Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparse Learning

Ryota Tomioka **Taiji Suzuki**
Department of Mathematical Informatics,
University of Tokyo.
tomioka@mist.i.u-tokyo.ac.jp
s-taiji@stat.t.u-tokyo.ac.jp

Masashi Sugiyama
Department of Computer Science,
Tokyo Institute of Technology.
sugi@cs.titech.ac.jp

Abstract

We analyze the convergence behaviour of a recently proposed algorithm for sparse learning called Dual Augmented Lagrangian (DAL). We theoretically analyze under some conditions that DAL converges super-linearly in a non-asymptotic and global sense. We experimentally confirm our analysis in a large scale ℓ_1-regularized logistic regression problem and compare the efficiency of DAL algorithm to existing algorithms.

1 Introduction

Sparse learning through convex regularization has become a common practice in many application areas including bioinformatics and natural language processing. However facing the rapid increase in the size of data-sets that we analyze everyday, clearly needed is the development of optimization algorithms that are tailored for machine learning application with diverse loss functions, possibly dense and poorly conditioned input matrices, and large number of unknowns compared to number of observations.

In this paper we consider a particular formulation of sparse learning based on the following optimization problem:

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w}) =: f(\boldsymbol{w}), \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ is the coefficient vector to be estimated, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the design matrix, and $f_\ell(\cdot)$ is a loss function. We assume that f_ℓ is a closed, proper convex function and is at least twice differentiable. Moreover, we assume that the convex conjugate $f_\ell^*(\boldsymbol{\alpha}) = \sup_{\boldsymbol{z}\in\mathbb{R}^m}\left(\boldsymbol{\alpha}^\top\boldsymbol{z} - f_\ell(\boldsymbol{z})\right)$ is strongly convex with modulus γ, i.e.

$$f_\ell^*(\boldsymbol{\alpha}') \geq f_\ell^*(\boldsymbol{\alpha}) + (\boldsymbol{\alpha}'-\boldsymbol{\alpha})^\top\nabla f_\ell^*(\boldsymbol{\alpha}) + \frac{\gamma}{2}\|\boldsymbol{\alpha}'-\boldsymbol{\alpha}\|^2. \tag{2}$$

The regularization term $\phi_\lambda(\boldsymbol{w})$ is a convex possibly non-differentiable function, e.g., $\phi_\lambda(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1$ for lasso. In addition we assume $\eta\phi_\lambda(\boldsymbol{w}) = \phi_{\eta\lambda}(\boldsymbol{w})$.

Various methods have been proposed to efficiently solve the optimization problem (1). Commonly, the *non-differentiability* of the regularization term $\phi_\lambda(\boldsymbol{w})$ has been considered as the major challenge in minimizing Eq. (1). Iteratively reweighted shrinkage (IRS) (see [1, 2, 3]; see also [4, 5] for more advanced examples of IRS) converts the minimization of Eq. (1) into a sequence smooth minimization problems by introducing a quadratic upper-bound on the regularization term; the upper-bound is tightened after every minimization. Orthant-wise limited-memory quasi-Newton (OWLQN) method [6] explicitly handles the non-differentiability of ℓ_1-regularizer; however there is no obvious way to generalize this algorithm to more general sparse regularizers.

We consider in contrast that the *coupling between variables* introduced by the design matrix $\boldsymbol{A}$ is the main source of difficulty. In fact, if $\boldsymbol{A} = \boldsymbol{I}_n$ ($n \times n$ identity matrix) and for simplicity let $f_\ell(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{b}\|^2$ and $\phi_\lambda(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1$, the solution of Eq. (1) is obtained analytically as follows:

$$w_j^* = \mathrm{ST}_\lambda(b_j) := \begin{cases} b_j - \lambda & (\text{if } b_j > \lambda), \\ 0 & (\text{if } \lambda \geq b_j \geq -\lambda), \\ b_j + \lambda & (\text{if } -\lambda > b_j) \end{cases} \qquad (j = 1, \ldots, n), \tag{3}$$

because Eq. (1) can be decomposed as follows:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{1}{2}\|\boldsymbol{w} - \boldsymbol{b}\|^2 + \lambda\|\boldsymbol{w}\|_1 \right) = \sum_{j=1}^{n} \min_{w_j\in\mathbb{R}} \left(\frac{(w_j - b_j)^2}{2} + \lambda|w_j| \right).$$

The operation in Eq. (3) is called *soft thresholding* [7, 8, 9]; we use the same notation: $\mathrm{ST}_\lambda(\boldsymbol{z}) = \mathrm{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \left(\frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 + \phi_\lambda(\boldsymbol{x})\right)$ also for the general regularization term in Eq. (1). It is known that the soft-thresholding operation can be computed analytically for the group lasso regularization [10] and also for the trace norm regularization [11].

Iterative Shrinkage Thresholding (IST) methods have recently been actively studied [7, 8, 9]. Basically IST computes the following minimization at every iteration

$$\begin{aligned} \boldsymbol{w}^{t+1} &= \operatorname*{argmin}_{\boldsymbol{w}} \left(f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + \phi_\lambda(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right) && (4) \\ &= \mathrm{ST}_{\lambda\eta_t}\left(\boldsymbol{w}^t - \eta_t \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)\right). && (5) \end{aligned}$$

One can see that there is no coupling between variables in Eq. (4) because the loss term is linearly approximated; thus what we need to do at every iteration is only to compute gradient at the current point, take a gradient step, and then perform the soft-thresholding operation (Eq. (5)).

The downside of the IST approach is the difficulty to choose the parameter η_t, which can be considered as a step-size (see Eq. (5)); this issue is especially problematic when the design matrix $\boldsymbol{A}$ is poorly conditioned. SpaRSA [12] uses approximate second order curvature information for the selection of the step-size parameter η_t. TwIST [13] and FISTA [14] are "two-step" approaches that try to alleviate the poor efficiency of IST when the design matrix is poorly conditioned. Recently proposed is the dual augmented Lagrangian (DAL) method [15], which uses an IST-like update equation with a "gradient" that is obtained by solving an inner minimization problem at every iteration. Despite being empirically promising and having the possibility to be generalized to more general sparse learning problems, DAL is difficult to interpret because it is derived as the augmented Lagrangian method [16, 17, 18] of the dual problem of Eq. (1) and the convergence of DAL is only known in an asymptotic sense.

In this paper we derive DAL algorithm from the proximal minimization framework [16] and based on that framework, analyze the convergence of DAL. We improve the general result of [16] in the case of DAL algorithm and show under certain conditions that DAL converges super-linearly in a global, non-asymptotic sense. One of the key assumptions in our analysis is the strong convexity of the dual problem of Eq. (1).

This paper is organized as follows. In Sec. 2, the DAL algorithm is derived from the proximal minimization framework. In Sec. 3 we theoretically analyze the convergence behaviour of DAL algorithm. The analysis is confirmed by empirical results in Sec. 4. Finally we summarize the paper in Sec. 5.

2 Proximal minimization view

2.1 Proximal minimization algorithm

Let us consider the following iterative algorithm called the proximal minimization algorithm [16] for the minimization of Eq. (1):

1. Choose some initial solution $\boldsymbol{w}^0$ and a sequence of non-decreasing positive numbers $\eta_0 \leq \eta_1 \leq \eta_2 \leq \cdots$.

2. Repeat until some criterion is satisfied:

$$\boldsymbol{w}^{t+1} = \mathop{\mathrm{argmin}}_{\boldsymbol{w}\in\mathbb{R}^n} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right), \tag{6}$$

   where $f(\boldsymbol{w})$ is the objective function in Eq. (1).

Although at this point it is not clear how we are going to carry out the above minimization, by definition we have $f(\boldsymbol{w}^{t+1}) + \frac{1}{2\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \leq f(\boldsymbol{w}^t)$; i.e., provided that the step-size is positive, the function value decreases monotonically at every iteration.

2.2 DAL algorithm from proximal minimization framework

IST approach can be considered to be linearly approximating the loss term in Eq. (6) at the *current point* $\boldsymbol{w}^t$ to perform the minimization (see Eq. (4)). In this paper we propose a technique to minimize Eq. (6) that uses a parametrized *linear lower bound* that can be adjusted to be the tightest at the *next point* $\boldsymbol{w}^{t+1}$, which precisely (to some finite precision) minimizes Eq. (6). Our approach is based on the convexity of f_ℓ and ϕ_λ. First note that these functions can be rewritten as follows:

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left(-\boldsymbol{\alpha}^\top \boldsymbol{A}\boldsymbol{w} - f_\ell^*(-\boldsymbol{\alpha}) \right), \quad \phi_\lambda(\boldsymbol{w}) = \max_{\boldsymbol{v}\in\mathbb{R}^n} \left(\boldsymbol{v}^\top \boldsymbol{w} - \phi_\lambda^*(\boldsymbol{v}) \right), \tag{7}$$

where f_ℓ^* and ϕ_λ^* are the convex conjugate functions of f_ℓ and ϕ_λ, respectively. Now we substitute these expression into Eq. (6) as follows:

$$\boldsymbol{w}^{t+1} = \mathop{\mathrm{argmin}}_{\boldsymbol{w}\in\mathbb{R}^n} \max_{\boldsymbol{\alpha}\in\mathbb{R}^m, \boldsymbol{v}\in\mathbb{R}^n} \left\{ (\boldsymbol{v} - \boldsymbol{A}^\top\boldsymbol{\alpha})^\top \boldsymbol{w} - f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right\}. \tag{8}$$

Note that now the loss term is expressed as a *linear* function (plus a quadratic term) as in the IST approach (see Eq. (4)). Now we exchange the order of minimization and maximization because the function to be minimaxed in Eq. (8) is a saddle function [19] (i.e., convex wrt $\boldsymbol{w}$ and concave wrt $(\boldsymbol{\alpha}, \boldsymbol{v})$), as follows:

$$\begin{aligned} \min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha},\boldsymbol{v}} \text{Eq. (8)} &= \max_{\boldsymbol{\alpha},\boldsymbol{v}} \min_{\boldsymbol{w}} \left\{ (\boldsymbol{v} - \boldsymbol{A}^\top\boldsymbol{\alpha})^\top \boldsymbol{w} - f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right\} & (9) \\ &= \max_{\boldsymbol{\alpha},\boldsymbol{v}} \left\{ -f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v}) + (\boldsymbol{v} - \boldsymbol{A}^\top\boldsymbol{\alpha})^\top \boldsymbol{w}^t - \frac{\eta_t}{2}\|\boldsymbol{v} - \boldsymbol{A}^\top\boldsymbol{\alpha}\|^2 \right\}. & (10) \end{aligned}$$

Moreover, the minimization with respect to $\boldsymbol{w}$ in Eq. (9) leads to an update equation $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \eta_t(\boldsymbol{A}^\top\boldsymbol{\alpha} - \boldsymbol{v})$. This is nothing but the dual augmented Lagrangian (DAL) algorithm proposed in [15] for ℓ_1-regularized minimization problems. After explicitly maximizing Eq. (10) with respect to $\boldsymbol{v}$ and removing it (see supplementary information), we obtain the following update equations:

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t), \tag{11}$$

where $\boldsymbol{\alpha}^t$ is an approximate minimizer of the *augmented Lagrangian function* $\varphi_t(\boldsymbol{\alpha})$ as follows:

$$\boldsymbol{\alpha}^t \simeq \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha}\in\mathbb{R}^m} \Bigg(\underbrace{f_\ell^*(-\boldsymbol{\alpha}) + \frac{1}{\eta_t}\Phi_{\lambda\eta_t}^*(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha})}_{=:\varphi_t(\boldsymbol{\alpha})} \Bigg). \tag{12}$$

In Eq. (12) the minimization is terminated by the following condition:

$$\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| \leq \epsilon_t \sqrt{\frac{\gamma}{\eta_t}} \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|, \tag{13}$$

where $\nabla\varphi_t(\boldsymbol{\alpha}^t)$ is the gradient of the augmented Lagrangian function at $\boldsymbol{\alpha}^t$ and γ is the modulus of strong convexity of f_ℓ^* (see Eq. (2)); ϵ_t is a positive sequence, which we specify in detail in the next section. The outer loop (Eq. (11)) can be terminated by monitoring the duality gap (see [15, 12]).

The function $\Phi_{\lambda\eta_t}^*$ in Eq. (12) is called the Moreau envelope of ϕ_λ^* ([20, 19]) and is defined as follows:

$$\Phi_\lambda^*(\boldsymbol{w}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} \left(\phi_\lambda^*(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{w}\|^2 \right). \tag{14}$$

For many "sparse" regularizers, the envelope function Φ^*_λ is easy to compute. For example if ϕ_λ is the support function[1] of some convex set C, ϕ^*_λ is the indicator function of C (see [19]), namely,

$$\phi^*_\lambda(\boldsymbol{v}) = \begin{cases} 0 & (\text{if } \boldsymbol{v} \in C), \\ +\infty & (\text{otherwise}). \end{cases}$$

Consequently, in this case we have (see supplementary information):

$$\Phi^*_\lambda(\boldsymbol{w}) = \frac{1}{2}\|\mathrm{ST}_\lambda(\boldsymbol{w})\|^2. \tag{15}$$

This is a continuously differentiable function. Moreover, it can be shown that the first and the second derivative of $\varphi_t(\boldsymbol{\alpha})$ with the above envelope function Φ^*_λ only depends on the columns of $\boldsymbol{A}$ that corresponds to non-zero elements of $\mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha})$ [15]. Thus when we are aiming for a sparse solution, the minimization of Eq. (12) can be performed efficiently; in fact when the Newton method is used the computational complexity for the minimization of $\varphi_t(\boldsymbol{\alpha})$ is roughly $\mathrm{O}(m^2 n^+)$, where m is the number of samples and n^+ is the number of non-zero elements of $\boldsymbol{w}^{t+1}$; quasi-Newton methods [18] can also be used if m is very large.

Note that the above update (Eqs. (11) and (12)) is very similar to the one in the IST approach (Eq. (5)). However, $-\boldsymbol{\alpha}$, which is the slope of the lower-bound of f_ℓ in Eq. (7) is optimized in Eq. (12) so that the lower-bound is the tightest at the *next point* $\boldsymbol{w}^{t+1}$. In fact, we can show $\nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1}) \simeq -\boldsymbol{\alpha}^t$ (see supplementary information).

The general connection between augmented Lagrangian methods and proximal minimization algorithms and (asymptotic) convergence results can be found in [16]. The derivation we show above is a special case when the objective function $f(\boldsymbol{w})$ can be split into a part that is easy to handle (regularization term $\phi_\lambda(\boldsymbol{w})$) and the rest (loss term $f_\ell(\boldsymbol{A}\boldsymbol{w})$).

3 Analysis

In this section we present two theorems that show super-linear convergence of DAL algorithm for the case the inner minimization (Eq. (12)) is carried out exactly ($\epsilon_t = 0$) and approximately ($\epsilon_t = 1$). Our result is inspired partly by [14] and is similar to the one given in [21] but is tighter and is valid for every iteration. The proofs of the theorems are given in the supplementary information due to lack of space.

Theorem 1. *Let $\boldsymbol{w}^*$ be the minimizer of Eq.* (1)*. We assume that there is a positive constant σ and a scalar α ($1 \leq \alpha \leq 2$) such that*

$$f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^\alpha, \qquad (t = 0, 1, 2, \ldots). \tag{16}$$

If the inner minimization is solved exactly (i.e., $\epsilon_t = 0$ in Eq. (13)*), we have the following convergence result:*

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| + \sigma\eta_t\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^{\alpha-1} \leq \|\boldsymbol{w}^t - \boldsymbol{w}^*\|.$$

Moreover, this implies that

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^{\frac{1+(\alpha-1)\sigma\eta_t}{1+\sigma\eta_t}} \leq \frac{1}{1+\sigma\eta_t}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|. \tag{17}$$

I.e., $\boldsymbol{w}^t$ converges to $\boldsymbol{w}^*$ super-linearly *if $\alpha < 2$ or $\alpha = 2$ and η_t is increasing, in a* global and non-asymptotic *sense.*

Theorem 2. *Let us assume that $\epsilon_t = 1$ in Eq.* (13) *and the same condition as in Theorem 1 (Eq.* (16)*). Then we have,*

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2 + 2\sigma\eta_t\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^\alpha \leq \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$

Moreover, this implies that

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^{\frac{1+\alpha\sigma\eta_t}{1+2\sigma\eta_t}} \leq \frac{1}{\sqrt{1+2\sigma\eta_t}}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|. \tag{18}$$

I.e., $\boldsymbol{w}^t$ converges to $\boldsymbol{w}^*$ super-linearly if $\alpha < 2$ or $\alpha = 2$ and η_t is increasing.

[1] Support function of a convex set C is defined as $\sigma_C(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in C} \boldsymbol{x}^\top \boldsymbol{y}$. For example, the ℓ_1-norm is the support function of the ℓ_∞ unit ball (see [19]). Group lasso, and the trace norm can also be considered as support functions of some convex sets.

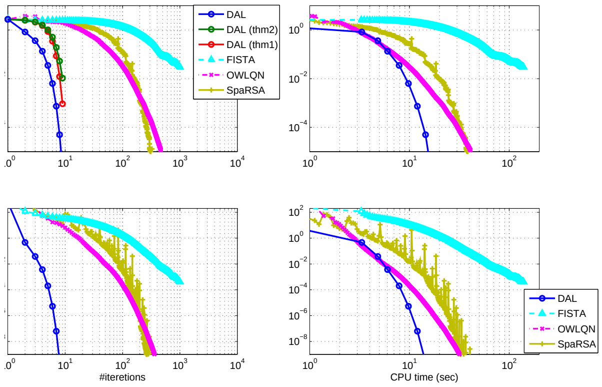

Figure 1: Empirical comparison of DAL [15], FISTA [14], OWLQN [6], and SpaRSA [12]. Top left: residual norm vs. number of iterations. Also the theoretical guarantees for DAL from Theorems 1 and 2 are shown. Top right: residual norm vs. CPU time. Bottom left: residual in the function value vs. number of iterations. Bottom right: residual in the function value vs. CPU time.

4 Empirical results

In this section we empirically confirm the validity of the convergence results obtained in the previous section and compare the efficiency of DAL [15], FISTA [14], OWLQN [6], and SpaRSA [12]. We randomly generated an ℓ_1-regularized logistic regression problem with $m = 1,024$ training examples and $n = 16,384$ features (see supplementary information for details). We use regularization constant $\lambda = 1$; $\boldsymbol{w}^0$ is an all zero vector and η^t is increased as $1, 2, 4, 8, \ldots$. First in order to obtain the true minimizer $\boldsymbol{w}^*$ of Eq. (1), we ran DAL algorithm to obtain a solution with high precision (relative duality gap$< 10^{-9}$). Assuming that the support of this solution is correct, we performed one Newton step of Eq. (1) in the subspace of active variables. The solution $\boldsymbol{w}^*$ we obtained in this way satisfied $\|\nabla f(\boldsymbol{w}^*)\| < 10^{-13}$, where $\nabla f(\boldsymbol{w}^*)$ is the minimum norm subgradient of f at $\boldsymbol{w}^*$. The parameter σ in Eq. (16) was estimated at $\boldsymbol{w}^*$ from the minimum eigenvalue of the Hessian of the loss term in Eq. (1) in the subspace of active variables. The modulus of strong convexity $\gamma = 4$ from simple calculation (see Eq. (2)). We used Eqs. (17) and (18) with $\alpha = 2$ and the initial residual $\|\boldsymbol{w}^0 - \boldsymbol{w}^*\|$ to generate sequences of predicted residual norm $\|\widehat{\boldsymbol{w}^t - \boldsymbol{w}^*}\|$ (red and green curves in Fig. 1, respectively). All algorithms were implemented in MATLAB and run on a LINUX workstation with two dual-core 3.3 GHz Xeon processors and 8 GB of RAM. We used preconditioned conjugate gradient method for solving the Newton system that arise in the minimization of Eq. (12).

In the top left panel of Fig. 1, we can see that the convergence in terms of the norm of the residual vector $\boldsymbol{w}^t - \boldsymbol{w}^*$ happens indeed rapidly as predicted by the theory in Sec. 3. The red curve shows the result of Theorem 1, which assumes exact minimization of Eq. (12), and the green curve shows the result of Theorem 2, which allows some error in the minimization of Eq. (12). We can see that the difference between the optimistic analysis of Theorem 1 and the realistic analysis of Theorem 2 is negligible. In this problem, in order to reach the quality of solution DAL achieves in 10 iterations OWLQN and SpaRSA take at least 100 iterations and FISTA takes 1,000 iterations.

The bottom left panel of Fig. 1 shows comparison of four algorithms DAL, FISTA, OWLQN, and SpaRSA in terms of the decrease in the function value. The convergence of DAL is the fastest also in terms of function value. OWLQN and SpaRSA are the next after DAL and are faster than FISTA.

DAL needs to solve a minimization problem at every iteration, which is heavier than the operation required in the three earlier studies. Thus we compare the total CPU time spent by the algorithms in the right part of Fig. 1. It can be seen that DAL can obtain a solution that is much more accurate in about 10 seconds than the solution FISTA obtained after 100 seconds. In terms of computation time, DAL and OWLQN seem to be on par at low precision. However as the precision becomes higher DAL becomes clearly faster than OWLQN. SpaRSA seems to be slightly slower than DAL and OWLQN.

5 Discussion

In this paper, We have presented a new view on DAL algorithm [15] for sparsity-regularized minimization problems; the new interpretation is based on the proximal minimization framework [16]. Generalizing the recent result from [14] we improved the general result on super-liner convergence of augmented Lagrangian methods in [16]. In both the noiseless case (Theorem 1) and the noisy case (Theorem 2), the new results hold without the assumption that η_t is sufficiently large. The strong convexity in the dual is a natural assumption for many loss functions including the ones that are not strongly convex in the primal (e.g., logistic loss). This makes DAL approach a promising direction in machine learning applications of sparse methods. The theoretically predicted rapid convergence of DAL algorithm is also empirically confirmed in a simulated ℓ_1-regularized logistic regression problem.

References

[1] I. F. Gorodnitsky and B. D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm", *IEEE Trans. Signal Process.*, 45(3), 1997.

[2] J. Bioucas-Dias, "Bayesian wavelet-based image deconvolution: A GEM algorithm exploiting a class of heavy-tailed priors", *IEEE Trans. Image Process.*, 15: 937–951, 2006.

[3] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-Minimization Algorithm for Wavelet-Based Image Restoration", *IEEE Trans. Image Process.*, 16(12), 2007.

[4] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-Task Feature Learning", in: B. Schölkopf, J. Platt, and T. Hoffman, eds., *Advances in NIPS 19*, 41–48, MIT Press, Cambridge, MA, 2007.

[5] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL", *JMLR*, 9: 2491–2521, 2008.

[6] G. Andrew and J. Gao, "Scalable training of L1-regularized log-linear models", in: *Proc. of the 24th international conference on Machine learning*, 33–40, ACM, New York, NY, USA, 2007.

[7] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration", *IEEE Trans. Image Process.*, 12: 906–916, 2003.

[8] I. Daubechies, M. Defrise, and C. D. Mol, "An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint", *Commun. Pur. Appl. Math.*, LVII: 1413–1457, 2004.

[9] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting", *Multiscale Modeling and Simulation*, 4(4): 1168–1200, 2005.

[10] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables", *J. Roy. Stat. Soc. B*, 68(1): 49–67, 2006.

[11] M. Fazel, H. Hindi, and S. P. Boyd, "A Rank Minimization Heuristic with Application to Minimum Order System Approximation", in: *Proc. of the American Control Conference*, 2001.

[12] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse Reconstruction by Separable Approximation", *IEEE Trans. Signal Process.*, 2009.

[13] J. Bioucas-Dias and M. Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration", *IEEE Trans. Image Process.*, 16(12): 2992–3004, 2007.

[14] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems", *SIAM J. Imaging Sciences*, 2(1): 183–202, 2009.

[15] R. Tomioka and M. Sugiyama, "Dual Augmented Lagrangian Method for Efficient Sparse Reconstruction", *IEEE Signal Processing Letters*, 16(12): 1067–1070, 2009.

[16] R. T. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming", *Math. of Oper. Res.*, 1: 97–116, 1976.

[17] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, 1982.

[18] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, 1999.

[19] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[20] J. J. Moreau, "Proximité et dualité dans un espace hilbertien", *Bulletin de la S. M. F.*, 93: 273–299, 1965.

[21] R. Rockafellar, "Monotone operators and the proximal point algorithm", *SIAM Journal on Control and Optimization*, 14: 877–898, 1976.

Supplementary Information for “Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparse Learning”

A Preliminaries on proximal operation

This section contains basic results on proximal operation, which we use in the next section and is based on [20, 19, 9].

Let f be a closed proper convex function over $\mathbb{R}^n$ that takes values in $\mathbb{R} \cup \{+\infty\}$. The *proximal operator* with respect to f is defined as follows:

$$\mathrm{prox}_f(\boldsymbol{z}) = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \left(f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|^2 \right).$$

Moreover, the minimum attained above is called the Moreau envelope of f.

$$F(\boldsymbol{z}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} \left(f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|^2 \right). \tag{A.1}$$

Proximal operation can be considered as a generalization of the projection on a convex set. In fact, if we take f as the indicator function of the ℓ_∞ unit-ball, i.e., $f(\boldsymbol{z}) = \delta_\infty(\boldsymbol{z})$, where

$$\delta_\infty(\boldsymbol{z}) = \begin{cases} 0 & (\text{if } z_j \leq 1 \text{ for all } j = 1, \ldots, n), \\ +\infty & (\text{otherwise}), \end{cases} \tag{A.2}$$

then the proximal operation with respect to δ_∞ is the projection on the unit ball as follows:

$$\left(\mathrm{prox}_{\delta_\infty}(\boldsymbol{z})\right)_j = \min\left(\frac{1}{|z_j|}, 1\right) z_j. \tag{A.3}$$

If $f(\boldsymbol{z}) = \sum_{j=1}^n |z_j|$, the proximal operation with respect to f is the soft-thresholding operation in Eq. (3):

$$\left(\mathrm{prox}_{\|\cdot\|_1}(\boldsymbol{z})\right)_j = \max\left(\frac{|z_j|-1}{|z_j|}, 0\right) z_j. \tag{A.4}$$

Comparing Eqs. (A.3) and (A.4) we notice that

$$\mathrm{prox}_{\delta_\infty}(\boldsymbol{z}) + \mathrm{prox}_{\|\cdot\|_1}(\boldsymbol{z}) = \boldsymbol{z}.$$

This is because the indicator function δ_∞ and the ℓ_1-norm function $\|\cdot\|_1$ are conjugate to each other. This relation holds indeed for general closed proper convex functions and can be stated as follows:

$$\mathrm{prox}_f(\boldsymbol{z}) + \mathrm{prox}_{f^*}(\boldsymbol{z}) = \boldsymbol{z}, \tag{A.5}$$

where f^* is the convex conjugate function of f. The proof can be found in [20].

Equation (15) is implied from Eq. (A.5) if f is a support function of some convex set C, i.e., $f(\boldsymbol{x}) = \sup_{\boldsymbol{y}\in C} \boldsymbol{x}^\top \boldsymbol{y}$. In fact,

$$\begin{aligned} F^*(\boldsymbol{z}) &= \min_{\boldsymbol{x}\in\mathbb{R}^n} \left(f^*(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|^2 \right) \\ &= f^*(\mathrm{prox}_{f^*}(\boldsymbol{z})) + \frac{1}{2}\|\mathrm{prox}_{f^*}(\boldsymbol{z}) - \boldsymbol{z}\|^2 \\ &= \frac{1}{2}\|\mathrm{prox}_f(\boldsymbol{z})\|^2. \end{aligned}$$

where we used $f^*(\mathrm{prox}_{f^*}(\boldsymbol{z})) = 0$ because f^* is the indicator function of the set C (as in Eq. (A.2)).

Another important property is related to the derivative of a Moreau envelope. A Moreau envelope $F(\boldsymbol{z})$ (see Eq. (A.1)) is always differentiable and its derivative is given as follows:

$$\nabla F(\boldsymbol{z}) = \mathrm{prox}_{f^*}(\boldsymbol{z}). \tag{A.6}$$

The proof can be found in [20].

B Derivation of Equations (11)-(14)

Proof. After pushing the maximization with respect to $\boldsymbol{v}$ in Eq. (10) inside and completing the square we have,

$$\begin{aligned}
\boldsymbol{\alpha}^t &\simeq \underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\operatorname{argmax}} \left\{ -f_\ell^*(-\boldsymbol{\alpha}) - \min_{\boldsymbol{v}\in\mathbb{R}^n} \left(\phi_\lambda^*(\boldsymbol{v}) + \frac{\eta_t}{2}\|\boldsymbol{v} - \boldsymbol{A}^\top\boldsymbol{\alpha} - \boldsymbol{w}^t/\eta_t\|^2 \right) + \frac{\|\boldsymbol{w}^t\|^2}{2\eta_t} \right\} \\
&= \underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\operatorname{argmax}} \left\{ -f_\ell^*(-\boldsymbol{\alpha}) - \min_{\boldsymbol{v}\in\mathbb{R}^n} \left(\phi_\lambda^*(\boldsymbol{v}) + \frac{\eta_t}{2}\|\boldsymbol{v} - \boldsymbol{A}^\top\boldsymbol{\alpha} - \boldsymbol{w}^t/\eta_t\|^2 \right) \right\} \qquad \text{(omit the constant term)} \\
&= \underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\operatorname{argmax}} \left\{ -f_\ell^*(-\boldsymbol{\alpha}) - \frac{1}{\eta_t} \min_{\tilde{\boldsymbol{v}}\in\mathbb{R}^n} \left(\phi_{\lambda\eta_t}^*(\tilde{\boldsymbol{v}}) + \frac{1}{2}\|\tilde{\boldsymbol{v}} - (\eta_t\boldsymbol{A}^\top\boldsymbol{\alpha} + \boldsymbol{w}^t)\|^2 \right) \right\} \qquad (\tilde{\boldsymbol{v}} = \eta_t\boldsymbol{v}) \\
&= \underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\operatorname{argmax}} \left(-f_\ell^*(-\boldsymbol{\alpha}) - \frac{1}{\eta_t}\Phi_{\lambda\eta_t}^*(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}) \right),
\end{aligned}$$

We used $\phi_{\lambda\eta_t}^*(\boldsymbol{v}) = (\eta_t\phi_\lambda)^*(\boldsymbol{v}) = \eta_t\phi_\lambda^*(\boldsymbol{v}/\eta_t)$ in the third line. Furthermore, the above minimum is attained at:

$$\tilde{\boldsymbol{v}} = \eta_t\boldsymbol{v} = \operatorname{prox}_{\phi_{\lambda\eta_t}^*}\left(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}\right).$$

Combining this expression with the update equation (see Eqs. (9) and (10))

$$\begin{aligned}
\boldsymbol{w}^{t+1} &= \boldsymbol{w}^t + \eta_t(\boldsymbol{A}^\top\boldsymbol{\alpha}^t - \boldsymbol{v}^t) \\
&= \left(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}^t\right) - \operatorname{prox}_{\phi_{\lambda\eta_t}^*}\left(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}^t\right) \\
&= \operatorname{prox}_{\phi_{\lambda\eta_t}}\left(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}^t\right) \qquad \text{(due to Eq. (A.5))} \\
&= \operatorname{ST}_{\lambda\eta_t}\left(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}^t\right).
\end{aligned}$$

This completes the derivation. □

C Proof of Theorem 1

Proof. The first step of the proof is a generalization of Lemma 2.3 in [14]. First we show that:

$$(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t \in \partial f(\boldsymbol{w}^{t+1}) = \boldsymbol{A}^\top\nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1}) + \partial\phi_\lambda(\boldsymbol{w}^{t+1}). \tag{C.1}$$

In fact, because $\boldsymbol{\alpha}^t$ minimizes $\varphi_t(\boldsymbol{\alpha})$ exactly ($\epsilon_t = 0$) we have,

$$\nabla\varphi_t(\boldsymbol{\alpha}^t) = -\nabla f_\ell^*(-\boldsymbol{\alpha}^t) + \boldsymbol{A}\boldsymbol{w}^{t+1} = 0,$$

where the derivative of the envelope function $\Phi_{\lambda\eta}^*(\boldsymbol{w})$ (see Eq. (14)) is computed using Eq. (A.6) with $\operatorname{prox}_{\phi_{\lambda\eta_t}}(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}) = \operatorname{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}) = \boldsymbol{w}^{t+1}$. Thus,

$$\nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1}) = \nabla f_\ell(\nabla f_\ell^*(-\boldsymbol{\alpha}^t)) = -\boldsymbol{\alpha}^t. \tag{C.2}$$

Additionally, from the definition of soft-thresholding operation and $\phi_{\lambda\eta_t} = \eta_t\phi_\lambda$, we have,

$$\frac{\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}^t - \boldsymbol{w}^{t+1}}{\eta_t} \in \partial\phi_\lambda(\boldsymbol{w}^{t+1}). \tag{C.3}$$

Combining Eqs. (C.2) and (C.3) we obtain Eq. (C.1).

Next due to the convexity of $f(\boldsymbol{w})$ we have,

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle \boldsymbol{w}^* - \boldsymbol{w}^{t+1}, (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t \right\rangle, \tag{C.4}$$

because $(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t \in f(\boldsymbol{w}^{t+1})$ (see Eq. (C.1)). Rewriting the above inequality, we obtin,

$$\begin{aligned}
\eta_t(f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1})) &\geq \left\langle \boldsymbol{w}^* - \boldsymbol{w}^{t+1}, \boldsymbol{w}^t - \boldsymbol{w}^* + \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle \\
&= \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 - \left\langle \boldsymbol{w}^* - \boldsymbol{w}^{t+1}, \boldsymbol{w}^t - \boldsymbol{w}^* \right\rangle \\
&\geq \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 - \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\| \|\boldsymbol{w}^t - \boldsymbol{w}^*\| \\
&\geq \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 - \left(\frac{\mu}{2} \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + \frac{1}{2\mu} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2 \right) \quad (\forall \mu > 0) \\
&= \left(1 - \frac{\mu}{2}\right) \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2 - \frac{1}{2\mu} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2. \qquad (\star)
\end{aligned}$$

Note that by setting $\mu = 1$ in $(\star)$ we can recover (a special case of) Lemma 2.3 in [14]. Now using the assumption (16) we obtain the following expression:

$$\left(2\mu - \mu^2\right) \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2 + 2\mu\sigma\eta_t \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^\alpha \leq \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$

Maximizing the left hand side with respect to μ, we have $\mu = 1 + \sigma\eta_t \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^{\alpha-2}$ and accordingly,

$$\left(1 + \sigma\eta_t \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^{\alpha-2}\right)^2 \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2 \leq \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$

Taking the square-root of both sides we obtain,

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| + \sigma\eta_t \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^{\alpha-1} \leq \|\boldsymbol{w}^t - \boldsymbol{w}^*\|. \tag{C.5}$$

The last part of the theorem is obtained by lower-bounding the lhs of the above inequality. Let $b_{t+1} = \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|$. The lhs of the above inequality can be rewritten as follows:

$$\begin{aligned}
b_{t+1} + \sigma\eta_t b_{t+1}^{\alpha-1} &= (1 + \sigma\eta_t) \left(\frac{1}{1 + \sigma\eta_t} b_{t+1} + \frac{\sigma\eta_t}{1 + \sigma\eta_t} b_{t+1}^{\alpha-1} \right) \\
&\geq (1 + \sigma\eta_t) b_{t+1}^{\frac{1}{1+\sigma\eta_t}} \cdot b_{t+1}^{\frac{(\alpha-1)\sigma\eta_t}{1+\sigma\eta_t}} \\
&= (1 + \sigma\eta_t) b_{t+1}^{\frac{1+(\alpha-1)\sigma\eta_t}{1+\sigma\eta_t}},
\end{aligned}$$

where we used Young's inequality to obtain the second line. Substituting this relation back into Eq. (C.5) completes the proof of the theorem. □

D Proof of Theorem 2

Proof. First let us define $\boldsymbol{\delta}^t \in \mathbb{R}^m$ as follows:

$$\boldsymbol{\delta}^t := \nabla\varphi_t(\boldsymbol{\alpha}^t) = -\nabla f_\ell^*(-\boldsymbol{\alpha}^t) + \boldsymbol{A}\boldsymbol{w}^{t+1},$$

where $\|\boldsymbol{\delta}^t\| \leq \sqrt{\frac{\gamma}{\eta_t}} \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|$ from Eq. (13) with $\epsilon_t = 1$. Therefore we obtain an analogue of Eq. (C.2) in the previous section as follows:

$$-\boldsymbol{\alpha}^t = \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1} - \boldsymbol{\delta}^t). \tag{D.1}$$

Next, we note that the strong duality of the conjugate loss function f_ℓ^* implies that the Hessian of the loss function f_ℓ is uniformly bounded from above, i.e., for any $\boldsymbol{z}, \boldsymbol{z}_0 \in \mathbb{R}^m$,

$$f_\ell(\boldsymbol{z}) \leq f_\ell(\boldsymbol{z}_0) + \langle \boldsymbol{z} - \boldsymbol{z}_0, \nabla f_\ell(\boldsymbol{z}_0) \rangle + \frac{1}{2\gamma} \|\boldsymbol{z} - \boldsymbol{z}_0\|^2. \tag{D.2}$$

Moreover, if $\boldsymbol{w} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{y})$,

$$(\boldsymbol{y} - \boldsymbol{w})/\eta_t \in \partial\phi_\lambda(\boldsymbol{w}), \tag{D.3}$$

because $\partial\phi_{\lambda\eta_t}(\boldsymbol{w}) + (\boldsymbol{w} - \boldsymbol{y}) \ni 0$ and $\phi_{\lambda\eta_t} = \eta_t\phi_\lambda$.

Now we are ready to derive an analogue of Eq. (C.4) as follows:

$$\begin{aligned}\eta_t(f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1})) = \eta_t(&\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}^*) - f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1} - \boldsymbol{\delta}^t)}_{\text{(A)}})\\ &+ \eta_t(\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1} - \boldsymbol{\delta}^t) - f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1})}_{\text{(B)}})\\ &+ \eta_t(\underbrace{\phi_\lambda(\boldsymbol{w}^*) - \phi_\lambda(\boldsymbol{w}^{t+1})}_{\text{(C)}}).\end{aligned}$$

The above terms (A), (B), and (C) can be separately bounded using the convexity of f_ℓ and ϕ_λ as follows:

$$\text{(A)}: \quad f_\ell(\boldsymbol{A}\boldsymbol{w}^*) - f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1} - \boldsymbol{\delta}^t) \geq \left\langle \boldsymbol{A}(\boldsymbol{w}^* - \boldsymbol{w}^{t+1}) + \boldsymbol{\delta}^t, -\boldsymbol{\alpha}^t \right\rangle \qquad \text{(due to Eq. (D.1))}$$

$$\text{(B)}: \quad f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1} - \boldsymbol{\delta}^t) - f_\ell(\boldsymbol{A}\boldsymbol{w}^{t+1}) \geq -\left\langle \boldsymbol{\delta}^t, -\boldsymbol{\alpha}^t \right\rangle - \frac{1}{2\gamma}\|\boldsymbol{\delta}^t\|^2 \qquad \text{(due to Eq. (D.2))}$$

$$\text{(C)}: \quad \phi_\lambda(\boldsymbol{w}^*) - \phi_\lambda(\boldsymbol{w}^{t+1}) \geq \left\langle \frac{\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t - \boldsymbol{w}^{t+1}}{\eta_t}, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle \qquad \text{(due to Eq. (D.3))}$$

The last inequality is because $\boldsymbol{w}^{t+1} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t)$. Combining (A), (B), and (C), we have the following expression:

$$\eta_t(f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1})) \geq \left\langle \boldsymbol{w}^t - \boldsymbol{w}^{t+1}, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle - \frac{\eta_t}{2\gamma}\|\boldsymbol{\delta}^t\|^2.$$

Note that the above inequality is identical to Eq. (C.4) except for the last term. Using Eq. (13) and after some manipulations similar to the proof of Theorem 1, we obtain,

$$\begin{aligned}\eta_t(f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1})) &\geq \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + \left\langle \boldsymbol{w}^t - \boldsymbol{w}^*, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle - \frac{1}{2}\|\boldsymbol{w}^t - \boldsymbol{w}^{t+1}\|^2\\ &= \frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 - \frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}^t\|^2\end{aligned}$$

Finally using Eq. (16) we obtain,

$$\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + 2\sigma\eta_t\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^\alpha \leq \|\boldsymbol{w}^* - \boldsymbol{w}^t\|^2.$$

The last part of the proof is identical to that of Theorem 1. □

E Generation of a random ℓ_1-regularized logistic regression problem

The elements of the design matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ was randomly sampled from the standard normal distribution and the class label $(y_i)_{i=1}^n$ was generated as $y_i = \mathrm{sign}(\boldsymbol{A}_i\boldsymbol{\beta} + 0.01\xi_i)$ $(i = 1, \ldots, m)$ where $\boldsymbol{A}_i$ is the i-th row vector of $\boldsymbol{A}$, and ξ_i is sampled independently and identically from the standard normal distribution. The true regression coefficient vector $\boldsymbol{\beta}$ was randomly generated by filling roughly 4% of its components by either $+1$ or -1 at random; the remaining elements of $\boldsymbol{\beta}$ were set to zero.

The regularization constant $\lambda = 1$ was chosen to reproduce the same level of sparsity as the true $\boldsymbol{\beta}$; the solution had about 4.7% of its elements filled.

The whole procedure was repeated ten times but since the variation was negligibly small, we show results from a single experiment, because the theoretical guarantee also depends on the data.