford et al., 2009), those greedy selection algorithms often exhibit similar accuracy guarantees but more attractive computational efficiency.

The least square error used in compressed sensing, however, is not an appropriate measure of discrepancy in a variety of applications beyond signal processing. For example, in statistical machine learning the log-likelihood function is commonly used in logistic regression problems (Bishop, 2006) and graphical models learning (Jalali et al., 2011; Ravikumar et al., 2011). It is thus desirable to investigate theory and algorithms applicable to a broader class of sparsity-constrained learning problems as given in (1). To this end, several forward selection algorithms have been proposed to select nonzero entries in a sequential fashion (Kim & Kim, 2004; Shalev-Shwartz et al., 2010; Yuan & Yan, 2013). This category of methods date back

*Proceedings of the $31^{st}$ International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. 

to the Frank-Wolfe method (Frank & Wolfe, 1956). The forward greedy selection method has also been generalized to minimize a convex objective over the linear hull of a collection of atoms (Yuan & Yan, 2013). To make the greedy selection procedure more adaptive, Zhang (2008) proposed a forward-backward algorithm which takes backward steps adaptively whenever beneficial. Jalali et al. (2011) applied this forward-backward selection algorithm to learn the structure of a sparse graphical model. More recently, Bahmani et al. (2013) proposed a gradient hard-thresholding method which generalizes compressive sampling matching pursuit (Needell & Tropp, 2009) from compressed sensing to general sparsity-constrained optimization problems. The hard-threshholding-type methods have also been shown to be statistically and computationally efficient for sparse principal component analysis (Yuan & Zhang, 2013; Ma, 2013).

### 1.1. Our contribution

In this paper, inspired by the success of Hard Thresholding Pursuit (HTP) (Foucart, 2011; 2012) in compressed sensing, we propose the Gradient Hard Thresholding Pursuit (GraHTP) method to encompass the sparse estimation problems arising from applications with general nonlinear models. At each iteration, GraHTP performs standard gradient descent followed by a hard truncation operation which first selects the top $k$ (in magnitude) entries of the resultant vector and then (optionally) conducts debiasing on the selected entries. We prove that under mild conditions GraHTP (with or without debiasing) has strong theoretical guarantees analogous to HTP in terms of convergence rate and parameter estimation accuracy. We have applied GraHTP to two popular machine learning models, sparse logistic regression and sparse support vector machines, verifying that the guarantees of HTP are valid for these models. Empirically we demonstrate that GraHTP is comparable or superior to the state-of-the-art greedy selection methods in these two sparse learning models.

### 1.2. Notation and outline

**Notation**: In the following, $x \in \mathbb{R}^p$ is a vector and $F$ is an index set. We denote $[x]_i$ its $i$-th entry, $x_F$ the restriction of $x$ to index set $F$, and $x_k$ the restriction of $x$ to the top $k$ (in modulus) entries, $\mathrm{supp}(x)$ the index set of non-zero entries of $x$, $\mathrm{supp}(x, k)$ the index set of its top $k$ (in modulus) entries, $\|x\| = \sqrt{x^\top x}$ the Euclidean norm, $\|x\|_1 = \sum_{i=1}^d |x_i|$ the $\ell_1$-norm, and $\|x\|_0$ the number of nonzero of vector $x$.

**Outline**: We present in §2 the GraHTP algorithm. The convergence guarantees of GraHTP are provided in §3. The specializations of GraHTP in logistic regression and support vector machines are investigated in §4. Monte-Carlo simulations and experimental results on real data are presented in §5. Finally, we conclude the paper in §6.

## 2. Gradient Hard Thresholding Pursuit

GraHTP is an iterative greedy selection procedure for approximately optimizing the non-convex problem (1). A high level summary of GraHTP is described in the top panel of Algorithm 1. The procedure generates a sequence of intermediate $k$-sparse vectors $x^{(0)}, x^{(1)}, \ldots$ from an initial sparse approximation $x^{(0)}$ (typically $x^{(0)} = 0$). At the $t$-th iteration, the first step (**S1**), $\tilde{x}^{(t)} = x^{(t-1)} - \eta \nabla f(x^{(t-1)})$, computes the gradient descent at the point $x^{(t-1)}$ with stepsize $\eta$. Then in the second step (**S2**), the $k$ coordinates of the vector $\tilde{x}^{(t)}$ that have the largest magnitude are chosen as the support in which pursuing the minimization will be most effective. In the third step (**S3**), we find a vector with this support that minimizes the objective function, which becomes $x^{(t)}$ for the next iteration. This last step, often referred to as *debiasing*, has been shown to improve the performance in other algorithms too (see, e.g., Shalev-Shwartz et al., 2010). The iterations continue until the algorithm reaches a terminating condition, e.g., on the change of the cost function or the change of the estimated minimum from the previous iteration. A natural criterion here is $F^{(t)} = F^{(t-1)}$ (see **S2** for the definition of $F^{(t)}$), since then $x^{(\tau)} = x^{(t)}$ for all $\tau \geq t$, although there is no guarantee that this should occur. It will be assumed throughout the paper that the cardinality $k$ is known. In practice this quantity may be regarded as a tuning parameter of the algorithm via, for example, cross-validations.

In the standard form of GraHTP, the debiasing step **S3** requires to minimize $f(x)$ over the support $F^{(t)}$. If this step is judged too costly, we may consider instead a fast variant of GraHTP, where the debiasing is replaced by a simple truncation operation $x^{(t)} = \tilde{x}^{(t)}_k$. This leads to the Fast GraHTP (FGraHTP) described in the bottom panel of Algorithm 1. It is interesting to note that FGraHTP can be regarded as a projected gradient descent procedure for optimizing the non-convex problem (1). Its per-iteration computational overload is almost equal to that of the standard gradient descent procedure. While in this paper we only study the Fast GraHTP outlined in Algorithm 1, we should mention that other fast variants of GraHTP can also be considered. For instance, to reduce the computational cost of **S3**, we can take a restricted Newton step or a restricted gradient descent step to calculate $x^{(t)}$.

We close this section by pointing out that, in the special case where the cost function is the squared error $f(x) = \frac{1}{2}\|y - Ax\|^2$, GraHTP reduces to HTP (Foucart, 2011). Specifically, the gradient descent step **S1** reduces to $\tilde{x}^{(t)} = x^{(t-1)} + \eta A^\top (y - Ax^{(t-1)})$ and the debiasing step **S3** reduces to the orthogonal projection $x^{(t)} = \arg\min\{\|y - Ax\|, \mathrm{supp}(x) \subseteq F^{(t)}\}$. In the meanwhile, FGraHTP re-

duces to IHT (Blumensath & Davies., 2009) in which the iteration becomes $x^{(t)} = (x^{(t-1)} + \eta A^\top(y - Ax^{(t-1)}))_k$.

**Algorithm 1:** Gradient Hard Thresholding Pursuit (GraHTP).

**Initialization:** $x^{(0)}$ *with* $\|x^{(0)}\|_0 \leq k$ *(typically* $x^{(0)} = 0$*),* $t = 1$.
**Output**: $x^{(t)}$.
**repeat**
  (**S1**) Compute $\tilde{x}^{(t)} = x^{(t-1)} - \eta\nabla f(x^{(t-1)})$;
  (**S2**) Let $F^{(t)} = \text{supp}(\tilde{x}^{(t)}, k)$ be the indices of $\tilde{x}^{(t)}$ with the largest $k$ absolute values;
  (**S3**) Compute $x^{(t)} = \arg\min\{f(x), \text{supp}(x) \subseteq F^{(t)}\}$;
  $t = t + 1$;
**until** *halting condition holds*;
————————★ *Fast GraHTP* ★————————
**repeat**
  Compute $\tilde{x}^{(t)} = x^{(t-1)} - \eta\nabla f(x^{(t-1)})$;
  Compute $x^{(t)} = \tilde{x}^{(t)}_k$ as the truncation of $\tilde{x}^{(t)}$ with top $k$ (in magnitude) entries preserved;
  $t = t + 1$;
**until** *halting condition holds*;

## 3. Theoretical Analysis

In this section, we analyze the theoretical properties of GraHTP and FGraHTP. We first study the convergence of these two algorithms. Then we investigate their sparse recovery performance in terms of convergence rate and parameter estimation accuracy. The technical proofs of the main theoretical results can be found in an extended version of this paper online available at *arxiv:1311.5750*.

We require the following key technical condition under which the convergence and parameter estimation accuracy of GraHTP/FGraHTP can be guaranteed. To simplify the notation in the following analysis, we abbreviate $\nabla_F f = (\nabla f)_F$ and $\nabla_s f = (\nabla f)_s$ (Recall the definition of $x_F$ and $x_s$ of a vector $x$ in §1.2).

**Definition 1** (Condition $C(s, \zeta, \rho_s)$)**.** *For any integer $s > 0$, we say $f$ satisfies Condition $C(s, \zeta, \rho_s)$ if for any index set $F$ with cardinality $|F| \leq s$ and any $x, y$ with supp$(x) \cup$ supp$(y) \subseteq F$, the following inequality holds for some $\zeta > 0$ and $0 < \rho_s < 1$:*

$$\|x - y - \zeta\nabla_F f(x) + \zeta\nabla_F f(y)\| \leq \rho_s\|x - y\|.$$

**Remark 1.** *In the special case where $f(x)$ is least square loss function and $\zeta = 1$, Condition $C(s, \zeta, \rho_s)$ reduces to the well-known* Restricted Isometry Property *(RIP) condition in compressed sensing.*

We may establish the connections between condition $C(s, \zeta, \rho_s)$ and the conditions of restricted strong convexity/smoothness which are extensively used in the analysis of previous greedy selection methods (Zhang, 2008; Shalev-Shwartz et al., 2010; Yuan & Yan, 2013; Bahmani et al., 2013).

**Definition 2** (Restricted Strong Convexity/Smoothness)**.** *For any integer $s > 0$, we say $f(x)$ is restricted $m_s$-strongly convex and $M_s$-strongly smooth if there exist $\exists m_s, M_s > 0$ such that for all $\|x - y\|_0 \leq s$,*

$$\frac{m_s}{2}\|x - y\|^2 \leq \Delta f(x, y) \leq \frac{M_s}{2}\|x - y\|^2, \qquad (2)$$

*where $\Delta f(x, y) := f(x) - f(y) - \langle\nabla f(y), x - y\rangle$ is the Bregman divergence associated with $f$ for $x, y$.*

The following proposition connects condition $C(s, \zeta, \rho_s)$ to the restricted strong convexity/smoothness conditions.

**Proposition 1.** *Assume that $f$ is a differentiable function.*

*(a) If $f$ satisfies Condition $C(s, \zeta, \rho_s)$, then for all $\|x - y\|_0 \leq s$ the following two inequalities hold:*

$$\frac{1 - \rho_s}{\zeta}\|x - y\| \leq \|\nabla_F f(x) - \nabla_F f(y)\| \leq \frac{1 + \rho_s}{\zeta}\|x - y\|,$$

$$\Delta f(x, y) \leq \frac{1 + \rho_s}{2\zeta}\|x - y\|^2.$$

*(b) If $f$ is $m_s$-strongly convex and $M_s$-strongly smooth, then $f$ satisfies condition $C(s, \zeta, \rho_s)$ with any*

$$\zeta < 2m_s/M_s^2, \quad \rho_s = \sqrt{1 - 2\zeta m_s + \zeta^2 M_s^2}.$$

**Remark 2.** *Part (a) of Proposition 1 indicates that if Condition $C(s, \zeta, \rho_s)$ holds, then $f$ is strongly smooth. Part (b) of Proposition 1 shows that the strong smoothness/convexity conditions imply Condition $C(s, \zeta, \rho_s)$. Therefore, Condition $C(s, \zeta, \rho_s)$ is no stronger than the strong smoothness/conveixy conditions.*

### 3.1. Convergence

We now analyze the convergence properties of GraHTP and FGraHTP. First and foremost, we make a simple observation about GraHTP: since there is only a finite number of subsets of $\{1, ..., p\}$ of size $k$, the sequence defined by GraHTP is eventually periodic. The importance of this observation lies in the fact that, as soon as the convergence of GraHTP is established, then we can certify that the limit is exactly achieved after a finite number of iterations. We establish below the convergence of GraHTP and FGraHTP under proper conditions.

**Theorem 1.** *Assume $f$ satisfies Condition $C(2k, \zeta, \rho_{2k})$ and the step-size $\eta < \zeta/(1 + \rho_{2k})$. Then the sequence $\{x^{(t)}\}$ defined by GraHTP terminates after a finite number of iterations. Moreover, the sequence $\{f(x^{(t)})\}$ defined by FGraHTP converges.*

**Remark 3.** *Since $\rho_{2k} \in (0,1)$, we have that the convergence results in Theorem 1 hold whenever the step-size $\eta < \zeta/2$. If $f$ is $m_{2k}$-strongly convex and $M_{2k}$-strongly smooth, then from Part(b) of Proposition 1, we know that Theorem 1 holds if we choose the step-size $\eta < m_{2k}/M_{2k}^2$.*

### 3.2. Sparse recovery performance

The following theorem is our main result on the parameter estimation accuracy of GraHTP and FGraHTP when the target solution is sparse.

**Theorem 2.** *Let $\bar{x}$ be an arbitrary $\bar{k}$-sparse vector and $k \geq \bar{k}$. Let $s = 2k+\bar{k}$. Assume $f$ satisfies Condition $C(s, \zeta, \rho_s)$ and the step size $\eta < \zeta$.*

*(a) If $\mu_1 = \sqrt{2}(1-\eta/\zeta+(2-\eta/\zeta)\rho_s)/(1-\rho_s) < 1$, then at iteration $t$, GraHTP will recover an approximation $x^{(t)}$ satisfying*

$$\|x^{(t)} - \bar{x}\| \leq \mu_1^t \|x^{(0)} - \bar{x}\| + \frac{2\eta+\zeta}{(1-\mu_1)(1-\rho_s)}\|\nabla_k f(\bar{x})\|.$$

*(b) If $\mu_2 = 2(1-\eta/\zeta+(2-\eta/\zeta)\rho_s) < 1$, then at iteration $t$, FGraHTP will recover an approximation $x^{(t)}$ satisfying*

$$\|x^{(t)} - \bar{x}\| \leq \mu_2^t \|x^{(0)} - \bar{x}\| + \frac{2\eta}{1-\mu_2}\|\nabla_s f(\bar{x})\|.$$

Note that we did not make any attempt to optimize the constants $(\mu_1, \mu_2)$ in Theorem 2, which are relatively loose. In the following discussion, we ignore the constants and focus on the main message Theorem 2 conveys.

Part (a) of Theorem 2 indicates that under proper conditions, the estimation error of GraHTP to a target sparse vector $\bar{x}$ is determined by the multiple of $\|\nabla_k f(\bar{x})\|$, and the rate of convergence before reaching this error level is *geometric*. Particularly, if the sparse vector $\bar{x}$ is sufficiently close to an unconstrained minimum of $f$ then the estimation error floor is negligible because $\nabla_k f(\bar{x})$ has small magnitude. In the ideal case where $\nabla f(\bar{x}) = 0$ (i.e., the sparse vector $\bar{x}$ is an unconstrained minimum of $f$), this result guarantees that we can recover $\bar{x}$ to arbitrary precision. In this case, if we further assume that $\eta$ satisfies the conditions in Theorem 1, then exact recovery is guaranteed in a *finite* number of iterations which is at most

$$T = \left\lceil \frac{\ln\left(\min_i |[\bar{x}]_i| / \|x^{(0)} - \bar{x}\|\right)}{\ln \mu_1} \right\rceil.$$

Indeed, we have $\|x^{(T)} - \bar{x}\| \leq \min_i |[\bar{x}]_i|$ and together with $k \geq \bar{k}$ we know that $\mathrm{supp}(\bar{x}) \subseteq \mathrm{supp}(x^{(T)})$, and thus $x^{(T)} = \bar{x}$ due to the global optimality of $\bar{x}$.

Part (b) of Theorem 2 shows that FGraHTP enjoys a similar geometric rate of convergence and the estimation error is determined by the multiple of $\|\nabla_s f(\bar{x})\|$ with $s = 2k + \bar{k}$. The shrinkage rates $\mu_1 < 1$ (see Part (a)) and $\mu_2 < 1$ (see Part (b)) respectively control the convergence rate of GraHTP and FGraHTP. For GraHTP, the condition $\mu_1 < 1$ implies

$$\eta > \frac{((2\sqrt{2}+1)\rho_s + \sqrt{2} - 1)\zeta}{\sqrt{2} + \sqrt{2}\rho_s}. \tag{3}$$

By combining this condition with $\eta < \zeta$, we can see that $\rho_s < 1/(\sqrt{2}+1)$ is a necessary condition to guarantee $\mu_1 < 1$. On the other side, if $\rho_s < 1/(\sqrt{2}+1)$, then we can always find a step-size $\eta < \zeta$ satisfying (3) such that $\mu_1 < 1$. This condition of $\rho_s$ is analogous to the RIP condition for estimation from noisy measurements in compressed sensing (Candès et al., 2006; Needell & Tropp, 2009; Foucart, 2011). Indeed, in this setup our GraHTP algorithm reduces to HTP which requires weaker RIP condition than prior compressed sensing algorithms. The guarantees of GraHTP and HTP for compressed sensing are almost identical, although we did not make any attempt to optimize the RIP sufficient constants, which are $1/(\sqrt{2}+1)$ (for GraHTP) versus $1/\sqrt{3}$ (for HTP). We would like to emphasize that the condition $\rho_s < 1/(\sqrt{2}+1)$ derived for GraHTP also holds in fairly general setups beyond compressed sensing. For FGraHTP we have very similar discussions.

For the general sparsity-constrained optimization problem, we note that a similar estimation error bound has been established for the GraSP (Gradient Support Pursuit) method (Bahmani et al., 2013) which is another hard-thresholding-type method. At time stamp $t$, GraSP first conducts debiasing over the union of the top $k$ entries of $x^{(t-1)}$ and the top $2k$ entries of $\nabla f(x^{(t-1)})$, then it selects the top $k$ entries of the resultant vector and updates their values via debiasing, which becomes $x^{(t)}$. Our GraHTP is connected to GraSP in the sense that the $k$ largest absolute elements after the gradient descent step (see **S1** and **S2** of Algorithm 1) will come from some combination of the largest elements in $x^{(t-1)}$ and the largest elements in the gradient $\nabla f(x^{(t-1)})$. Although the convergence rate are of the same order, the per-iteration cost of GraHTP is cheaper than GraSP: at each debiasing step, GraSP minimizes the objective over a support of size $3k$ while that size for GraHTP is $k$. FGraHTP is even cheaper for iteration as it does not need any debiasing operation. We will compare the actual numerical performance of these methods in our empirical study. We also point out that our FGraHTP algorithm is identical [1] to the nonlinear-IHT investigated in (Blumensath, 2013). The estimation error bound there, however, is different from ours in the sense that the bound there is dependent on the objective value at the target solu-

[1] Our work was initially submitted in May 2013, which was later posted online at *arxiv:1311.5750*. We appreciate Blumensath for pointing out to us that FGraHTP in *arxiv:1311.5750* is closely related to the new work (Blumensath, 2013).

tion; whereas ours is dependent on the restricted norm of gradient at the target solution. As we will show in the next section, our bound is especially useful when applied to analyze the statistical efficiency of sparse learning problems. Also, the results in (Blumensath, 2013) are obtained under a stronger condition of restricted strict convexity property which implies the $C(s, \zeta, \rho_s)$ condition in our analysis.

Last but not least, the bounds in Theorem 2 are relying on the values of $\zeta$ and $\rho_s$. Also, the step-size $\eta$ is required to satisfy certain conditions relying on $(\zeta, \rho_s)$ to guarantee the convergence. As we will show in the next section, the values of $\zeta$ and $\rho_s$ can be estimated from data, and thus is $\eta$, for several popular machine learning models.

# 4. Applications to Sparsity-Constrained M-estimation

In this section, we will specialize GraHTP/FGraHTP to M-estimation (maximum likelihood type estimator) which is a poplar class of statistical learning models. Given a set of $n$ independently drawn data samples $\{x^{(i)}\}_{i=1}^n$, the essential form of the M-estimation problem is defined as to minimize the following risk function averaged over the samples:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \phi(x^{(i)} \mid w),$$

where and $\phi$ is a cost function and $w$ is a set of adjustable parameters. The sparsity-constrained M-estimation problem is then given by:

$$\min_w f(w), \quad \text{subject to } \|w\|_0 \leq k. \tag{4}$$

In the subsequent subsections, we will consider two instances of this model: sparse logistic regression and sparse support vector machines (SVMs).

## 4.1. Sparsity-constrained $\ell_2$-regularized logistic regression

Logistic regression is one of the most popular models in statistics and machine learning (Bishop, 2006). In this model the relation between the random feature vector $u \in \mathbb{R}^p$ and its associated random binary label $v \in \{-1, +1\}$ is determined by the conditional probability

$$\mathbb{P}(v|u; \bar{w}) = \frac{\exp(2v\bar{w}^\top u)}{1 + \exp(2v\bar{w}^\top u)}, \tag{5}$$

where $\bar{w} \in \mathbb{R}^p$ denotes a parameter vector. Given a set of $n$ independently drawn data samples $\{(u^{(i)}, v^{(i)})\}_{i=1}^n$, logistic regression learns the parameters $w$ so as to minimize the logistic loss given by

$$l(w) := \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-2v^{(i)}w^\top u^{(i)})).$$

It is well-known that $l(w)$ is convex. Unfortunately, in high-dimensional setting, i.e., $n < p$, the problem can be underdetermined and thus its minimum is not unique. A conventional way to avoid singularity is to impose $\ell_2$-regularization, resulting in the following sparsity-constrained $\ell_2$-regularized logistic regression problem:

$$\min_w f(w) = l(w) + \frac{\lambda}{2}\|w\|^2, \quad \text{subject to } \|w\|_0 \leq k, \tag{6}$$

where $\lambda > 0$ is the regularization strength parameter. Obviously $f(w)$ is $\lambda$-strongly convex and hence it has a unique minimum. The presence of the cardinality constraint enforces the solution to be sparse.

### 4.1.1. Verifying condition $C(s, \zeta, \rho_s)$

Let $U = [u^{(1)}, ..., u^{(n)}] \in \mathbb{R}^{p\times n}$ be the design matrix and $\sigma(z) = 1/(1 + \exp(-z))$ be the sigmoid function. In the case of $\ell_2$-regularized logistic loss, we have

$$\nabla f(w) = Ua(w)/n + \lambda w,$$

where the vector $a(w) \in \mathbb{R}^n$ is given by $[a(w)]_i = -2v^{(i)}(1 - \sigma(2v^{(i)}w^\top u^{(i)}))$. The following result verifies $f(w)$ satisfies Condition $C(s, \zeta, \rho_s)$ under mild conditions.

**Proposition 2.** *Given an integer $s$, let $R_s := \max_i \|(u^{(i)})_s\|$. Then the $\ell_2$-regularized logistic loss satisfies Condition $C(s, \zeta, \rho_s)$ with any*

$$\zeta < \frac{2\lambda}{(4\sqrt{s}R_s^2 + \lambda)^2},\ \rho_s = \sqrt{1 - 2\zeta\lambda + \zeta^2(4\sqrt{s}R_s^2 + \lambda)^2}.$$

**Remark 4.** *Since $R_s$ is known for a given data and $\lambda$ is fixed in the model, Proposition 2 indicates that the values of $\zeta$ and $\rho_s$ can be explicitly calculated in a data-driven way. For example, we may set $\zeta = \lambda/(4\sqrt{s}R_s^2 + \lambda)^2$ and $\rho_s = \sqrt{1 - \lambda^2/(4\sqrt{s}R_s^2 + \lambda)^2}$ to guarantee the $C(s, \zeta, \rho_s)$ condition. By Theorem 2, these two values further guide us to select the step-size $\eta = O(\lambda/(4\sqrt{s}R_s^2 + \lambda)^2)$.*

### 4.1.2. Bounding the estimation error

We are going to bound $\|\nabla_s f(\bar{w})\|$ which we obtain from Theorem 2 that controls the estimation error bounds of GraHTP (with $s = k$) and FGraHTP (with $s = 2k + \bar{k}$). In the following deviation, we assume that the joint density of the random vector $(u, v) \in \mathbb{R}^{p+1}$ is given by the following exponential family distribution:

$$\mathbb{P}(u, v; \bar{w}) = \exp\left(v\bar{w}^\top u + B(u) - A(\bar{w})\right),$$

where $A(\bar{w})$ is the log-partition function. The term $B(u)$ characterizes the marginal behavior of $u$. Obviously, the conditional distribution of $v$ given $u$, $\mathbb{P}(v \mid u; \bar{w})$, is given by the logistical model (5). By trivial algebra we can obtain the following standard result which shows

that the first derivative of the logistic loss $l(w)$ yields the cumulants of the random variables $v[u]_j$ (see, e.g., Wainwright & Jordan, 2008):

$$\frac{\partial l}{\partial [w]_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{-v^{(i)}[u^{(i)}]_j + \mathbb{E}_v[v[u^{(i)}]_j \mid u^{(i)}]\right\}.$$

Here the expectation $\mathbb{E}_v[\cdot \mid u]$ is taken over the conditional distribution (5). We introduce the following sub-Gaussian condition on the random variate $v[u]_j$.

**Assumption 1.** *For all $j$, we assume that there exists constant $\sigma > 0$ such that for all $\eta$, $\mathbb{E}[\exp(\eta v[u]_j)] \leq \exp\left(\sigma^2\eta^2/2\right)$.*

This assumption holds when $[u]_j$ are sub-Gaussian (e.g., Gaussian or bounded) random variables. The following result establishes the bound of $\|\nabla_s f(\bar{w})\|$.

**Proposition 3.** *If Assumption 1 holds, then with probability at least $1 - 4p^{-1}$,*

$$\|\nabla_s f(\bar{w})\| \leq 4\sigma\sqrt{s\ln p/n} + \lambda\|\bar{w}_s\|.$$

**Remark 5.** *If we choose $\lambda = O(\sqrt{\ln p/n})$, then with overwhelming probability, the term $\|\nabla_s f(\bar{w})\|$ vanishes at the rate of $O(\sqrt{s\ln p/n})$. This bound is superior to the bound provided by Bahmani et al. (2013, Section 4.2) which is non-vanishing.*

### 4.2. Sparsity-constrained SVMs

We now consider applying our algorithms to fast and scalable parameter learning of linear SVMs with sparsity constraint. SVMs usually map instance vectors into a high-dimensional (even infinite-dimensional) space, and solve the dual problem with a nonlinear kernel. However, dual approaches may face trouble dealing with explosion of variables when datasets scale up. Moreover in practical domains such as document analysis, data come intrinsically with high dimensions (e.g., bag of words), so that SVMs with/without nonlinear mappings often yield similar performance. As concluded in (Chappelle, 2007), when the goal is to find an approximate solution, primal linear optimizations are superior. Henceforth we focus on linear SVMs in the primal form. Particularly, we are interested in the following sparsity-constrained $L_2$-SVMs:

$$\min_w f(w) = h(w) + \frac{\lambda}{2}\|w\|^2, \quad \text{subject to } \|w\|_0 \leq k, \quad (7)$$

where

$$h(w) := \frac{1}{2n}\sum_{i=1}^{n}\left(\max\left\{0, 1 - v^{(i)}w^\top u^{(i)}\right\}\right)^2$$

is the $L_2$-hinge loss. For this class of SVMs, the objective $f(w)$ is smooth and $\lambda$-strongly convex, and the cardinality constraint enforces the solution to be sparse. We refer the $f(w)$ defined in (7) as regularized $L_2$-hinge loss in the remaining text.

#### 4.2.1. Verifying condition $C(s, \zeta, \rho_s)$

It is easy to verify that

$$\nabla f(w) = -\frac{1}{n}\sum_{i=1}^{n}\max\left\{0, 1 - v^{(i)}w^\top u^{(i)}\right\}v^{(i)}u^{(i)} + \lambda w.$$

Similar to Proposition 2, we can prove the following result which confirms that $L_2$-hinge loss $f(w)$ satisfies Condition $C(s, \zeta, \rho_s)$ under mild conditions.

**Proposition 4.** *Given an integer $s$, let $R_s := \max_i \|(u^{(i)})_s\|$. Then the $L_2$-hinge loss satisfies Condition $C(s, \zeta, \rho_s)$ with any*

$$\zeta < \frac{2\lambda}{(R_s + \lambda)^2}, \quad \rho_s = \sqrt{1 - 2\zeta\lambda + \zeta^2(R_s + \lambda)^2}.$$

Similar to the arguments in Remark 4, this proposition suggests that the values of $\zeta$ and $\rho_s$ can be selected in a data-driven way and these values in turn guide us to select the step-size $\eta = O(\lambda/(R_s + \lambda)^2)$.

#### 4.2.2. Bounding the estimation error

Recall that $\forall i$, $\|(u^{(i)})_s\| \leq R_s$. For any $\bar{k}$-sparse vector $\bar{w}$, it is known from Theorem 2 that the estimation error is controlled by

$$\|\nabla_s f(\bar{w})\| \leq \frac{1}{n}\sum_{i=1}^{n}\max\left\{0, 1 - v^{(i)}\bar{w}^\top u^{(i)}\right\}R_s + \lambda\|\bar{w}_s\|.$$

If the samples can be well separated by $\bar{w}$, then $\|\nabla_s f(\bar{w})\|$ tends to be small. More preciously, let $U(n)$ be the number of points in $\{(u^{(i)}, v^{(i)})\}_{i=1}^n$ with margin $v^{(i)}\bar{w}^\top u^{(i)} < 1$, then we have $\|\nabla_s f(\bar{w})\| \leq (U(n)R_s^2/n + \lambda)\|\bar{w}_s\| + U(n)R_s/n$. By choosing $\lambda = O(U(n)/n)$, we will have $\|\nabla_s f(\bar{w})\| = O(U(n)/n)$ which vanishes as long as $U(n)/n$ vanishes with respect to $n$.

## 5. Experimental Results

This section is devoted to evaluating the empirical performance of GraHTP and FGraHTP when these two methods are applied to sparse logistic regression and sparse $L_2$-SVMs learning tasks. All the considered algorithms are implemented in Matlab 7.12 running on a desktop with Intel Core i7 3.2G CPU and 16G RAM.

### 5.1. Sparsity-constrained logistic regression

#### 5.1.1. Monte-Carlo simulation

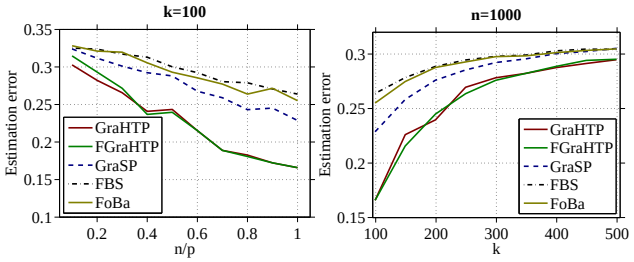
We consider a synthetic data model in which the sparse parameter $\bar{w}$ is a $p = 1000$ dimensional vector that has $\bar{k} = 100$ nonzero entries drawn independently from the standard Gaussian distribution. Each data sample $u$ is a normally distributed vector. The data labels, $v \in \{-1, 1\}$, are then generated randomly according to the Bernoulli distribution $\mathbb{P}(v = 1|u; \bar{w}) = \exp(2\bar{w}^\top u)/(1 + \exp(2\bar{w}^\top u))$. We fix the regularization parameter $\lambda = 10^{-4}$ in the objective of (6). We are interested in the following two cases:

1. **Case 1**: Cardinality $k$ is fixed and sample size $n$ is varying: we test with $k = 100$ and $n \in \{100, 200, ..., 1000\}$.

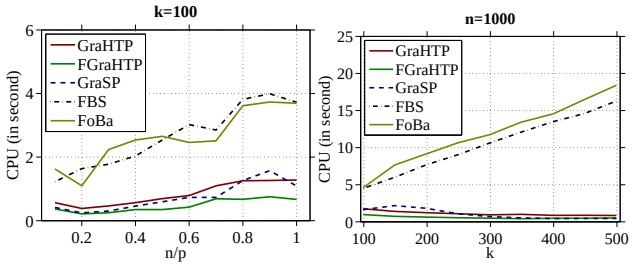
2. **Case 2**: Sample size $n$ is fixed and cardinality $k$ is varying: we test with $n = 1000$ and $k \in \{100, 150, ..., 500\}$.

For each case, we compare GraHTP and FGraHTP with three state-of-the-art greedy selection methods: GraSP (Bahmani et al., 2013) as a hard-thresholding-type method, FBS (Forward Basis Selection) (Yuan & Yan, 2013) as a forward-selection-type method, and FoBa (Zhang, 2008) as an adaptive forward backward selection method. Note that all these considered algorithms have geometric rate of convergence. We will compare the computational efficiency of these methods in this study. Theorem 2 suggests that under proper conditions GraHTP/FGraHTP are insensitive to initialization. Therefore, we simply initialize $w^{(0)} = 0$ and set the stopping criterion as $\|w^{(t)} - w^{(t-1)}\|/\|w^{(t-1)}\| \leq 10^{-4}$.

**Results.** Figure 1(a) presents the estimation errors of the considered algorithms. From the left panel of Figure 1(a) (for Case 1) we observe that: (i) when cardinality $k$ is fixed, the estimation errors of all the considered algorithms tend to decrease as sample size $n$ increases; and (ii) in this case, GraHTP and FGraHTP are comparable and they significantly outperform the other considered algorithms. From the right panel of Figure 1(a) (for Case 2) we observe that: (i) when $n$ is fixed, the estimation errors of all the considered algorithms tend to increase as $k$ increases; and (ii) in this case, GraHTP and FGraHTP are comparable and they are significantly superior to the other considered algorithms. Figure 1(b) shows the execution time of the considered algorithms. From this group of results we observe that in most cases, GraHTP/FGraHTP and GraSP are comparable and they are faster than FBS and FoBa. Also, the overall computational time of GraHTP/FGraHTP and GraSP is relatively insensitive to $k$. This is potentially because as $k$ increases, fewer iterations are needed to converge.




(a) Estimation Error




(b) CPU Running Time

*Figure 1.* Logistic regression on simulated data: estimation error and CPU time of the considered algorithms.

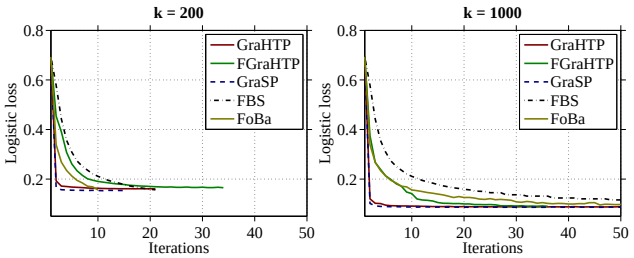
#### 5.1.2. Real data

The algorithms are also compared on two real datasets: rcv1.binary ($p = 47,236$) and news20.binary ($p = 1,355,191$). For rcv1.binary, a training subset of size 20,242 and a testing subset of size 20,000 are used. For news20.binary, a training subset of size 10,000 and a testing subset of size 9,996 are used. We test with sparsity parameters $k \in \{100, 200, ..., 1000\}$ and fix the regularization parameter $\lambda = 10^{-5}$. All the considered algorithms are initialized with $w^{(0)} = 0$ and terminated when $\|w^{(t)} - w^{(t-1)}\|/\|w^{(t-1)}\| \leq 10^{-4}$ or $t > k$.

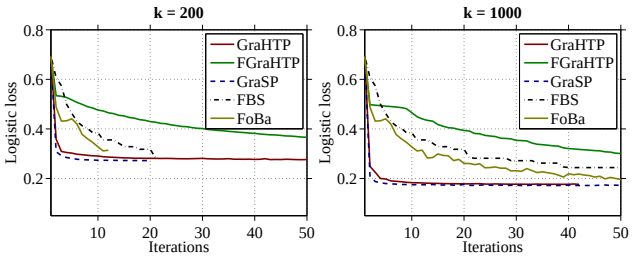
Figure 2 shows the evolving curves of empirical logistic loss for $k \in \{200, 1000\}$. It can be observed that on both datasets, GraHTP and GraSP converge much sharper than FGraHTP/FBS/FoBa. The testing classification errors and CPU running time of the considered algorithms are provided in Figure 3: (i) in terms of accuracy, all the considered methods are comparable, although GraSP is slightly more favorable; and (ii) in terms of overall execution time, FGraHTP and GraHTP are the top two ones and their computational advantage becomes significant on the relatively larger data news20.binary. To summarize, GraHTP and FGraHTP achieve favorable trade-offs between accuracy and efficiency on the considered data.

### 5.2. Sparsity-constrained $L_2$-SVMs

For this empirical study, we consider the same two real datasets and experimental protocols as used in the previous experiment. We fix the regularization parameter $\lambda = 10^{-4}$

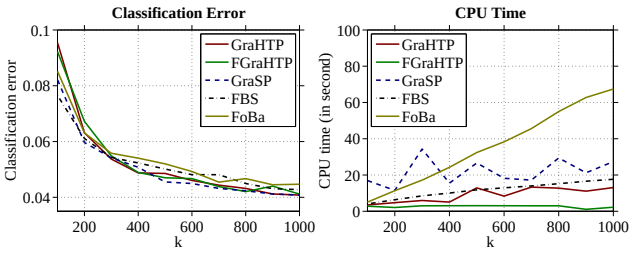



(a) rcv1.binary

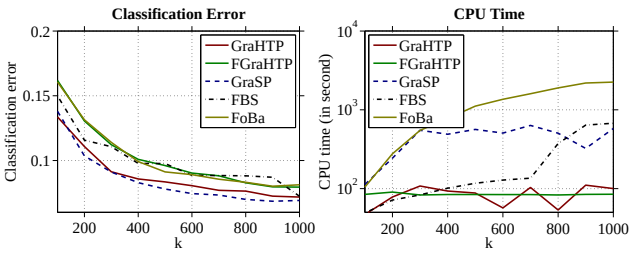



(b) news20.binary

*Figure 2.* Logistic regression on real data: $\ell_2$-regularized logistic loss *versus* number of iterations.



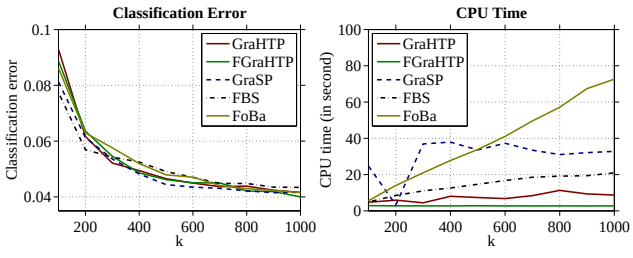

(a) rcv1.binary



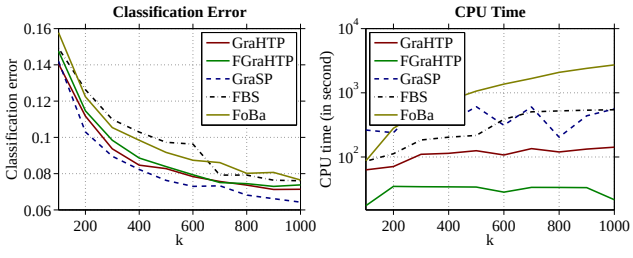

(b) news20.binary

*Figure 3.* Logistic regression on real data: Classification error and CPU running time of the considered algorithms.

and set the initial vector $w^{(0)} = 0$ for all the considered algorithms. The testing classification errors and CPU running time of the considered algorithms are provided in Figure 4. From these figures, we make very similar observations as from the previous experiment: GraHTP/FGraHTP make better trade-off between computational efficiency and classification accuracy on the used datasets.




(a) rcv1.binary




(b) news20.binary

*Figure 4.* $L_2$-SVMs on real data: Classification error and CPU running time of the considered algorithms.

## 6. Conclusions

In this paper, we have proposed GraHTP as a generalization of HTP from compressed sensing to a generic setup of sparsity-constrained minimization. The main idea is to force the gradient descent iteration to be sparse via hard truncation. Theoretically, we have proved that under mild conditions, GraHTP converges *geometrically* in finite steps of iteration and its estimation error is controlled by the restricted norm of gradient at the target sparse solution. We have also proposed and analyzed the FGraHTP algorithm as a fast variant of GraHTP without the debiasing step. Empirically, we compared GraHTP and FGraHTP with several representative greedy selection methods when applied to sparse logistic regression and sparse SVMs learning tasks. Our theoretical results and empirical evidences show that simply combing gradient descent with a truncation operation, with or without debiasing, leads to efficient and accurate computational procedures for estimating sparsity-constrained models.

## Acknowledgements

The authors thank the constructive review comments from NIPS 2013 and ICML 2014. Xiao-Tong Yuan was a postdoctoral research associate supported by NSF-DMS 0808864, NSF-EAGER 1249316, ONR-N00014-13-1-0764 and NUIST-Startup S8113029001. Ping Li is supported by ONR-N00014-13-1-0764, AFOSR-FA9550-13-1-0137, NSF III 1360971, and NSF BIGDATA 1419210. Tong Zhang is supported by NSF-IIS 1016061, NSF-DMS 1007527, and NSF-IIS 1250985.

## References


Bahmani, S., Raj, B., and Boufounos, P. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14:807–841, 2013.

Beck, A. and Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Bishop, C.M. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 978-0-387-31073-2.

Blumensath, T. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Transactions on Information Theory*, 59(6):3466–3474, 2013.

Blumensath, T. and Davies., M. E. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

Candès, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

Chappelle, O. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, 2007.

Dai, W. and Milenkovic, O. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Foucart, S. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

Foucart, S. Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. In *Approximation Theory XIII: San Antonio 2010, volume 13 of Springer Proceedings in Mathematics*, pp. 65–77, 2012.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, 5:95–110, 1956.

Jalali, A., Johnson, C. C., and Ravikumar, P. K. On learning discrete graphical models using greedy methods. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS'11)*, 2011.

Kim, Yongdai and Kim, Jinseog. Gradient lasso for feature selection. In *Proceedings Of The Twenty-First International Conference On Machine Learning (ICML'04)*, 2004.

Langford, J., Li, L., , and Zhang, T. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

Ma, Z. Sparse principal component analysis and iterative thresholding. *Annals of Statistics*, 41(2):772–801, 2013.

Mallat, S. and Zhang, Zhifeng. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Needell, D. and Tropp, J. A. Cosamp: iterative signal recovery from incomplete and inaccurate samples. *IEEE Transactions on Information Theory*, 26(3):301–321, 2009.

Pati, Y.C., Rezaiifar, R., and Krishnaprasad, P.S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, pp. 40–44, 1993.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Shalev-Shwartz, Shai, Srebro, Nathan, and Zhang, Tong. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20:2807–2832, 2010.

Tropp, J. and Gilbert, A. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

Wainwright, M.J. and Jordan, M.I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Yuan, X.-T. and Yan, S. Forward basis selection for pursuing sparse representations over a dictionary. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 35(12):3025–3036, 2013.

Yuan, X.-T. and Zhang, T. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14:899–925, 2013.

Zhang, T. Adative forward-backward greedy algorithm for sparse learning with linear models. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS'08)*, 2008.