Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima

Authored by:

Martin J. Wainwright
Po-Ling Loh

Abstract

We establish theoretical results concerning all local optima of various regularized M-estimators, where both loss and penalty functions are allowed to be nonconvex. Our results show that as long as the loss function satisfies restricted strong convexity and the penalty function satisfies suitable regularity conditions, any local optimum of the composite objective function lies within statistical precision of the true parameter vector. Our theory covers a broad class of nonconvex objective functions, including corrected versions of the Lasso for errors-in-variables linear models; regression in generalized linear models using nonconvex regularizers such as SCAD and MCP; and graph and inverse covariance matrix estimation. On the optimization side, we show that a simple adaptation of composite gradient descent may be used to compute a global optimum up to the statistical precision epsilon in log(1/epsilon) iterations, which is the fastest possible rate of any first-order method. We provide a variety of simulations to illustrate the sharpness of our theoretical predictions.

1 Paper Body

We establish theoretical results concerning local optima of regularized M estimators, where both loss and penalty functions are allowed to be nonconvex. Our results show that as long as the loss satisfies restricted strong convexity and the penalty satisfies suitable regularity conditions, any local optimum of the composite objective lies within statistical precision of the true parameter vector. Our theory covers a broad class of nonconvex objective functions, including corrected versions of the Lasso for errors-in-variables linear models and regression in generalized linear models using nonconvex regularizers such as SCAD and MCP. On the optimization side, we show that a simple adaptation of composite gradient descent may be used to compute a global optimum up to the statistical precision stat in log(1/stat) iterations, the fastest possible rate for any first-order method. We provide simulations to illustrate the sharpness of our theoretical predictions.

1

Introduction

Optimization of nonconvex functions is known to be computationally intractable in general [11, 12]. Unlike convex functions, nonconvex functions may possess local optima that are not global optima, and standard iterative methods such as gradient descent and coordinate descent are only guaranteed to converge to local optima. Although statistical results regarding nonconvex M -estimation often only provide guarantees about the accuracy of global optima, it is observed empirically that the local optima obtained by various estimation algorithms seem to be well-behaved. In this paper, we study the question of whether it is possible to certify ?good? behavior, in both a statistical and computational sense, for various nonconvex M -estimators. On the statistical level, we provide an abstract result, applicable to a broad class of (potentially nonconvex) M -estimators, which bounds the distance between any local optimum and the unique minimum of the population risk. Although local optima of nonconvex objectives may not coincide with global optima, our theory shows that any local optimum is essentially as good as a global optimum from a statistical perspective. The class of M -estimators covered by our theory includes the modified Lasso as a special case, but our results are much stronger than those implied by previous work [6]. In addition to nonconvex loss functions, our theory also applies to nonconvex regularizers, shedding new light on a long line of recent work involving the nonconvex SCAD and MCP regularizers [3, 2, 13, 14]. Various methods have been proposed for optimizing convex loss functions with nonconvex penalties [3, 4, 15], but these methods are only guaranteed to generate local optima of the composite objective, which have not been proven to be well-behaved. In contrast, our work provides a set of regularity conditions under which all local optima are guaranteed to lie within a small ball of the population-level minimum, ensuring that standard methods such as projected and composite gradient descent [10] are sufficient for obtaining estimators that lie within statistical error of the 1

truth. In fact, we establish that under suitable conditions, a modified form of composite gradient descent only requires log(1/stat) iterations to obtain a solution that is accurate up to the statistical precision stat . Notation. For functions f (n) and g(n), we write f (n) - g(n) to mean that f (n) ? cg(n) for some universal constant c ? (0, ?), and similarly, f (n) % g(n) when f (n) ? c0 g(n) for some universal constant c0 ? (0, ?). We write f (n) g(n) when f (n) - g(n) and f (n) % g(n) hold simultaneously. For a function h : Rp ? R, we write ?h to denote a gradient or subgradient, if it exists. Finally, for q, r ¿ 0, let Bq (r) denote the ‘q -ball of radius r centered around 0.

2

Problem formulation

In this section, we develop some general theory for regularized M -estimators. We first establish notation, then discuss assumptions for nonconvex regularizers and losses studied in our paper. 2.1

Background

Given a collection of n samples Z1n = {Z1 , . . . , Zn }, drawn from

a marginal distribution P over a space Z, consider a loss function Ln : Rp ? (Z)n ? R. The value Ln (?; Z1n) serves as a measure of the ?fit? between a parameter vector ? ? Rp and the observed data. This empirical loss function should be viewed as a surrogate to the population risk function L : Rp ? R, given by

L(?) := EZ Ln (?; Z1n) . Our goal is to estimate the parameter vector ? ? := arg minp L(?) that minimizes the population ??R

risk, assumed to be unique. To this end, we consider a regularized M -estimator of the form ?b ? arg min {Ln (?; Z1n) + ?? (?)} , g(?)?R

(1)

where ?? : Rp ? R is a regularizer, depending on a tuning parameter ? ¿ 0, which serves to enforce a certain type of structure on the solution. In all cases, we consider regularizers that are Pp separable across coordinates, and with a slight abuse of notation, we write ?? (?) = j=1 ?? (?j). Our theory allows for possible nonconvexity in both the loss function Ln and the regularizer ?? . Due to this potential nonconvexity, our M -estimator also includes a side constraint g : Rp ? R+ , which we require to be a convex function satisfying the lower bound g(?) ? k?k1 , for all ? ? Rp . Consequently, any feasible point for the optimization problem (1) satisfies the constraint k?k1 ? R, and as long as the empirical loss and regularizer are continuous, the Weierstrass extreme value theorem guarantees that a global minimum ?b exists. 2.2

Nonconvex regularizers

We now state and discuss conditions on the regularizer, defined in terms of ?? : R ? R. Assumption 1. (i) The function ?? satisfies ?? (0) = 0 and is symmetric around zero (i.e., ?? (t) = ?? (?t) for all t ? R). (ii) On the nonnegative real line, the function ?? is nondecreasing. (iii) For t ¿ 0, the function t 7?

?? (t) t

is nonincreasing in t.

(iv) The function ?? is differentiable for all t 6= 0 and subdifferentiable at t = 0, with nonzero subgradients at t = 0 bounded by ?L. (v) There exists ? ¿ 0 such that ??,? (t) := ?? (t) + ?t2 is convex. Many regularizers that are commonly used in practice satisfy Assumption 1, including the ‘1 -norm, ?? (?) = k?k1 , and the following commonly used nonconvex regularizers: 2

SCAD penalty: This penalty, due to Fan and Li [3], takes the form ? for —t— ? ?, ??—t—, 2 2 ?? (t) := ?(t ? 2a?—t— + ?)/(2(a ? 1)), for ? ¡ —t— ? a?, ? (a + 1)?2 /2, for —t— ¿ a?, where a ¿ 2 is a fixed parameter. Assumption 1 holds with L = 1 and ? =

(2)

1 a?1 .

MCP regularizer: This penalty, due to Zhang [13], takes the form Z —t— z 1? ?? (t) := sign(t) ? ? dz, ?b + 0

(3)

where b ¿ 0 is a fixed parameter. Assumption 1 holds with L = 1 and ? = 1b . 2.3

Nonconvex loss functions and restricted strong convexity

Throughout this paper, we require the loss function Ln to be differentiable, but we do not require it to be convex. Instead, we impose a weaker condition known as restricted strong convexity (RSC). Such conditions have been discussed in previous literature [9, 1], and involve a lower bound on the remainder in the first-order Taylor expansion of Ln . In particular, our main statistical result is based on the following RSC condition: ? log p ? 2 ? k?k21 , ?k?k2 ? 1, (4a) ? ?1 k?k2 ? ?1 n ? ? r h?Ln (? + ?) ? ?Ln (?), ?i ? ? ? ? ?2 k?k2 ? ?2 log p k?k1 , ?k?k2 ? 1, (4b) n where the ?j ?s are strictly positive constants and the ?j ?s are nonnegative constants. To understand this condition, note that if Ln were actually strongly convex, then both these RSC inequalities would hold with ?1 = ?2 ¿ 0 and ?1 = ?2 = 0. However, in the high-dimensional setting (p n), the empirical loss Ln can never be strongly convex, but the RSC condition may still hold with strictly positive (?j , ?j). On the other hand, if Ln is convex (but not strongly convex), the left-hand expression q in inequality (4) is always q nonnegative, so inequalities (4a) and (4b) hold k?k1 ?1 n ?2 n 1 trivially for k?k ? ? and k?k2 ?1 log p k?k2 ?2 log p , respectively. Hence, the RSC inequalities n o q n 1 only enforce a type of strong convexity condition over a cone set of the form k?k ? c k?k2 log p .

3

Statistical guarantees and consequences

We now turn to our main statistical guarantees and some consequences for various statistical models. Our theory applies to any vector ?e ? Rp that satisfies the first-order necessary conditions to be a local minimum of the program (1): e + ??? (?), e ? ? ?i e ? 0, h?Ln (?) for all feasible ? ? Rp . (5) When ?e lies in the interior of the constraint set, condition (5) is the usual zero-subgradient condition. 3.1

Main statistical results

Our main theorem is deterministic in nature, and specifies conditions on the regularizer, loss function, and parameters, which guarantee that any local optimum ?e lies close to the target vector ? ? = arg minp L(?). Corresponding probabilistic results will be derived in subsequent sections. For ??R

proofs and more detailed discussion of the results contained in this paper, see the technical report [7]. Theorem 1. Suppose the regularizer ?? satisfies Assumption 1, Ln satisfies the RSC conditions (4) with ?1 ¿ ?, and ? ? is feasible for the objective. Consider any choice of ? such that ( ) r 2 ?2 log p ? ? max k?Ln (?)k? , ?2 ? ? ? , (6) L n 6RL

3

2

2

2

16R max(?1 ,?2) log p. Then any vector ?e satisfying the first-order necessary conand suppose n ? ?22 ditions (5) satisfies the error bounds ? 7?L k 56?Lk ? e k? ? ? k2 ? , and k?e ? ? ? k1 ? , (7) 4(?1 ? ?) 4(?1 ? ?)

where k = k? ? k0 . From the bound (7), note that the squared ‘2 - error grows proportionally with k, the number of non2 zeros in the target q

parameter, and with ? . As will be clarified in the following sections, choosing ? proportional to

log p n

and R proportional to

1 ?

will satisfy the requirements of Theorem 1 w.h.p. for

many statistical models, in which case we have a squared ‘2 -error that scales as

k log p n ,

as expected.

Remark 1. It is worthwhile to discuss the quantity ?1 ? ? appearing in the denominator of the bound in Theorem 1. Recall that ?1 measures the level of curvature of the loss function Ln , while ? measures the level of nonconvexity of the penalty ?? . Intuitively, the two quantities should play opposing roles in our result: Larger values of ? correspond to more severe nonconvexity of the penalty, resulting in worse behavior of the overall objective (1), whereas larger values of ?1 correspond to more (restricted) curvature of the loss, leading to better behavior. We now develop corollaries for various nonconvex loss functions and regularizers of interest. 3.2

Corrected linear regression

We begin by considering the case of high-dimensional linear regression with systematically corrupted observations. Recall that in the framework of ordinary linear regression, we have the model yi = h? ? , xi i + i , — {z } P p j=1

for i = 1, . . . , n,

(8)

?j? xij

where ? ? ? Rp is the unknown parameter vector and {(xi , yi)}ni=1 are observations. Following Loh and Wainwright [6], assume we instead observe pairs {(zi , yi)}ni=1 , where the zi ?s are systematically corrupted versions of the corresponding xi ?s. Some examples include the following: (a) Additive noise: Observe zi = xi + wi , where wi ?? xi , E[wi] = 0, and cov[wi] = ?w . (b) Missing data: For ? ? [0, 1), observe zi ? Rp such that for each component j, we independently observe zij = xij with probability 1 ? ?, and zij = ? with probability ?. We use the population and empirical loss functions L(?) =

1 T ? ?x ? ? ? ?T ?x ?, 2

and

Ln (?) =

1 Tb ? ?? ? ? bT ?, 2

(9)

b ? where (?, b) are estimators for (?x , ?x ? ?) depending on {(zi , yi)}ni=1 . Then ? ? = arg min? L(?). From the formulation (1), the corrected linear regression estimator is given by

1 Tb ?b ? arg min ? ?? ? ? bT ? + ?? (?) . (10) g(?)?R 2 We now state a corollary in the case of additive noise (model (a)), where we take T b = Z Z ? ?w , ? n

and

? b=

ZT y . n

(11)

b in equation (11) is always negative-definite, so the empirical loss function When p n, the matrix ? b are applicable to missing data (model Ln previously defined (9) is nonconvex. Other choices of ? (b)), and also lead to nonconvex programs [6]. 4

Corollary 1. Suppose we have i.i.d. observations {(zi , yi)}ni=1 from a corrupted linear model with sub-Gaussian additive noise. Suppose (?, R) are chosen such that ? ? is feasible and r log p c0 c ??? . n R Then given a sample size n ? C max{R2 , k} log p, any local optimum ?e of the nonconvex program (10) satisfies the estimation error bounds ? c00 ?k c0 ? k ? e , and k?e ? ? ? k1 ? , k? ? ? k2 ? ?min (?x) ? 2? ?min (?x) ? 2? with probability at least 1 ? c1 exp(?c2 log p), where k? ? k0 = k. q ? Remark 2. When ?? (?) = ?k?k1 and g(?) = k?k1 , taking ? logn p and R = b0 k for some constant b0 ? k? ? k2 yields the required scaling n % k log p. Hence, the bounds in Corollary 1 agree with bounds in Theorem 1 of Loh and Wainwright [6]. Note, however, that the latter results are stated only for a global minimum ?b of the program (10), whereas Corollary 1 is a much stronger e Theorem 2 of our earlier paper [6] provides an indirect result holding for any local minimum ?. route for establishing similar bounds on k?e ? ? ? k1 and k?e ? ? ? k2 , since the projected gradient descent algorithm may become stuck in local minima. In contrast, our argument here does not rely on an algorithmic proof and applies to a more general class of (possibly nonconvex) penalties. Corollary 1 also has important consequences in the case where pairs {(xi , yi)}ni=1 from the linear model (8) are observed without corruption and ?? is nonconvex. Then the empirical loss Ln is equivalent to the least-squares loss, modulo a constant factor. Much existing work [3, 14] only establishes statistical consistency of global minima and then provides specialized algorithms for obtaining specific local optima that are provably close to global optima. In contrast, our results demonstrate that any optimization algorithm converging to a local optimum suffices. 3.3

Generalized linear models

Moving beyond linear regression, we now consider the case where observations are drawn from a generalized linear model (GLM). Recall that a GLM is characterized by the conditional distribution

yi h?, xi i ? ?(xTi ?) P(yi — xi , ?, ?) = exp , c(?) where ? ¿ 0 is a scale parameter and ? is the cumulant function. By standard properties of exponential families [8, 5], we have ? 0 (xTi ?) = E[yi — xi , ?, ?]. In our analysis, we assume there exists ?u ¿ 0 such that ? 00 (t) ? ?u for all t ? R. This boundedness assumption holds in various settings, including linear regression, logistic regression, and multinomial regression. The bound is required to establish both statistical consistency results in the present section and fast global convergence guarantees for our optimization algorithms in Section 4. We will assume that ? ? is sparse and optimize the penalized maximum likelihood program) (n X

1 T T ?(xi ?) ? yi xi ? + ?? (?) . ?b ? arg min n i=1 g(?)?R

(12)

We then have the following corollary: Corollary 2. Suppose we have i.i.d. observations {(xi , yi)}ni=1 from a GLM, where the xi ?s are sub-Gaussian. Suppose (?, R) are chosen such that ? ? is feasible and r c0 log p c ??? . n R 2 e Given a sample size n ? CR log p, any local optimum ? of the nonconvex program (12) satisfies ? c0 ? k c00 ?k ? e k? ? ? k2 ? , and k?e ? ? ? k1 ? , ?min (?x) ? 2? ?min (?x) ? 2? with probability at least 1 ? c1 exp(?c2 log p), where k? ? k0 = k. 5

4

Optimization algorithm

We now describe how a version of composite gradient descent may be applied to efficiently optimize the nonconvex program (1). We focus on a version of the optimization problem with the side function o 1n g?,? (?) := (13) ?? (?) + ?k?k22 , ? which is convex by Assumption 1. We may then write the program (1) as n o

?b ? arg min Ln (?) ? ?k?k22 +?g?,? (?) . (14) g?,? (?)?R — {z } ?n L

The objective function then decomposes nicely into a sum of a differentiable but nonconvex function and a possibly nonsmooth but convex penalty. Applied to the representation (14), the composite gradient descent procedure of Nesterov [10] produces a sequence of iterates {? t }? t=0 via the updates ( )

2 t

?L ? (?) 1 n

? ? ? t ?

+ g?,? (?) , ? t+1 ? arg min (15)

2 ? ? g?,? (?)?R 2 where

1 ?

is the stepsize. Define the Taylor error around ?2 in the direction ?1 ? ?2 by

T (?1 , ?2) := Ln (?1) ? Ln (?2) ? h?Ln (?2), ?1 ? ?2 i. (16) For all vectors ?2 ? B2 (3) ? B1 (R), we require the following form of restricted strong convexity: ? log p ? 2 ? k?1 ? ?2 k21 , ?k?1 ? ?2 k2 ? 3, (17a) ? ?1 k?1 ? ?2 k2 ? ?1 n r T (?1 , ?2) ? ? ? ? ?2 k?1 ? ?2 k2 ? ?2 log p k?1 ? ?2 k1 , ?k?1 ? ?2 k2 ? 3. (17b) n The conditions (17) are similar but not identical to the earlier RSC conditions (4). The main difference is that we now require the Taylor difference to be bounded below uniformly over ?2 ? B2 (3) ? B1 (R), as opposed to for a fixed ?2 = ? ? . We also assume an upper bound: log p T (?1 , ?2) ? ?3 k?1 ? ?2 k22 + ?3 k?1 ? ?2 k21 , for all ?1 , ?2 ? Rp , (18) n a condition referred to as restricted smoothness in past work [1]. Throughout this section, we assume ?i ¿ ? for all i, where ? is the coefficient ensuring the convexity of the function g?,? from equation (13). Furthermore, we define ? = min{?1 , ?2 } and ? = max{?1 , ?2 , ?3 }. The following theorem applies to any population loss function L for which the population minimizer ? ? is k-sparse and k? ? k2 ? 1, and under the scaling n ¿ Ck log p, for a constant C depending on the ?i ?s and ?i ?s. We show that the composite gradient updates (15) exhibit a type of globally geometric convergence in terms of the quantity ??? 4?

128? k logn p . 1 ? ?(n, p, k) ??? Under the stated scaling on the sample size, we are guaranteed that ? ? (0, 1). Let 0

b ?)

2 log ?(?)??(?2 log 2 ?RL ? T (?) := + 1+ log log , log(1/?) log(1/?) ?2 where ?(?) := Ln (?) + ?? (?), and define stat := k?b ? ? ? k2 . ? :=

1?

+ ?(n, p, k)

,

where ?(n, p, k) :=

(19)

(20)

Theorem 2. Suppose Ln satisfies the RSC/RSM conditions (17) and (18), and suppose ?? satisfies Assumption 1. Suppose ?b is any global minimum of the program (14), with ( ) r r log p 4 log p ? R ? c, and ? ? ? max k?Ln (?)k? , ? . n L n c2

stat Then for any stepsize ? ? 2 ? max{?3 ? ?, ?} and tolerance ? 2 ? 1?? , we have

4 2 ? k log p 2 b 2 ? k? t ? ?k ?2 + + 128? stat , ?t ? T ? (?). 2 ??? ? n

6

(21)

Remark 3. Note that for the optimal choice of tolerance parameter ? stat , the bound in inequalc2stat ity (21) takes the form ??? , meaning successive iterates are guaranteed to converge to a region b Combining Theorems 1 and 2, we have within statistical accuracy of the true global optimum ?. ! r n o k log p t t ? b , ?t ? T (c0 stat). max k? ? ?k2 , k? ? ? k2 = O n

5

Simulations

In this section, we report the results of simulations for two versions of the loss function Ln , corresponding to linear and logistic regression, and three penalty functions: Lasso, q SCAD, and MCP. In all cases, we chose regularization parameters R =

1.1 ?

? ?? (? ?) and ? =

log p n .

Linear regression: In the case of linear regression, we simulated covariates corrupted by additive noise according to the mechanism described in Section 3.2, giving the estimator

1 T XT X yT Z ?b ? arg min ? ? ?w ? ? ? + ?? (?) . (22) n n g?,? (?)?R 2 We generated i.i.d. samples xi ? N (0, I) and i ? N (0, (0.1)2), and set ?w = (0.2)2 I. Logistic regression: In the case of logistic regression, we generated i.i.d. samples xi ? N (0, I). Since ?(t) = log(1 + exp(t)), the program (12) becomes) (n X 1 {log(1 + exp(h?, xi i) ? yi h?, xi i} + ?? (?) . (23) ?b ? arg min n i=1 g?,? (?)?R We optimized the programs (22) and (23) using the composite gradient updates (15). Figure 1 shows the results of corrected linear regression with Lasso, SCAD, and MCP ? regularizers for three different problem sizes p. In each case, ? ? is a k-sparse vector with k = b pc, where the nonzero entries

were generated from a normal distribution and the vector was then rescaled so k? ? k2 = 1. As predicted by Theorem 1, the curves corresponding to the same penalty function stack up nicely n when the estimation error k?b ? ? ? k2 is plotted against the rescaled sample size k log p , and the ‘2 error decreases to zero as the number of samples increases, showing that the estimators (22) and (23) are statistically consistent. We chose the parameter a = 3.7 for SCAD and b = 3.5 for MCP. comparing penalties for corrected linear regression 0.5
0.3
0.2
0.1
0 0
p=128 p=256 p=512
0.9
l2 norm error
0.4 l2 norm error
comparing penalties for logistic regression 1
p=128 p=256 p=512
0.8 0.7 0.6 0.5
10
20 30 n/(k log p)
40
0.4 0
50
(a)
20
40 60 n/(k log p)
80
100
(b)

Figure 1. Plots showing statistical consistency of (a) linear and (b) logistic regression with Lasso, SCAD, and MCP. Each point represents an average over 20 trials. The estimation error k?b ? ? ? k2 n is plotted against the rescaled sample size k log . Lasso, SCAD, and MCP results are represented by p solid, dotted, and dashed lines, respectively.

The simulations in Figure 2 depict the optimization-theoretic conclusions of Theorem 2. Each panel shows two different families of curves, corresponding to statistical error (red) and optimization error 7

(blue). The vertical axis measures the ‘2 -error on a log scale, while the horizontal axis tracks the iteration number. The curves were obtained?by running composite gradient descent from 10 random starting points. We used p = 128, k = b pc, and n = b20k log pc. As predicted by our theory, the optimization error decreases at a linear rate until it falls to the level of statistical error. Panels (b) and (c) provide simulations for two values of the SCAD parameter a; the larger choice a = 3.7 corresponds to a higher level of curvature and produces a tighter cluster of local optima. log error plot for corrected linear regression with MCP, b = 1.5 2 opt err stat err 0

log error plot for corrected linear regression with SCAD, a = 3.7 2 opt err stat err 0

log error plot for corrected linear regression with SCAD, a = 2.5 1 opt err stat err 0

?2
?2
?1
?8
——2) ?
?3
t
?4
log(——
?6
log(——
log(——
t
t
?
?
*
?4
*
——2)
*
——2)
?2
?6
?4 ?5 ?6
?8
?10 ?12 0
200
400 600 iteration count
800
?10 0
1000
?7 200
400
(a)
600 800 iteration count
1000
?8 0
1200
200
400
(b)

600 800 iteration count
1000
1200
(c)

Figure 2. Plots illustrating linear rates of convergence for corrected linear regression with MCP and

SCAD. Red lines depict statistical error log k?b ? ? ? k2 and blue lines depict optimization error

b 2 . As predicted by Theorem 2, the optimization error decreases linearly up to statistical log k? t ? ?k accuracy. Each plot shows the solution trajectory for 10 initializations of composite gradient descent. Panel (a) shows results for MCP; panels (b) and (c) show results for SCAD with different values of a.

? Figure 3 provides analogous results to Figure 2 for logistic regression, using p = 64, k = b pc, and n = b20k log pc. The plot shows solution trajectories for 20 different initializations of composite gradient descent. Again, the log optimization error decreases at a linear rate up to the level of statistical error, as predicted by Theorem 2. Whereas the convex Lasso penalty yields a unique b SCAD and MCP produce multiple local optima. local/global optimum ?, log error plot for logistic regression with Lasso 1

log error plot for logistic regression with SCAD, a = 3.7 0.5
opt err stat err
0
——2)
?1
?6 500
1000 iteration count
(a)
1500
2000
?
?1.5
t
?
?5
?7 0
?1.5
t
?2
log(——
log(——
t
log(——
?4
?1
*
——2)

?0.5
*
*
2
——) ?
?3
opt err stat err
0
?0.5
?1 ?2
log error plot for logistic regression with MCP, b = 3 0.5
opt err stat err
0
?2.5
?2 ?2.5
?3
?3
?3.5
?3.5
?4 0
500
1000 iteration count
(b)
1500
2000
?4 0
500
1000 iteration count
1500
2000
(c)

Figure 3. Plots showing linear rates of convergence on a log scale for logistic regression. Red lines depict statistical error and blue lines depict optimization error. (a) Lasso penalty. (b) SCAD penalty. (c) MCP. Each plot shows the solution trajectory for 20 initializations of composite gradient descent.

6

Discussion

We have analyzed theoretical properties of local optima of regularized M -estimators, where both the loss and penalty function are allowed to be non-convex. Our results are the first to establish that all local optima of such non-convex problems are close to the truth, implying that any optimization method guaranteed to converge to a local optimum will provide statistically consistent solutions. We show that a variant of composite gradient descent may be used to obtain near-global optima in linear time, and verify our theoretical results with simulations. Acknowledgments PL acknowledges support from a Hertz Foundation Fellowship and an NSF Graduate Research Fellowship. MJW and PL were

also partially supported by grants NSF-DMS-0907632 and AFOSR09NL184. The authors thank the anonymous reviewers for helpful feedback. 8

2 References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. Annals of Statistics, 40(5):2452?2482, 2012. [2] P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Annals of Applied Statistics, 5(1):232?253, 2011. [3] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96:1348?1360, 2001. [4] D. R. Hunter and R. Li. Variable selection using MM algorithms. Annals of Statistics, 33(4):1617?1642, 2005. [5] E.L. Lehmann and G. Casella. Theory of Point Estimation. Springer Verlag, 1998. [6] P. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. Annals of Statistics, 40(3):1637?1664, 2012. [7] P. Loh and M.J. Wainwright. Regularized M -estimators with nonconvexity: Statistical and algorithmic theory for local optima. arXiv e-prints, May 2013. Available at http://arxiv. org/abs/1305.2436. [8] P. McCullagh and J. A. Nelder. Generalized Linear Models (Second Edition). London: Chapman & Hall, 1989. [9] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for highdimensional analysis of M -estimators with decomposable regularizers. Statistical Science, 27(4):538?557, December 2012. See arXiv version for lemma/propositions cited here. [10] Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Universit Catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007. [11] Y. Nesterov and A. Nemirovskii. Interior Point Polynomial Algorithms in Convex Programming. SIAM studies in applied and numerical mathematics. Society for Industrial and Applied Mathematics, 1987. [12] S. A. Vavasis. Complexity issues in global optimization: A survey. In Handbook of Global Optimization, pages 27?41. Kluwer, 1995. [13] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics, 38(2):894?942, 2010. [14] C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. Statistical Science, 27(4):576?593, 2012. [15] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics, 36(4):1509?1533, 2008.

9