Variable Screening for Sparse Online Regression

Jingwei Liang[*] Clarice Poon[†‡]

March 11, 2021

Abstract. Sparsity-promoting regularizers are widely used to impose low-complexity structure (e.g. ℓ_1-norm for sparsity) to the regression coefficients of supervised learning. In the realm of deterministic optimization, the sequence generated by iterative algorithms (such as proximal gradient descent) exhibit "finite activity identification", namely, they can identify the low-complexity structure in a finite number of iterations. However, many online algorithms (such as proximal stochastic gradient descent) do not have this property owing to the vanishing step-size and non-vanishing variance. In this paper, by combining with a screening rule, we show how to eliminate useless features of the iterates generated by online algorithms, and thereby enforce finite activity identification. One advantage of our scheme is that when combined with any convergent online algorithm, sparsity properties imposed by the regularizer can be exploited to improve computational efficiency. Numerically, significant acceleration can be obtained.

1 Introduction

1.1 Background

Regression problems play a fundamental role in various fields including machine learning, data science and statistics and moreover, sparse regularisation, such as ℓ_1 regularisation [29] has been increasingly popular in recent years. In this paper, we combine variable screening techniques with online algorithms to solve sparsity-promoting regression problems of the following form

$$\min_{\beta\in\mathbb{R}^n} F(\beta) + \lambda\Omega(\beta), \quad \text{where} \quad F(\beta) \overset{\text{def}}{=} \mathbb{E}_{(x,y)}[f(x^\top\beta; y)]. \tag{$\mathcal{P}_\lambda$}$$

The expectation is taken over random variable (x, y) whose probability distribution Λ is supported on some compact domain $\mathcal{X}\times\mathcal{Y}\subset\mathbb{R}^n\times\mathbb{R}$ and $\lambda > 0$ is a trade-off parameter to balance the loss F and the sparsity promoting regularizer Ω.

Popular choices of the loss function f include the squared loss, logistic loss and the squared hinge loss, while popular choices for Ω include the ℓ_1-norm for enforcing sparsity [29], the $\ell_{1,2}$-norm for enforcing group sparsity [37] and the $\ell_1+\ell_{1,2}$-norms for enforcing sparsity within groups [27].

We assume throughout that:

(H.1) There exists $L > 0$ such that for each y, $f_y \overset{\text{def}}{=} f(\cdot; y) : \mathbb{R}\to\mathbb{R}$ is convex, differentiable, and has L-Lipschitz continuous gradient.

[*] School of Mathematical Sciences, Queen Mary University of London, London UK. E-mail: jl993@cam.ac.uk.

[†] Department of Mathematical Sciences, University of Bath, Bath UK. E-mail: cmhsp20@bath.ac.uk.

[‡] Corresponding author

(H.2) The regularization function $\Omega : \mathbb{R}^n \to [0, \infty)$ is a convex and group decomposable norm, that is, given $\beta \in \mathbb{R}^n$ and a partition $\mathcal{G}$ on $\{1, \ldots, n\}$ such that $\beta = (\beta_g)_{g\in\mathcal{G}}$, we have

$$\Omega(\beta) \overset{\text{def}}{=} \sum\nolimits_{g\in\mathcal{G}} \Omega_g(\beta_g)$$

where Ω_g is a norm on $\mathbb{R}^{n_g}$ with n_g being the cardinality of β_g.

Empirical loss minimization In practice, rather than minimizing the loss function F over the distribution Λ, one draws samples from Λ and deal with the *empirical loss*

$$F_\eta(\beta) \overset{\text{def}}{=} \sum\nolimits_{i=1}^{m} \eta_i f(x_i^\top \beta, y_i)$$

where m samples $\{x_i, y_i\}_{i=1}^m \in (\mathbb{R}^n)^m \times \mathbb{R}^m$ are drawn from Λ, with positive weights η_i which sum to 1. A popular choice of $(\eta_i)_i$ is uniform weights, *i.e.* $\eta_i \equiv \frac{1}{m}$. Correspondingly, $(\mathcal{P}_\lambda)$ becomes the following regularized empirical loss

$$\min_{\beta\in\mathbb{R}^n} \left\{ P_{\lambda,\eta}(\beta) = \lambda\Omega(\beta) + \sum\nolimits_{i=1}^{m} \eta_i f(x_i^\top \beta, y_i) \right\} \tag{1.1}$$

1.1.1 Dimension reduction via (safe) screening

The purpose of using sparsity-promoting regularizers is so that the solution of the optimization problem (1.1) has as few non-zeros coefficients as possible. In high dimensional statistics, (safe) screening techniques are popular approaches for filtering out features whose corresponding coefficients are 0, hence achieving dimension reduction; See [8, 30, 22] and the references therein. Safe feature elimination was first proposed by El Ghaoui et al. [8] for ℓ_1 regularization problems. The rules introduced were static rules, where features are screened out as a preprocessing step, and sequential rules where one solves a sequence of optimization problems with a decreasing list of parameters $\{\lambda_k\}_k$, so that solutions of an optimization problem with λ_{k+1} are used to screen out features when solving with λ_k. Since this work, several extensions to these rules have been proposed [18, 32, 35].

Dynamic screening rules were later proposed by [3], where the safe screening region are updated along the iterates of a solver. Another work in this direction are so-called gap-safe rules [20, 22] where the calculation of the safe regions along the iterates are done via primal-dual gap computations. The present article is largely inspired by [22], where we dynamically construct safe regions by computing an 'online' primal-dual gap.

1.2 Our contributions

Although screening techniques are algorithm agnostic, they have been investigated mostly in the context of deterministic or batched algorithms (where one evaluates the full gradient ∇F or ∇F_η at each iteration). For large-scale problems, batched optimization (e.g. proximal gradient descent) may be impractical and it is preferable to use online algorithms [4]: at each iteration t, draw a sample (x_t, y_t) randomly from the distribution Λ and perform the update

$$\beta_{t+1} = \beta_t - \gamma_t\big(f'_{y_t}(x_t^\top \beta_t)x_t + \lambda Z_t\big) \tag{1.2}$$

where $Z_t \in \partial\Omega(\beta_t)$ is a subgradient (see Eq. (A.1)). This is a special instance of stochastic gradient descent, which can be traced back to [25, 11].

To deal with the non-smoothness imposed by the regularization term Ω, various stochastic algorithms have been proposed in the literature, such as *truncated gradient* [13] or stochastic versions of proximal gradient descent [7] (Prox-SGD)

$$\beta_{t+1} = \text{prox}_{\lambda\gamma_t\Omega}\big(\beta_t - \gamma_t f'_{y_t}(x_t^\top \beta_t)x_t\big),$$

where $\text{prox}_{\lambda\gamma_t\Omega}(\cdot)$ is called the *proximal operator* which is defined by $\text{prox}_{\gamma\Omega}(\cdot) = \text{argmin}_\beta \gamma\Omega(\beta) + \frac{1}{2}\|\beta - \cdot\|^2$, and has closed form expressions for the aforementioned sparsity promoting regularization terms [5]. However, for the standard Prox-SGD, due to the vanishing step-size γ_t and non-vanishing variance in the stochastic gradient estimates, the generated sequence $\{\beta_t\}_{t\in\mathbb{N}}$ tend to have full support for all t even though the sought after solution is sparse [36, 14]; see [23] for an explicit example. As a result, one cannot easily exploit the sparsity-promoting structure of Ω for computational gains.

In this paper, we address the non-sparseness problem (*i.e.* no support identification) of online optimization algorithms by combining them with the idea of safe screening rules [8, 22]. More precisely, our contributions are as follows.

- By adapting gap-safe screening rules of [22] to online algorithms, we propose an online-screening rule. The proposed rule only needs to evaluate function values at the sampled data, hence has low per iteration complexity. In particular, we show how to construct a "dual certificate" along the iterations which allows us to apply gap-safe rules to screen out certain features. Moreover, this certificate can be built alongside any convergent online algorithms.
- The consequence of screening rules for online optimization is support identification of β_t, *i.e.* dimension reduction. It allows us to locate exactly which features are of interests. More importantly, significant computational gains can be obtained since the per iteration complexity scales from n the dimension of the variable to κ the sparsity of the solution.

Remark 1.1. An interesting feature of many batched optimization methods such as proximal gradient descent [16, 15, 17] and coordinate descent [12], is that they exhibit "finite activity identification", where after a finite number of iterations, all iterates will have the same sparsity structure as the solution. Although one cannot check a-priori whether activity identification has been achieved, one can heuristically still exploit this property for computational gains by switching to higher order methods once the support is sufficiently small.

One the other hand, while activity identification is not present in many online algorithms (with the exception of regularized dual averaging [36]), we make use of screening in this work to enforce the identification property. This idea is inspired by the recent work [28], where the authors developed a gap-safe rule for conditional gradient descent. One highlight of their work is that through safe screening, identification is achieved whereas simply running conditional gradient descent will never achieve identification. In this work, we draw on this idea from [28] to enforce identification for online algorithms. We extend the gap-safe rules developed in [22] to the stochastic setting, where we dynamically construct safe regions by computing an "online" primal-dual gap.

Paper organization The rest of the paper is organized as follows. We recall the basic derivation of screening rules for sparsity-promoting regression problems in Section 2. Theoretical analysis of our online-screening rule is presented in Section 3. Numerical experiments on LASSO and sparse logistic regression problems are provided in Section 4. Finally, in the appendix we collect some basic concepts of convex analysis and the proofs of the main theorems.

2 Safe screening

Before introducing our algorithm, we first provide some background on screening rules, with particular focus on the gap-safe rule from [22].

Given a sparsity-promoting norm Ω, its dual norm is defined by

$$\Omega^D(x) \overset{\text{def}}{=} \sup_{\Omega(z)\leq 1} \langle z,\, x\rangle$$

and its sub-differential can be expressed as

$$\partial\Omega(\beta) = \big\{Z :\; \langle Z,\, \beta\rangle = \Omega(\beta), \quad \Omega^D(Z) \leq 1\big\}.$$

Note that if Ω is group decomposable, then we have $\Omega^D(x) = \sup_{g\in\mathcal{G}} \Omega_g^D(x)$.

2.1 Safe screening

We summarise here a few key facts about the support of solutions to (1.1), and refer to [9, 31] for further details. Let $\beta^\star$ be a global minimizer of the regularized empirical loss minimization (1.1), the first-order optimality condition is

$$0 \in \nabla F_\eta(\beta^\star) + \lambda\partial\Omega(\beta^\star). \tag{2.1}$$

This is equivalent to saying that $\beta^\star$ is a minimizer if and only if

$$Z^\star \overset{\text{def}}{=} -\frac{1}{\lambda}\sum_{i=1}^m \eta_i\theta_i^\star x_i \in \partial\Omega(\beta^\star), \quad \text{where} \quad \theta_i^\star = \nabla f_i(x_i^\top\beta^\star),\;\; i = 1, ..., m. \tag{2.2}$$

The fundamental idea behind screening rule comes from the optimality condition of the minimizers. Given any group $g \in \mathcal{G}$,

$$\Omega_g^D(Z_g^\star) < 1 \quad \Longrightarrow \quad \beta_g^\star = 0. \tag{2.3}$$

The converse is also true under the non degeneracy condition $0 \in \mathrm{ri}\big(\nabla F_\eta(\beta^\star) + \lambda\partial\Omega(\beta^\star)\big)$, where ri denotes the relative interior [9, 31]. In the case where $\Omega(\beta) = \|\beta\|_1$, this means that $\lambda^{-1}\nabla F_\eta(\beta^\star)$ takes absolute value ±1 only on the support of β^*. Moreover, $\theta^\star$ in $Z^\star$ is precisely the solution of the *dual problem* of (1.1)

$$\max_{\theta\in\mathcal{K}_{\lambda,\eta}} \Big\{D_{\lambda,\eta}(\theta) \overset{\text{def}}{=} -\sum_{i=1}^m \eta_i f_{y_i}^*(\theta_i)\Big\} \tag{2.4}$$

where

$$\mathcal{K}_{\lambda,\eta} \overset{\text{def}}{=} \big\{\theta :\; \Omega^D\big(\textstyle\sum_{i=1}^m \theta_i\eta_i x_i\big) \leq \lambda\big\} \subset \mathbb{R}^m$$

is the dual constraint set. Since the vector $Z^\star$ certifies the support of $\beta^\star$, it is called the *dual certificate.*

The above message implies that, if $Z^\star$ is known, then we can identify an index set which includes the support of the solution:

$$\mathcal{I} \overset{\text{def}}{=} \big\{g \in \mathcal{G} :\; \Omega_g^D(Z^\star) = 1\big\} \supseteq \mathrm{supp}(\beta^\star).$$

Consequently, one can restrict to optimization over $\beta_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ instead. This can lead to huge computation gains if $\mathcal{I}$ tightly estimates the true support $\mathrm{supp}(\beta^\star)$ which very often is much smaller than the dimension of the problem.

While computing $Z^\star$ is generally as difficult as finding $\beta^\star$, the positions where $Z^\star$ saturates can be estimated more readily. This is exactly the idea of safe screening, which constructs a "safe region" $\mathcal{Z}$ such that $Z^\star \in \mathcal{Z}$. Then instead of using (2.3) to determine the zero entries of $\beta^\star$, one can consider the relaxed criteria: $\beta^\star_g = 0$ if $\sup_{Z\in\mathcal{Z}} \Omega^D_g(Z) < 1$. The following result, which can be found in [28], illustrates how to perform screening rules based on a safe region $\mathcal{Z}$ with center c.

Proposition 2.1 (Safe screen rule). *Let $\beta^\star$ be a minimizer to* (1.1) *and suppose that*

$$Z^\star \in \mathcal{Z} \overset{\text{def}}{=} \big\{ Z : \ \Omega^D_g(Z - c) \le r_g, \quad g \in \mathcal{G} \big\}.$$

Then, $\beta^\star_g = 0$ if $1 - \Omega^D_g(c) > r_g$.

There are several ground rules for constructing a safe region $\mathcal{Z}$:

- The supremum of the dual norm over the safe region $\sup_{Z\in\mathcal{Z}} \Omega^D_g(Z)$ is easy to compute;
- The size of the safe region should be as small as possible: as the most trivial safe region is the whole space which screens out nothing, while the best one is $\mathcal{Z} = \{Z^\star\}$ which screens out all useless features.

In the literature, various safe regions have been proposed. The very first safe screening work [8] introduced the idea of static screening and sequential screening. For *static safe screening*, screening is only implemented as a pre-processing of data, so the amount of discarded features are fixed, and it is very crucial to construct a good safe region such that the amount of discarded features is as many as possible. If we have a finite sequence of regularization parameter λ_j for $j = 0, ..., J$ such that $\lambda_0 \ge \lambda_1 \ge \cdots \ge \lambda_J = \lambda$. Then static screening can be applied to each λ_j which results in sequential screening. For both static and sequential screening, the volume of the safe regions is always bounded way from 0 which limits the potential of screening. In addition to the safe region proposed in [8], other safe regions include dual polytope projection safe sphere [33] and *safe dome* [34]. Dynamic screening rules were later proposed in [20, 22], where they combine screening rules and numerical methods such that the constructed safe regions are generated by the sequence generated by the numerical scheme. As a result, the safe region can eventually converge to the dual certificate and screen out all useless features. Our approach will follow the idea of dynamic screening.

2.1.1 Gap-safe screening

In a series of work [20, 21, 22], the authors develop a gap-safe rule for screening, where the "gap" here refers to the use of the primal and dual function value gap in constructing the safe region. Recall the primal and dual problems (1.1) and (2.4). For any $\beta \in \mathbb{R}^n$ and $\theta \in \mathcal{K}_{\lambda,\eta}$, the duality gap is defined by

$$G_{\lambda,\eta}(\beta, \theta) = P_{\lambda,\eta}(\beta) - D_{\lambda,\eta}(\theta).$$

Let $\beta^\star$ and $\theta^\star$ be a primal solution and the dual solution respectively, then strong duality holds and

$$\forall \beta \in \mathbb{R}^n, \theta \in \mathbb{R}^m, \quad D_{\lambda,\eta}(\theta) \le D_{\lambda,\eta}(\theta^\star) = P_{\lambda,\eta}(\beta^\star) \le P_{\lambda,\eta}(\beta).$$

As a result, the duality gap $G_{\lambda,\eta}(\beta,\theta)$ is always non-negative.

Since the loss function is differentiable with gradient being L-Lipschitz continuous, the dual problem $D_{\lambda,\eta}(\theta)$ is μ-strongly concave with $\mu = 1/L$, and one has for any $\beta \in \mathbb{R}^n$ and θ in the dual feasible set $\mathcal{K}_{\lambda,\eta}$, $\frac{\mu^2}{2}\|\theta - \theta^\star\|^2 \leq P_{\lambda,\eta}(\beta) - D_{\lambda,\eta}(\theta)$ [22, Theorem 6]. Therefore, letting

$$r_t \stackrel{\text{def}}{=} \mu^{-1}\sqrt{2G_{\lambda,\eta}(\beta_t,\theta_t)}, \tag{2.5}$$

one obtains the following safe sphere:

$$\mathcal{Z} \stackrel{\text{def}}{=} \Big\{ -\tfrac{1}{\lambda}\textstyle\sum_{i=1}^m \eta_i\theta_i^\star x_i, \ \ \forall\theta \in \mathcal{Z}_\theta \Big\} \quad \text{with} \quad \mathcal{Z}_\theta \stackrel{\text{def}}{=} \big\{\theta \in \mathbb{R}^m : \|\theta - \theta_t\| \leq r_t\big\}.$$

Now given a numerical scheme, at each iteration, β_t is explicitly available and one can compute a dual feasible variable θ_t by projecting $\bar\theta_t \stackrel{\text{def}}{=} (f'_{y_i}(x_i^\top\beta_t))_{i=1}^m$ on to the dual feasible set $\mathcal{K}_{\lambda,\eta}$. In particular, define

$$Z \stackrel{\text{def}}{=} c^{-1}\sum_i \eta_i(\bar\theta_t)_i x_i, \quad \text{where} \quad c \stackrel{\text{def}}{=} \max\{1, \Omega^D(\frac{1}{\lambda}\sum_{i=1}^m \eta_i(\bar\theta_t)_i x_i)\}. \tag{2.6}$$

It follows by applying Holder's inequality and using the fact that Ω^D is positive homogeneous that

$$\forall g \in \mathcal{G}, \quad \Omega_g^D(Z - Z^*) \leq \|\theta_t - \theta^*\|\sqrt{\sum_i \eta_i^2\Omega_g^D(x_{i,g})^2}$$

and hence, $(\beta_t)_g = 0$ if $\Omega_g^D(Z) < 1 - r_t\sqrt{\sum_i \eta_i^2\Omega_g^D(x_{i,g})^2} \stackrel{\text{def}}{=} 1 - r_g$.

2.2 Gap-safe screening for Prox-SGD

Screening rules are algorithm agnostic. That is to say, given an algorithm which outputs β_t, one can always compute a safe region $\mathcal{Z}$ for screening. As a result, we can incorporate screening to proximal stochastic gradient descent when the problem to solve has a finite sum empirical loss of the form (1.1). This is the most straightforward way of carrying out variable screening, and indeed, a similar screening strategy was recently proposed for ordered weighted ℓ_1 regularised regression in [?].

For the finite sum problem, consider an algorithm of the following general form: for each t, sample (x_t, y_t) uniformly at random from the finite data

$$\begin{aligned} \theta_t &= f'_{y_t}(x_t^\top\beta_t) \\ \beta_{t+1} &= \mathcal{T}(\beta_t, \theta_t x_t, \gamma_t) \end{aligned} \tag{2.7}$$

In the case of SGD, $\mathcal{T}(\beta_t, \phi_t, \gamma_t) = \beta_t - \gamma_t(\theta_t + \lambda Z_t)$ with $Z_t \in \partial\Omega(\beta_t)$, while for Prox-SGD we have $\mathcal{T}(\beta_t, \theta_t, \gamma_t) = \mathrm{prox}_{\gamma_t\tau\Omega}(\beta_t - \gamma_t\phi_t)$. We have the following algorithm which combines (2.7) with safe screening rules.

Algorithm 1: Safe screening for finite sum problem

```
Given: $T > 0$, step-size $\{\gamma_t\}_{t\in}$;
initialization $t = 1$, $\bar{\beta}_0 \in \mathbb{R}^n$;
while not terminate do
    $\beta_0 = \bar{\beta}_{t-1}$ ;                                        // set the anchor point
    for $j = 0, \dots, T-1$ do
        Sample $(x_j, y_j)$ uniformly at random ;                            // random sampling
        $\beta_j = \mathcal{T}(\beta_{j-1}, f'_{y_j}(x_j^\top \beta_{j-1})x_j, \gamma_j)$ ;    // standard gradient update
        $j = j + 1$;
    end
    $\bar{\beta}_t = \beta_T, \bar{\theta}_t = (f'_{y_i}(x_i^\top \bar{\beta}_t))_{i=1,\dots,m}$ ;   // primal and dual variables
    Compute safe centre $Z$ and radius $r_g$ ;                 // e.g.  as in (2.6) and (2.5)
    $\mathcal{S} = \left\{ g \in \mathcal{G} : \Omega_g^D(Z) < 1 - r_g \right\}$ ;                // screening set
    $(\bar{\beta}_t)_{\mathcal{S}} = 0$ ;                                   // pruning the primal point
    $t = t + 1$;
end
```

Remark 2.2. The above algorithm has two loops of iteration: the inner loop is the standard stochastic gradient update, while for the outer loop, screening with certain safe rules is applied. Note that the outer loop makes use of $\bar{\theta}_t$ which is evaluated over the entire dataset. Such a setting is reminiscent of the SVRG algorithm [10], where the full gradient of the loss function at an anchor point needs to be computed. Likewise, the choice of steps for inner loop, the value of T in Algorithm 1 should be of the order of m, to balance the overhead of computing θ_t.

Remark 2.3. For Algorithm 1, all the aforementioned safe screening rules can be applied; see [8, 18, 20, 22] and the references therein. However, for online learning, this is no longer true, since for online learning it is expensive to obtain the projected point $\bar{\theta}_t$, let alone construct the safe region $\mathcal{Z}$ using (2.6) with the projected dual point. Hence, in what follows, we propose an approach to compute an online gap and construct a safe region without the need to project onto the constraint set.

3 Screening for online algorithms

For large-scale problems online optimization methods, it is unrealistic or impossible to compute the dual variable $\bar{\theta}_t$. Consequently, one cannot construct the safe region $\mathcal{Z}$ for screening. However, for the gap-safe screening rule, since its safe region is built on function duality gap, it is possible to generalize the rule to the online setting via stochastic approximation. The purpose of this section is to build this generalization. The roadmap of this section reads

1. We first describe how to construct online dual certificates and primal/dual objectives, which consist of the following aspects: a) the dual problem of the online problem ($\mathcal{P}_\lambda$); b) online duality gap via stochastic approximation; c) online dual certificate; d) convergence guarantees. These are provided in Section 3.1.
2. With the online duality gap and dual certificate obtained in the first stage, we then can extend the gap-safe screening rule to the online setting. This extension is described in Section 3.2.

3. Finally in Section 3.3, we summarize our online-screening scheme for proximal stochastic gradient descent in Algorithm 2.

3.1 Online dual certificates and objectives

Given an online optimization method, at each iteration, we sample (x_t, y_t) from the distribution Λ and evaluate

$$\theta_t \overset{\text{def}}{=} f'_{y_t}(x_t^\top \beta_t) \tag{3.1}$$

In what follows, we define an online dual point $\bar{Z}_t$ which is constructed as weighted average of the past evaluated points $\{\theta_s\}_{s\leq t}$ and define online primal and dual objectives which are again weighted averages of the past selected functions $\{f_{y_s}\}_{s\leq t}$ and $\{f^*_{y_s}\}_{s\leq t}$, where f^*_y denotes the convex conjugate of loss function f_y.

Dual problem We first consider the dual problem of the primal problem ($\mathcal{P}_\lambda$), which reads

$$\max_v \left\{ \mathcal{D}(v) \overset{\text{def}}{=} -\mathbb{E}_{(x,y)}[f^*_y(v(x,y))] :\ \Omega^D\left(\mathbb{E}_{(x,y)}[v(x,y)x]\right) \leq \lambda \right\} \tag{$\mathcal{D}_\lambda$}$$

where we maximize over Λ-measurable functions v. Note that it admits a unique maximizer, since f_y is L-Lipschitz smooth which implies that f^*_y is $\frac{1}{L}$-strongly convex. The problems ($\mathcal{P}_\lambda$) and ($\mathcal{D}_\lambda$) are referred as primal and dual problems and their solutions are related: Any minimizer $\beta^\star$ of ($\mathcal{P}_\lambda$) is related to the optimal solution $v^\star$ of ($\mathcal{D}_\lambda$) by $v^\star(x,y) = f'(x^\top\beta^\star, y)$ and

$$Z^\star \overset{\text{def}}{=} -\frac{1}{\lambda}\mathbb{E}_{(x,y)}[v^\star(x,y)x] \in \partial\Omega(\beta^\star). \tag{3.2}$$

Online duality gap Observe that the primal and dual objective functions are expectations. We now discuss their online ergodic estimations over the sampled data. For each iteration step $t \in \mathbb{N}$, let $\mu_t \in (0,1)$. Given a primal variable $\beta \in \mathbb{R}^n$ and a dual variable $\zeta \in \mathbb{R}^t$, define the online primal objective

$$\bar{P}^{(t)}(\beta) \overset{\text{def}}{=} \bar{F}^{(t)}(\beta) + \lambda\Omega(\beta) \tag{3.3}$$

where for $\beta \in \mathbb{R}^n$,

$$\bar{F}^{(t)}(\beta) \overset{\text{def}}{=} \mu_t f_{y_t}(x_t^\top\beta) + (1-\mu_t)\bar{F}^{(t-1)}(\beta), \qquad \bar{F}^{(1)}(\beta) = f_{y_1}(x_1^\top\beta)$$

and the online dual objective for $\zeta \in \mathbb{R}^t$,

$$\bar{D}^{(t)}(\zeta) \overset{\text{def}}{=} -\mu_t f^*_{y_t}(\zeta_t) + (1-\mu_t)\bar{D}^{(t-1)}((\zeta_s)_{s\leq t-1}), \qquad \bar{D}^{(1)}(\zeta_1) = -f^*_{y_1}(\zeta_1)$$

Remark 3.1. Note that the primal variable has a fixed dimension which is n, however the dimension of the dual variable ζ grows with iteration t.

We make the following standard assumption on μ_t. Typical choices are $\mu_t = t^{-u}$ for $u \in (0.5, 1]$.

Assumption 3.2 ([25]). *We assume that $\mu_t \in (0,1)$ and*

$$\sum\nolimits_t \mu_t = +\infty \quad \textit{and} \quad \sum\nolimits_t \mu_t^2 < \infty \tag{3.4}$$

It is straightforward to check (see Lemma A.4 (i)) that there exists a decreasing sequence $\eta_s^{(t)} > 0$ such that

$$\textstyle\sum_{s\leq t} \eta_s^{(t)} = 1 \quad \text{and} \quad \eta_s^{(t)} \overset{\text{def}}{=} \mu_s \prod_{i=s+1}^{t} (1-\mu_i) \tag{3.5}$$

so that in our previous notation of (1.1) and (2.4), we have $\bar{P}^{(t)} = P^{\eta^{(t)}}$ and $\bar{D}^{(t)} = D^{\eta^{(t)}}$, and they are related by

$$\min_{\beta} \bar{P}^{(t)}(\beta) = \max_{\xi \in \mathcal{K}_{\lambda,\eta^{(t)}}} \bar{D}^{(t)}(\xi)$$

Definition 3.1 (Online duality gap). We define the online duality gap at $\beta \in \mathbb{R}^n$ as

$$\mathrm{Gap}_t(\beta) \overset{\text{def}}{=} \bar{P}^{(t)}(\beta) - \bar{D}^{(t)}((\theta_s)_{s\leq t}). \tag{3.6}$$

where we recall the definition of θ_s from (3.1).

Remark 3.3. Since θ_s is not necessarily a dual feasible point, the online duality gap $\mathrm{Gap}_t(\beta)$ is not guaranteed to be non-negative. As we shall see in the next paragraph, while the gap Gap_t can be computed in an online fashion, the feasible point $\bar{\theta}_s$, in $\bar{P}^{(t)}(\beta) - \bar{D}^{(t)}((\bar{\theta}_s)_{s\leq t})$, which is the projection of θ_s onto the constraint set cannot be computed online.

An online estimate of the dual certificate With the online duality gap, we now construct a dual certificate from β_t and θ_t. For the primal variable, since β_t converges to $\beta^\star$, it is natural to define a candidate point, for $\mu_t > 0$, as

$$\bar{Z}_t \overset{\text{def}}{=} -\frac{1}{\lambda}\mu_t \theta_t x_t + (1-\mu_t)\bar{Z}_{t-1}, \quad \text{and} \quad \bar{Z}_1 \overset{\text{def}}{=} -\frac{1}{\lambda}\theta_1 x_1. \tag{3.7}$$

In the notation introduced in (3.5), we can write

$$\bar{Z}_t = -\frac{1}{\lambda} \textstyle\sum_{s=1}^{t} \eta_s^{(t)} \theta_s x_s = -\frac{1}{\lambda} \sum_{s=1}^{t} \eta_s^{(t)} f'_{y_s}(x_s^\top \beta_s) x_s. \tag{3.8}$$

Convergence results Before presenting our online-screening rule, we provide some theoretical convergence analysis of the above online estimates. We first establish uniform convergence of the online objective $\bar{P}^{(t)}$ to its expectation $\mathcal{P}$, and uniform convergence of the corresponding dual certificate $Z^{*,(t)}$ to $Z^\star$.

Proposition 3.4 (Convergence of online objectives). *The following result holds*

(i) let $\mathcal{O}$ be a compact set of $\mathbb{R}^n$,

$$\sup_{\beta \in \mathcal{O}} |\bar{P}^{(t)}(\beta) - \mathcal{P}(\beta)| \to 0, \qquad t \to +\infty$$

almost surely.

(ii) Let $Z^{*,(t)}$ be the dual certificate associated to $\bar{P}^{(t)}$, we have uniform convergence to $Z^\star$ defined in (3.2):

$$\|Z^{*,(t)} - Z^\star\|_\infty \to 0, \qquad t \to +\infty.$$

We also have convergence of the online certificate $\bar{Z}_t$ to Z^*, and the online gap evaluated at converging points also converges to zero.

Proposition 3.5 (Convergence of the online estimate). *Let $\beta^\star$ be a minimizer of $\mathcal{P}$ and assume that $\beta_t \to \beta^\star$ almost surely. Then with probability 1,*

(i) $\bar{Z}_t$ converge to Z^* as $t \to \infty$.

(ii) If $\bar{\beta}_t \to \beta^\star$, then $\mathrm{Gap}_t(\bar{\beta}_t) \to 0$ as $t \to \infty$.

The proofs can be found in the appendix.

3.2 Online-screening

In this section, we derive a screening rule for solutions to the online objective $\bar{P}^{(t)}$ based on the certificate $\bar{Z}_t$ and Gap_t. In the following, let $\bar{Z}_t$ be as defined in (3.8), θ_t be as in (3.1) and let $\bar{Z}_t^* = -\frac{1}{\lambda}\sum_{s=1}^t \eta_t^{(t)}\theta_s^{(t),*}x_s$ where $\theta^{(t),*}$ is the maximizer of $\bar{D}^{(t)}$.

Lemma 3.6 (Screen gap). *Let $\bar{\beta} \in \mathbb{R}^n$, then there holds*

$$\frac{1}{2L}\sum\nolimits_{s=1}^t \eta_s^{(t)}|\theta_s - \theta_s^{(t),*}|^2 \leq \text{Gap}_t(\bar{\beta}) + \sum\nolimits_{s=1}^t \eta_s^{(t)} f_{y_s}(0)\,\left(\Omega^D(\bar{Z}_t) - 1\right)_+$$

Moreover, for all $g \in \mathcal{G}$,

$$\Omega_g^D(\bar{Z}_t - \bar{Z}_t^*) \leq r_g^{(t)}(\bar{\beta}, \bar{Z}_t)$$

where

$$r_g^{(t)}(\bar{\beta}, \bar{Z}_t) \overset{\text{def}}{=} \frac{\sqrt{2LN_g}}{\lambda}\sqrt{\text{Gap}_t(\bar{\beta}) + \sum\nolimits_{s=1}^t \eta_s^{(t)} f_{y_s}(0)\,(\Omega^D(\bar{Z}_t) - 1)_+}$$

with $N_g \overset{\text{def}}{=} \sum_{s=1}^t \eta_s^{(t)}\Omega_g^D(x_s)^2$.

By combining Lemma 3.6 with Proposition 2.1, we obtain the following online-screening rule.

Corollary 3.7 (Screen rule). *Let $\beta^{(t),*} \in \text{argmin}_\beta \bar{P}^{(t)}(\beta)$. Then, given any $\bar{\beta} \in \mathbb{R}^n$, $\beta_g^{(t),*} = 0$ if*

$$1 - \Omega_g^D(\bar{Z}_t) > r_g^{(t)}(\bar{\beta}, \bar{Z}_t).$$

Remark 3.8.

(i) Compared with the gap-safe rules of [20, 21, 22], we do not project $\bar{Z}_t$ onto the dual feasible set $\mathcal{K}$ and hence have an additional term $(\Omega^D(\bar{Z}_t) - 1)_+$ in $r_g^{(t)}(\beta, \bar{Z}_t)$.

(ii) The above screening is safe for the online problem $\bar{P}^{(t)}$ in the sense that screening will not falsely remove features which are in the solution $\beta^{(t),*}$ of $\bar{P}^{(t)}$, however, the support of $\beta^{(t),*}$ may not necessarily coincide with that of the global minimizer β^* of ($\mathcal{P}_\lambda$) and hence, our rule is not necessarily safe for the expectation ($\mathcal{P}_\lambda$). Further discussions on the safety of our rule can be found in Section 4.4.

We can directly apply Corollary 3.7 to screen out variables while running SGD. However, the effectiveness of this rule will depend on the proximity of $\bar{\beta}$ to the optimal point $\beta^\star$. We therefore propose to progressively update this *anchor point* $\bar{\beta}$.

Let $0 = t_0 < t_1 < t_2 < \cdots < t_k = T$ and denote $[t_{j-1}, t_j] \overset{\text{def}}{=} \{t_{j-1}+1, \ldots, t_j\}$. Let $\eta_s^{(T)} \in (0,1)$ for $s \in [0,T]$ be such that $\sum_{s\in[0,T]}\eta_s^{(T)} = 1$. Given $j \in \{1, \ldots, k\}$, let $(\theta_s^*)_{s\in[t_{j-1},t_j]}$ be the optimal dual solution to

$$\max_\zeta \sum\nolimits_{s\in[t_{j-1},t_j]} - \eta_s^{(T)} f_{y_s}^*(\zeta_s) \tag{3.9}$$

where

$$\Omega^D\left(\sum\nolimits_{s\in[t_{j-1},t_j]}\eta_s^{(T)}x_s\zeta_s\right) \leq \lambda\sum\nolimits_{s\in[t_{j-1},t_j]}\eta_s^{(T)}.$$

Note that (3.9) is dual to the primal problem

$$\min_{\beta} \sum\nolimits_{s\in[t_{j-1},t_j]} \eta_s^{(T)} \big(f_{y_s}(x_s^\top \beta) + \lambda\Omega(\beta) \big).$$

The corresponding dual certificate is

$$Z_j^* \overset{\text{def}}{=} -\frac{1}{\lambda} \sum\nolimits_{s\in[t_{j-1},t_j]} \frac{\eta_s^{(t)}}{\gamma_j} \theta_s^*, \quad \text{where} \quad \gamma_j = \sum\nolimits_{s\in[t_{j-1},t_j]} \eta_s^{(T)}.$$

For each j, by applying Lemma 3.6 with $\{\theta_s\}_{s\in[t_{j-1},t_j]}$, $\bar\beta \overset{\text{def}}{=} \beta_{t_{j-1}}$ and $\{\theta_s^*\}_{s\in[t_{j-1},t_j]}$, we obtain

$$\begin{aligned} \sum_{s\in[t_{j-1},t_j]} \eta_s^{(T)} |\theta_s - \theta_s^*|^2 \leq \sum_{s\in[t_{j-1},t_j]} \eta_s^{(T)} \Bigg(& f_{y_s}(x_s^\top \beta_{t_{j-1}}) \\ & + \lambda\Omega(\beta_{t_{j-1}}) - f_{y_s}^*(\theta_s) + f_{y_s}(0)(\Omega^D(Y_j) - 1)_+ \Bigg) \end{aligned} \tag{3.10}$$

where

$$Y_j \overset{\text{def}}{=} \frac{1}{\sum_{s\in[t_{j-1},t_j]} \eta_s^{(T)}} \sum\nolimits_{n\in[t_{j-1},t_j]} \eta_s^{(T)} \theta_s x_s.$$

Summing (3.10) over $j = 1, \ldots, k$ and denoting $\bar\beta_s \overset{\text{def}}{=} \beta_{t_{j-1}}$ for $s \in [t_{j-1}, t_j]$, we obtain

$$\sum_{s=1}^{T} \eta_s^{(T)} |\theta_s - \theta_s^*|^2 \leq R_T \tag{3.11}$$

where

$$R_t \overset{\text{def}}{=} \sum_{s=1}^{T} \eta_s^{(T)} \left(f_{y_s}(x_s^\top \bar\beta_s) + \lambda\Omega(\bar\beta_s) - f_{y_s}^*(\theta_s) \right) + \sum_{j=1}^{k} V_j (\Omega^D(Y_j) - 1)_+$$

with $V_j \overset{\text{def}}{=} \sum_{s\in[t_{j-1},t_j]} \eta_s^{(T)} f_{y_s}(0)$.

Define

$$\bar Z^* = \sum\nolimits_{j=1}^{k} \gamma_j Z_j^* = -\frac{1}{\lambda} \sum\nolimits_j \sum\nolimits_{s\in[t_{j-1},t_j]} \eta_s^{(T)} f'_{y_s}(x_s^\top \bar\beta_j^*) x_s,$$

and we have $\bar Z_T \overset{\text{def}}{=} \sum_{s=1}^{T} \eta_s^{(T)} \theta_s x_s$ satisfy

$$\Omega_g^D(\bar Z_T - \bar Z^*) \leq \frac{\sqrt{2LR_T N_{T,g}}}{\lambda}, \tag{3.12}$$

where $N_{T,g} \overset{\text{def}}{=} \sum_{s=1}^{T} \eta_s^{(T)} \Omega_g^D(x_s)^2$. Note that the residual term R_T is now dependent on the sequence β_{t_j} which converges to $\beta^\star$ as $j \to \infty$. We can therefore expect the RHS of (3.12) to converge to 0 as $T \to \infty$. Having established how to progressively update the anchor point, we finally present our online-screening rule for online optimization algorithms in the next section.

3.3 Screening procedure

Consider an online algorithm of the following form: for each $t = 1, 2, \ldots,$, draw sample $(x_t, y_t) \sim \Lambda$, and update

$$\begin{aligned} \theta_t &= f'_{y_t}(x_t^\top \beta_t) \\ \beta_{t+1} &= \mathcal{T}(\beta_t, \theta_t x_t, \gamma_t) \end{aligned} \tag{3.13}$$

where $\mathcal{T}$ is a fixed point operator. In the case of SGD, $\mathcal{T}(\beta_t, \phi_t, \gamma_t) = \beta_t - \gamma_t(\theta_t + \lambda Z_t)$, while for Prox-SGD we have $\mathcal{T}(\beta_t, \theta_t, \gamma_t) = \mathrm{prox}_{\gamma_t \tau \Omega}(\beta_t - \gamma_t \phi_t)$.

We state our screening framework for online optimization methods in Algorithm 2 below.

Algorithm 2: Online optimization algorithm with screening

```
Given: step-size $\{\gamma_t\}_{t\in\mathbb{N}}$, exponent $w$, $\mu_t \overset{\text{def}}{=} 1/t^w$, initial point $\bar{\beta} \in \mathbb{R}^n$;
initialization $t = 1$; $p_0 = d_0 = N_{0,g} = 0$;
$u_0 = 0$; $S = 0$; $Z = 0_n$;
while not terminate do
    $\beta_0 = \bar{\beta}$ ;                                              // set the anchor point
    $v_0 = 0$; $X_0 = 0$;
    for $t = 1, \ldots, T$ do
        $(x_t, y_t) \sim \Lambda$ ;                                           // random sampling
        $\theta_t = f'_{y_t}(x_t^\top \beta_{t-1})$; $\beta_t = \mathcal{T}(\beta_{t-1}, \theta_t x_t, \gamma_t)$ ;     // standard gradient update
        $X_t = -\frac{1}{\lambda}\mu_t \theta_t x_t + (1-\mu_t) X_{t-1}$ ;                   // certificate update
        $p_t = \mu_t(f_{y_t}(x_t^\top \bar{\beta}) + \lambda\Omega(\bar{\beta})) + (1-\mu_t) p_{t-1}$ ;          // primal value
        $d_t = -\mu_t f^*_{y_t}(\theta_t) + (1-\mu_t) d_{t-1}$ ;                         // dual value
        $\forall g \in \mathcal{G},\ N_{t,g} = \mu_s \Omega_g^D(x_t)^2 + (1-\mu_s) N_{t-1,g}$ ;
        $v_t = \mu_t f_{y_t}(0) + (1-\mu_t) v_{t-1}$;
        $u_t = (1-\mu_t) u_{t-1}$;
    end
    $\bar{\beta} = \beta_{t-1}$ ;                                          // update anchor point
    $Z = u_t Z + X_t$ ;                                             // estimated certificate
    $S = u_t S + v_t \Omega^D(X_t/(1-u_t) - 1)_+$;
    $R = p_t - d_t + S$;
    $\mathcal{S} = \left\{ g \in \mathcal{G} :\ \Omega_g^D(Z) < 1 - \frac{\sqrt{2LN_{t,g}R}}{\lambda} \right\}$ ;                 // screening set
    $(\beta_t)_{\mathcal{S}} = 0$ ;                                         // pruning the primal point
    $u_0 = 1$;
end
```

Next we provide some discussions on how to compute some key values of the algorithm, for instance the terms described in (3.11) and (3.12).

- It is straightforward to compute $\bar{Z}_T$ and $N_{T,g}$, as we have

$$\begin{aligned} \bar{Z}_s &\overset{\text{def}}{=} \mu_s \theta_s x_s + (1-\mu_s)\bar{Z}_{s-1}, && \text{where} \quad \bar{Z}_0 = 0, \\ N_{s,g} &\overset{\text{def}}{=} \mu_s \Omega_g^D(x_s)^2 + (1-\mu_s) N_{s-1,g}, && \text{where} \quad N_{0,g} = 0. \end{aligned}$$

- Next is the computation of

$$R_T \overset{\text{def}}{=} \underbrace{\sum_{s=1}^{T} \eta_s^{(T)}(f_{y_s}(x_s^\top \bar{\beta}_s) + \lambda\Omega(\bar{\beta}_s))}_{p_T} - \underbrace{\sum_{s=1}^{T} -\eta_s^{(T)} f_{y_s}^*(\theta_s)}_{d_T} + \underbrace{\sum_{j=1}^{k} V_j\big(\Omega^D(Y_j) - 1\big)_+}_{S_T}$$

The first two terms are straightforward: define $\bar{\beta}_s = \beta_{t_{j-1}}$ for all $s \in [t_{j-1}, t_j]$ and repeat this over $s = 1, \ldots, t$:

$$p_s \overset{\text{def}}{=} \mu_s\big(f_{y_s}(x_s^\top \bar{\beta}_s) + \lambda\Omega(\bar{\beta}_s)\big) + (1 - \mu_s)p_{s-1}, \qquad \text{where} \quad p_0 = 0,$$
$$d_s \overset{\text{def}}{=} -\mu_s f_{y_s}^*(\theta_s) + (1 - \mu_s)d_{s-1}, \qquad \text{where} \quad d_0 = 0.$$

To compute S_T: during the first $s = 1, \ldots, t_1$ iterations, define

$$X_s^{(1)} \overset{\text{def}}{=} \mu_s\theta_s x_s + (1 - \mu_s)X_{s-1}^{(1)}, \qquad \text{where} \quad X_0^{(1)} = 0,$$
$$v_s^{(1)} \overset{\text{def}}{=} \mu_s f_{y_s}(0) + (1 - \mu_s)v_{s-1}^{(1)}, \qquad \text{where} \quad v_0^{(1)} = 0.$$

and that

$$Y_1 \overset{\text{def}}{=} X_{t_1}, \qquad V_1 \overset{\text{def}}{=} v_{t_1}, \qquad S_{t_1} \overset{\text{def}}{=} V_1\big(\Omega^D(Y_1) - 1\big)_+.$$

Then for iteration $s \in [t_{j-1}, t_j]$, $j = 2, \ldots, k$, we have

$$X_s^{(j)} \overset{\text{def}}{=} \mu_s\theta_s x_s + (1 - \mu_s)X_{s-1}^{(j)}, \qquad \text{where} \quad X_{t_1}^{(j)} = 0,$$
$$v_s^{(j)} \overset{\text{def}}{=} \mu_s f_{y_s}(0) + (1 - \mu_s)v_{s-1}^{(j)}, \qquad \text{where} \quad v_{t_{j-1}}^{(j)} = 0.$$

At iteration t_j: define $\gamma_j \overset{\text{def}}{=} \prod_{s \in [t_{j-1}, t_j]}(1 - \mu_s)$ and

$$Y_j \overset{\text{def}}{=} \frac{1}{1 - \gamma_j} X_{t_j}^{(j)}, \qquad V_2^{(j)} \overset{\text{def}}{=} v_{t_j}^{(j)} \quad \text{and} \quad S_{t_j} \overset{\text{def}}{=} \gamma_j S_{t_{j-1}} + V_j(\Omega^D(Y_j) - 1)_+.$$

Note that we in fact have $\bar{Z}_{t_j} = X_{t_j}^{(j)} + \gamma_j \bar{Z}_{t_{j-1}}$.

We conclude this section by few remarks.

Remark 3.9 (Computational pains and gains). Our screening rule adds several computational overheads to the original online optimization problem, however, all of them are of $\mathcal{O}(n)$ complexity. Denote by n_t the dimension of the problem at current iteration.

- For the *inner loop* of Algorithm 2, `line 10-12` computing the dual certificate and primal/dual function values are of $\mathcal{O}(n_t)$ complexity.
- For the *outer loop* of Algorithm 2, all computations are at most $\mathcal{O}(n_t)$.

Overall, the computational overheads added by screening is $\mathcal{O}(n_t)$ per iteration where n_t is the dimension of β_t at iteration step t

On the other hand, since our screening rule can effectively remove useless features along iteration. Suppose the sparsity of $\beta^\star$ is κ which is much smaller than n and our screening rule manages to screen out all useless features, then we have eventually $n_t = \kappa$ for all t large enough, which in turn means the computational overheads are negligible.

Remark 3.10 (Effect of the exponent w). For Algorithm 2, the weight parameter μ_t, specified by the exponent w, determines how important the latest iterate is. As a result, w is crucial to the screening behaviour of Algorithm 2. In general the value of w lies in $]0.5, 1]$. As we shall see in the numerical experiments, the smaller the value of w, the more aggressive the screening rule which make Algorithm 2 unsafe. While for larger choice of w, the screening is much more passive, hence safer.

Remark 3.11 (Choices of T). For Algorithm 2, the inner loop iteration number is controlled by T. Similar to Algorithm 1, in practice, choices like ℓm with ℓ being small integers demonstrate good performance, and we refer to Section 4 the numerical experiments for more detailed discussions.

Remark 3.12 (Online-screening is *not* safe). Although our screening rule is adapted from gap-safe rule, which is guaranteed to be safe, *i.e.* only removes useless features and keeps all the active ones, algorithm 2 is *not* guaranteed to be safe. This is due to the fact that the rule we derive is with respect to the online objective $\bar{P}^{(t)}$ which is not the original objective $(\mathcal{P}_\lambda)$. As a result, potentially our screening rule can falsely remove useful features. However, this can be avoided by incorporating safe guard step, for instance, we can combine Algorithm 2 with the strong rules developed in [30] to avoid false removal.

4 Numerical results

To evaluate the effectiveness of our proposed screening scheme [1], in this section, we present numerical experiments for the following ℓ_1-regularized problem

$$\min_{\beta \in \mathbb{R}^n} F(\beta) + \|\beta\|_1$$

where $F(\beta) \overset{\text{def}}{=} \sum_{i=1}^m f(x_i^\top \beta; y_i)$ with f being either

(i) the quadratic loss $f(z; y) = \frac{1}{2}(z-y)^2$, aka the LASSO formulation,

(ii) the logistic loss $f(z; y) = \log(1+\exp(-yz))$ and $\Omega(\beta) = \|\beta\|_1$ with $y \in \{-1, 1\}$, aka sparse logistic regression.

For both problems, we compare the performances of the proposed online-screening (Algorithm 2) and the full-screening algorithm (Algorithm 1) using datasets from LIBSVM[2]. The details of the considered datasets are listed in Table 1.

As seen from the table, four relatively small-scale datasets and four large-scale datasets are considered. Three algorithms are compared: the standard proximal gradient descent (Prox-SGD), Algorithm 1 with Prox-SGD (FS-Prox-SGD), Algorithm 2 with Prox-SGD (OS-Prox-SGD). The details of settings of our experiments are as follows.

- SAGA [6] is used for computing the solution of the problems.
- Step-sizes are set as $\gamma_t = \frac{1}{mLt^{0.51}}$.
- The exponent w is set as 0.51 for all the tests. We refer to Section 4.4 for a discussion on this choice.

[1] Matlab code for reproducing our experiments are available at https://github.com/jliang993/sgd-screening

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Name	m	n
`colon-cancer`	62	2,000
`leukemia`	38	7,129
`breast-cancer`	44	7,129
`gisette`	6,000	5,000
`arcene`	200	10,000
`dexter`	600	20,000
`dorothea`	1,150	100,000
`rcv1`	20,242	47,236

Table 1: The considered data sets and their scales: m is the number of samples and n is the dimension of the problem.

- Regularization parameter: from the optimality condition, 0 is a solution if $\|\nabla F(0)\|_\infty \leq \lambda$. Therefore, to ensure non-trivial solutions, we should choose $\lambda < \lambda_{\max} \overset{\text{def}}{=} \|\nabla F(0)\|_\infty$. In our experiments for the LASSO problem, we choose $\lambda = \frac{\lambda_{\max}}{2}$. While for SLR problem, various choices are chosen and provided below.
- For both screening schemes, we set $T = 4m$, *i.e.* screening is applied every $4m$ steps.

In practice, we find that online screening is fairly robust to the choice of $T \geq m$, however, we remark that screening is computationally expensive for the full screening approach in Algorithm 1, and hence the choice of T should balance the number of screening and the total number of iteration of the method.

4.1 Dimension reduction of screening schemes

We first compare the support identification properties of Prox-SGD, FS-Prox-SGD and OS-Prox-SGD. The obtained results are shown in Figure 1 (LASSO) and Figure 2 (SLR), respectively.

For each figure, two quantities are provide: size of support over number of epochs of β_t for *solid lines* and elapsed time over number of epochs for *dashed lines*. For LASSO problem, we obtain the following observations,

- Prox-SGD, black lines in all figures, indeed does not have support identification property, as the size of support is oscillating and does not decrease.
- Except for `rcv1`, online-screening can effectively remove redundant features, while full-screening mainly works for smaller datasets. It should be noted that, limited by the number of iterations, support identification is not exactly reached.
- Between FS-Prox-SGD and OS-Prox-SGD, overall the latter demonstrates a better screening outcomes. In particular, OS-Prox-SGD can significantly reduce the dimension of the problem at the very early iteration stages.

For SLR problem, the choice of regularization parameter λ is provided under each sub-figure of Figure 2. The advantages of online-screening is similar to those of LASSO problem for the first six datasets. However, for `dorothea` and `rcv1` datasets, the behaviours are different

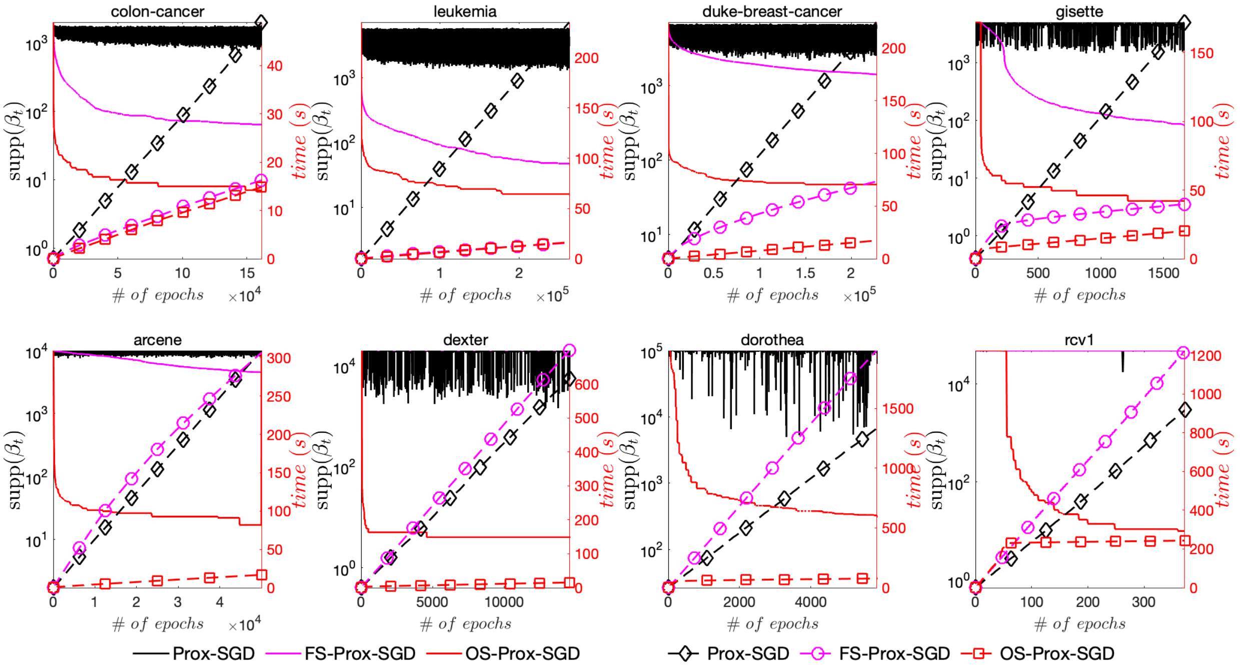


Figure 1: Comparison of support reduction and wall clock time for LASSO problems. For all datasets, the regularization parameter is $\lambda = \frac{\lambda_{\max}}{2}$. For each figure, two quantities are displayed: the *solid lines* show how the size of support of β_t decreases over number of epochs, while the *dashed lines* show the elapsed time increases over number of epochs. Online-screening substantially decreases the size of support and hence obtains substantial savings in computational time.

- For `dorothea` dataset, both screening schemes are slower than the pure Prox-SGD, while for LASSO problem, online-screening achieve dimension reduction.
- For `rcv1` dataset, online-screening eventually provides dimension reduction while for LASSO case, both schemes are not working.

We observe from above that, for both problems, when online-screening works, it can achieve dimension at the very early stage of the iteration, which means practically it is more attractive than the full-screening scheme, since in practice stochastic algorithms are run for limited number of epochs.

4.2 LASSO

In this part, we present absolute error $\|\beta_t - \beta^\star\|$ comparisons and solution quality comparisons for the LASSO problem. Error comparisons are displayed in Figure 3. Similarly to the wall-clock time comparisons in Figure 1, the faster algorithm yields faster error decays.

In Figure 4, we provide comparisons of the final outputs obtained by the algorithms, for the Prox-SGD schemes, the maximum number of iteration is set as 10^7. For the full-screening and online-screening variants of Prox-SGD, we run either to 10^7 iterations or when the wall-clock time exceeds that of Prox-SGD (whichever occurs first). Recall that, for reference, we use SAGA to compute the true solution and verify that it is indeed a solution using is the optimality condition.

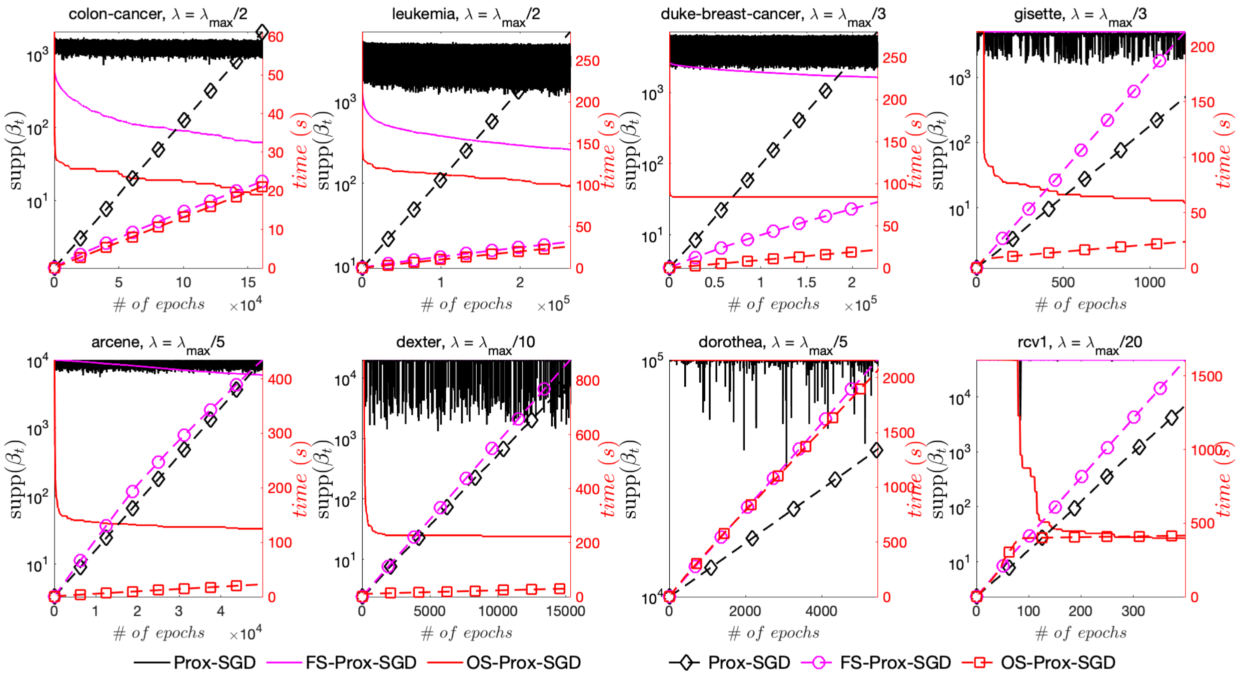


Figure 2: Comparison of support reduction and wall clock time for SLR problems. For each figure, two quantities are displayed: *solid lines* show the size of support β_t over number of epochs and *dashed lines* show the elapsed time over number of epochs.

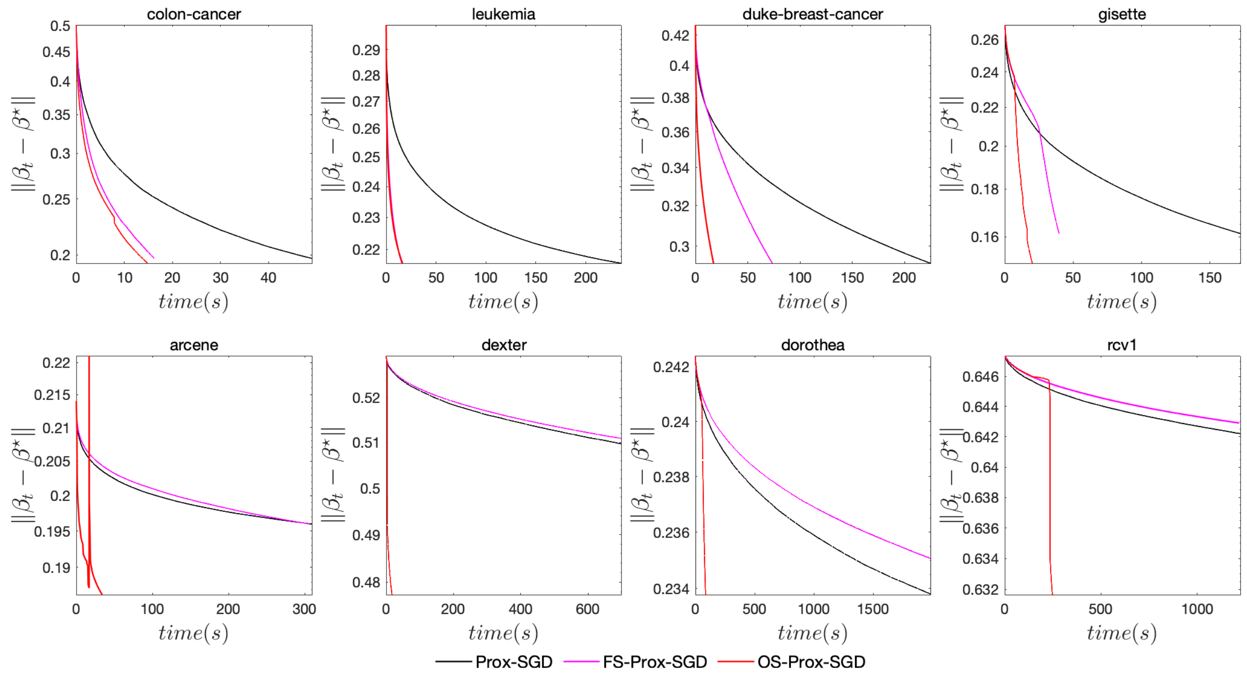


Figure 3: Comparison of errors $\|\beta_t - \beta^\star\|$ against wall clock time for LASSO problems. For all datasets, the regularization parameter is $\lambda = \frac{\lambda_{\max}}{2}$.

- It can be observed that for Prox-SGD, non-identification can be observed by the large number of tiny values around order 10^{-6}.

- In general, there are discrepancies between the outputs of Prox-SGD schemes and the solution by SAGA, which is caused by the vanishing step-size of SGD schemes.
- Screening can be effective in screening out these tiny values, with online-screening being better than full-screening.

Note that for the dexter dataset, online-screening over-screens the features which results in only 2-sparse output while the solution obtained by SAGA is 6-sparse. As we mentioned in Remark 3.10, the aggressiveness of online-screening is controlled by the exponent parameter w, in the last part of this section we provide a detailed discussion on this parameter.

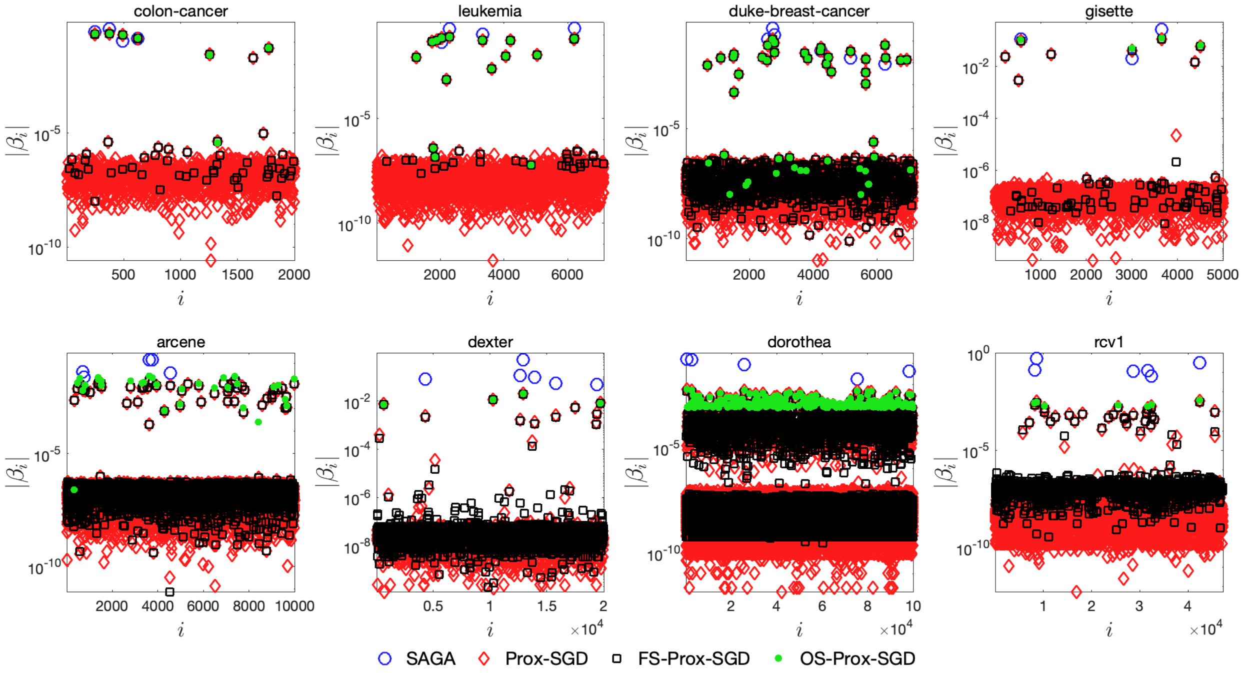


Figure 4: Comparison of the solutions for LASSO problems. The plots show the absolute value of the solution at each index. We display the 'ground truth' obtained by running SAGA to convergence, and the final outputs of prox-SGD, prox-SGD with full-screening and prox-SGD with online-screening. The number of iterations is capped at 10^7.

4.3 Sparse logistic regression

For SLR problems, the comparisons of error-time and solution quality are provided in Figure 5 and Figure 6 respectively. Similar to the LASSO problem, the error comparison is in consistent with time comparisons of Figure 2. For solution comparisons, for the dorothea dataset, the outputs of all Prox-SGD schemes are far from close to the solution obtained by SAGA.

It is worth noting that, this time for dexter dataset, the behaviour of online-screening is not as aggressive as that of LASSO case (see Figure 6 (f)), yet still the outputs of Prox-SGD schemes are quite far away from the solution obtained by SAGA. Moreover, for rcv1 data set, this time our online-screening is not safe where the 3rd feature (with smallest amplitude) of the output of SAGA is screened out.

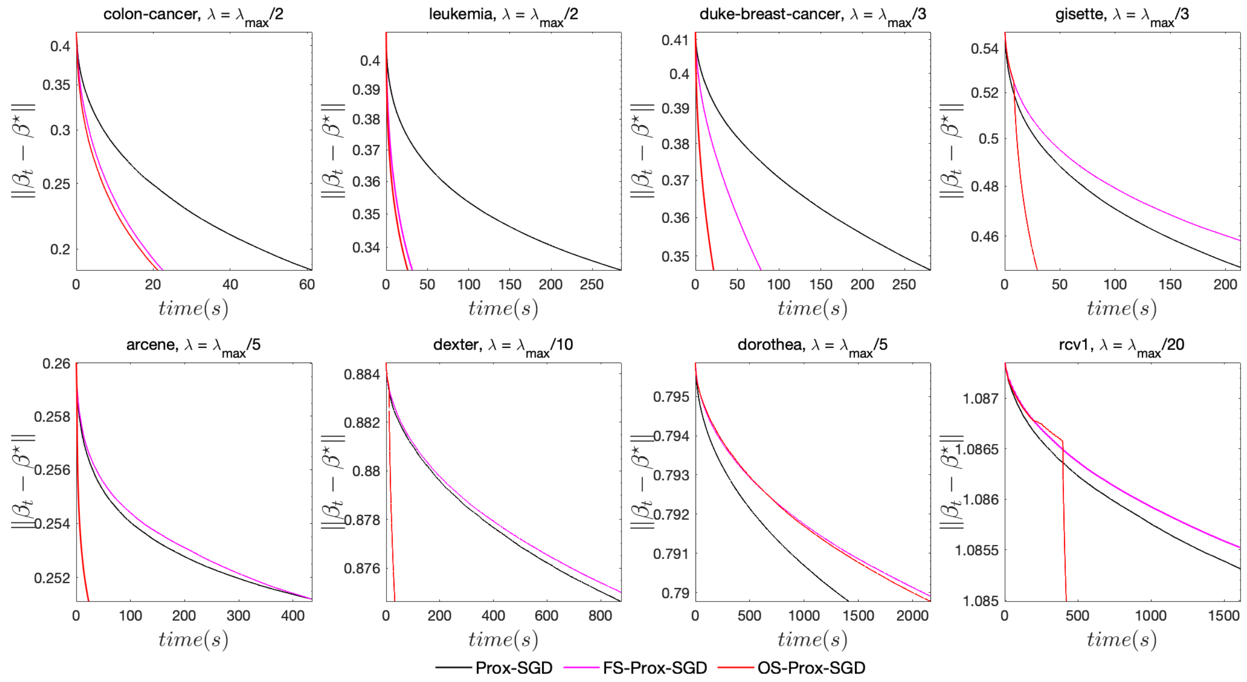


Figure 5: Comparison of error $\|\beta_t - \beta^\star\|$ against wall clock time for SLR problems.

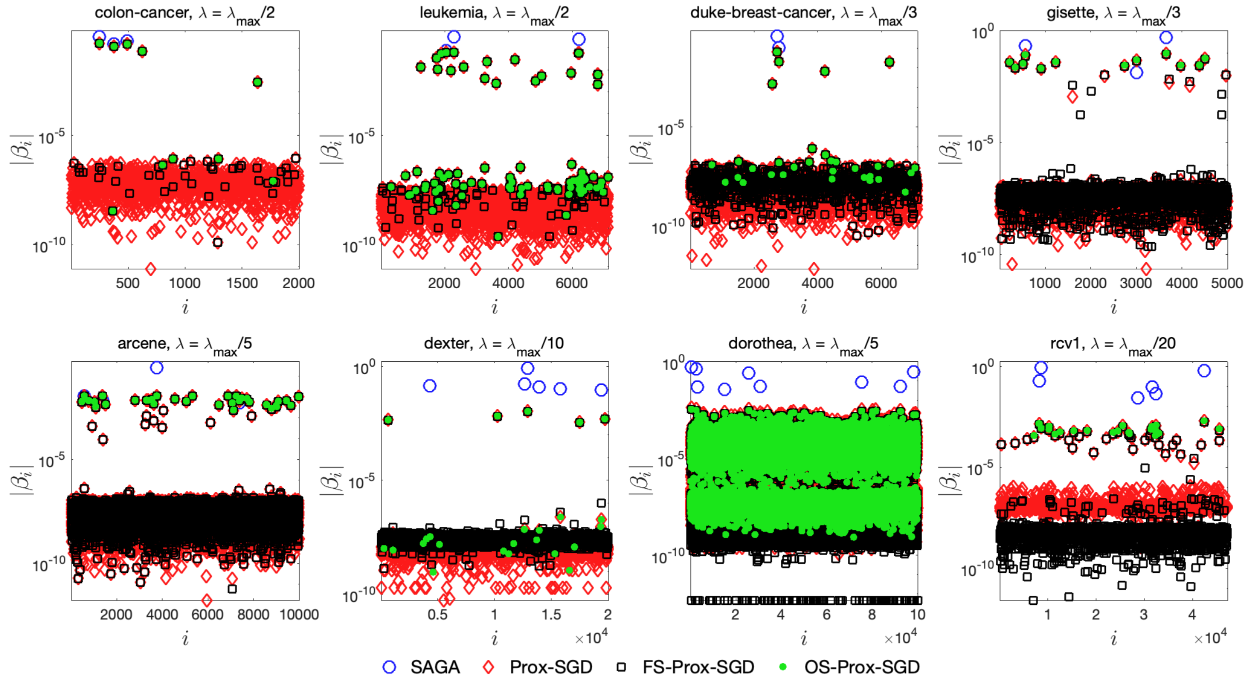


Figure 6: Comparison of the solutions for SLR problems. The plots show the absolute value of the solution at each index. We display the 'ground truth' obtained by running SAGA to convergence, and the final outputs of prox-SGD, prox-SGD with full-screening and prox-SGD with online-screening. The number of iterations is capped at 10^7.

4.4 Safety of online-screening

In this part we discuss the effect of the exponent parameter w in Algorithm 2 and comment on the non-safe behaviours of online-screening for LASSO on `dexter` dataset and SLR on rcv1 dataset For comparison purpose, three choices of the exponent w are tested, which are $w = 0.51, 0.6$ and 0.75. We describe the impact of this choice on three quantities: dimension reduction, solution and error decay. Since SAGA has the support identification property [23], we also include it for reference.

In the experiments, we only consider LASSO on dexter dataset and SLR on rcv1 dataset. Though limited, these two examples are demonstrative of the situation where online-screening can correctly identify the solution support and where small values of w results in false removal of solution support. Similar results can be observed for the other datasets described in the previous section.

LASSO problem The results for LASSO and dexter dataset are provided below in Figure 7, from which we observe the followings

- **Dimension reduction** For all choices of the exponent w, there is a sharp dimension reduction at the beginning stage of the iteration. Eventually, the smaller the value of w, the sparser the output of Prox-SGD. Moreover, the support identification of SAGA, see the *magenta line* which indicates the non-safe behaviour of $w = 0.51$.
- **Solution** As we have seen previously, $w = 0.51$ is not safe as it screens out true support[3]. For $w = 0.6, 0.75$, both choices retain the support solution, with $w = 0.6$ being the better one.
- **Error decay** Since $w = 0.51$ is not safe, the fast decay of the *black line* is meaningless. Both $w = 0.6$ and $w = 0.75$ produce almost the same error decay.

Finally, it is worth mentioning that the wall clock time for all three choices of w are very close and around 10 seconds.

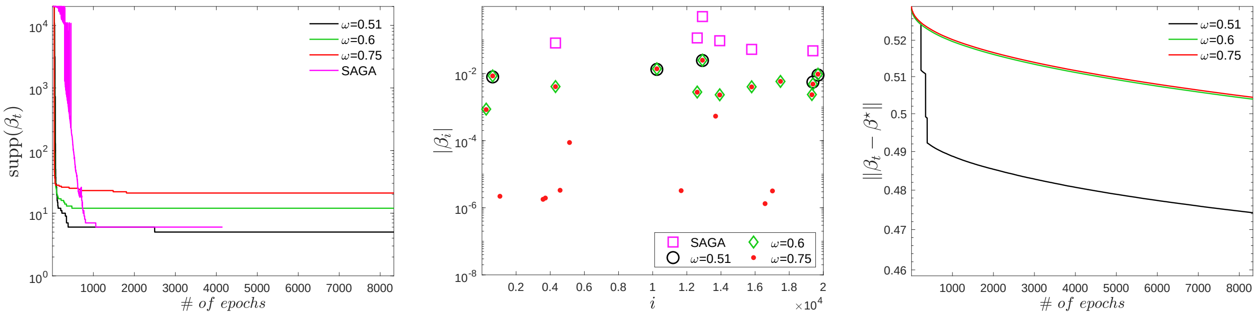
The main conclusion is that *smaller* values of w will lead to a more aggressive screening rule. This can however (in 2 of our tested datasets) lead to false removal of the solution support. To guarantee its output, certain safeguards need to be considered. For example, we can include the rules developed in [30], so that once false dimension reduction happens, the screening will be reset and the value of w will be increased by a certain margin until an upper bound is reached.

SLR problem To conclude this section, we present numerical results for SLR on rcv1 dataset, which are shown in Figure 8. The observations are similar to those we obtain from Figure 7, except that

- Online-screening only starts to be effective after certain number of epochs, which is already seen in Figure 2 (h).
- Only $w = 0.75$ is safe for this case, though the error decay of $w = 0.6$ is better than that of $w = 0.75$.

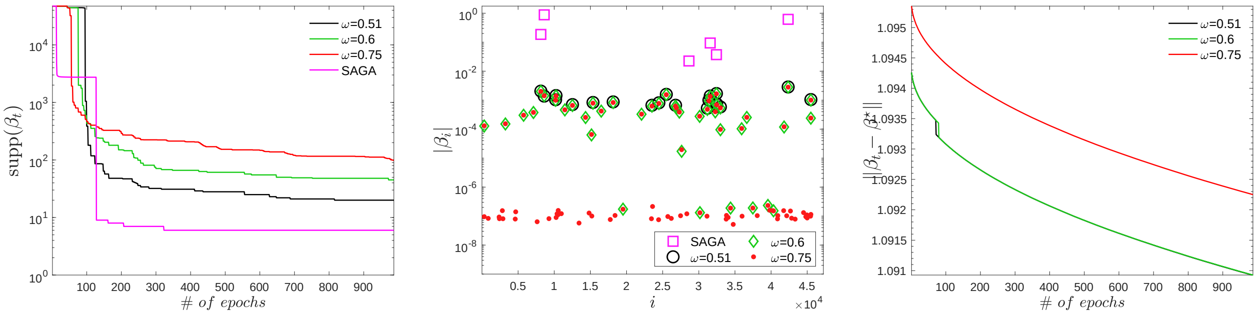
The wall clock time for these three choices of w are 325, 371 and 297 seconds, respectively. The reason that $w = 0.75$ has better wall clock time than other two, though it eventually provides largest support size, is that it has slightly more significant dimension reduction than the other two after 50'th epoch, see sub-figure (a) in Figure 8.

[3] We ran the method until convergence and checked its optimality condition, which showed that the output obtained by $w = 0.51$ is not a solution of the problem



(a) Support (b) Solution (c) Error

Figure 7: Comparison of exponent parameter w for LASSO problem and `dexter` dataset. Figure (a) shows how the support decays over number of epochs and figure (c) shows the error decay over the number of epochs. Figure (b) shows the final solutions to give a sense of the quality of solutions. Decreasing w leads to more aggressive screening behaviour. For this dataset, the choice of $w = 0.6$ and $w = 0.75$ are 'safe choices'.



(a) Support (b) Solution (c) Error

Figure 8: Comparison of exponent parameter w for SLR problem and `rcv1` dataset for online-screening.

5 Conclusion

Online optimization methods are widely used for solving large-scale problems arising from machine learning, data science and statistics. However, when combined with sparsity promoting regularizers, online methods can break the support identification property of these regularizers. In this paper, we combined the well established safe screening technique with online optimization methods which allows online methods to discard useless features along the iteration, hence achieving dimension reduction. Numerical result demonstrated that dramatic wall time gains can be achieved for classic regression tasks over real datasets.

A Appendix

A.1 Convex analysis

The sub-differential of a proper convex and lower semi-continuous function $\Omega : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is a set-valued mapping defined by

$$\partial\Omega : \mathbb{R}^n \rightrightarrows \mathbb{R}^n,\, \beta \mapsto \big\{ Z \in \mathbb{R}^n \,|\, \Omega(\beta') \geq \Omega(\beta) + \langle Z,\, \beta' - \beta\rangle,\ \forall \beta' \in \mathbb{R}^n \big\}. \tag{A.1}$$

Lemma A.1 (Descent lemma [2]). *Suppose that $F : \mathbb{R}^n \to \mathbb{R}$ is convex continuously differentiable and ∇F is L-Lipschitz continuous. Then, given any $\beta, \beta' \in \mathbb{R}^n$,*

$$F(\beta) \leq F(\beta') + \langle \nabla F(\beta'),\, \beta - \beta' \rangle + \tfrac{L}{2}\|\beta - \beta'\|^2.$$

Lemma A.2. *Fenchel-Young inequality Let $F : \mathbb{R}^n \to \mathbb{R}$ be convex, lower semicontinuous and proper, then for all $p, x \in \mathbb{R}^n$,*

$$F(x) + F^*(p) \geq \langle p,\, x \rangle$$

with equality if $p \in \partial F(x)$.

A.2 Derivation of the dual problem ($\mathcal{D}_\lambda$)

Write $f_y(z) \overset{\text{def}}{=} f(z; y)$ and $v_\beta(x, y) = f'_y(x^\top \beta)$, then applying the Fenchel-Young (in)equality twice,

$$\begin{aligned}\mathcal{P}(\beta) &= \mathbb{E}_{(x,y)}[v_\beta(x,y)x^\top\beta] - \mathbb{E}_{(x,y)}[f_y^*(v_\beta(x,y))] + \lambda\Omega(\beta)\\ &= -\mathbb{E}_{(x,y)}[f_y^*(v_\beta(x,y))] - \lambda\big(\langle -\tfrac{1}{\lambda}\mathbb{E}_{(x,y)}[v_\beta(x,y)x],\, \beta\rangle - \Omega(\beta)\big)\\ &\geq -\mathbb{E}_{(x,y)}[f_y^*(v_\beta(x,y))] - \lambda\Omega^*\big(-\tfrac{1}{\lambda}\mathbb{E}_{(x,y)}[v_\beta(x,y)x]\big)\end{aligned}$$

where the final inequality is an equality if $-\frac{1}{\lambda}\mathbb{E}_{(x,y)}[v_\beta(x,y)x] \in \partial\Omega(\beta)$, which is the case at the optimal point $\beta^\star$. Therefore, it follows that

$$\min_{\beta \in \mathbb{R}^n} \mathcal{P}(\beta) = -\mathbb{E}_{(x,y)}[f_y^*(v_{\beta^\star}(x,y))] - \lambda\Omega^*\big(-\tfrac{1}{\lambda}\mathbb{E}_{(x,y)}[v_{\beta^\star}(x,y)x]\big)$$

On the other hand, again by the Fenchel-Young inequality,

$$\min_\beta \mathcal{P}(\beta) \geq \max_v \mathcal{D}(v) \overset{\text{def}}{=} -\mathbb{E}_{(x,y)}[f_y^*(v(x,y))] - \lambda\Omega^*\big(-\tfrac{1}{\lambda}\mathbb{E}_{(x,y)}[v(x,y)x]\big)$$

where we maximize over Λ-measurable functions v. Therefore, the dual problem is $\max_v \mathcal{D}(v)$ and strong duality holds. Finally, note that Ω^* is the indicator function on the dual constraint set $\mathcal{K}_{\lambda,\eta}$.

A.3 Proofs of Section 3.1

We prove Propositions 3.4 and 3.5 in this section. The proofs are provided for completeness although they use standard techniques, see for example [14]. Similar results can be found in [19] (although we relax the condition of $\sum_t \mu_t^2\sqrt{t} < +\infty$ to simply $\sum_t \mu_t^2 < +\infty$). We make use of the following lemma.

Lemma A.3 (Super-martingale convergence [26]). *Let $\mathcal{F}_k$ be a set of random variables with $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$. Let Y_k, Z_k, W_k be non-negative random variables which are functions of random variables in $\mathcal{F}_k$, such that*

- $\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \leq Y_k + W_k - Z_k$
- $\sum_k W_k < +\infty$ *with probability 1*

Then, $\sum Z_k < +\infty$ and Y_k converges to a non-negative random variable Y with probability 1.

Lemma A.4. *For some* $\{\mu_n\}_n \subset (0,1)$ *and let* $\{f_n\}_n \subset \mathbb{R}^d$ *be random variables. Suppose* $\sum_j \mu_j = +\infty$ *and* $\sum_j \mu_j^2 < +\infty$. *Define* $\bar{f}_n \overset{\text{def}}{=} \mu_n f_n + (1-\mu_n)\bar{f}_{n-1}$ *with* $\bar{f}_1 = f_1$. *Let* $\eta_j^{(n)} \overset{\text{def}}{=} \mu_j \prod_{i=j+1}^n (1-\mu_i)$.

(i) $\bar{f}_n = \sum_{j=1}^n \eta_j^{(n)} f_j$

(ii) $\sum_j \eta_j^{(n)} = 1$

(iii) $\lim_{n\to+\infty} \sum_{j=1}^n (\eta_j^{(n)})^2 = 0$

(iv) Suppose that $\mathbb{E}[f_j] = \mathbb{E}[f_1]$, $\{f_j\}$ *are iid random variables and* $\|f_j\|_\infty \leq B$. *Then,* $\lim_{n\to+\infty} \bar{f}_n = \mathbb{E}[f_1]$ *with probability 1.*

Proof. The first two statements are a simple computation. For the third statement,

$$\sum_{j=1}^n (\eta_j^{(n)})^2 = \sum_{j=1}^n \mu_j^2 \prod_{i=j+1}^n (1-\mu_i)^2 \leq \sum_{j=m}^n \mu_j^2 + \prod_{i=m}^n (1-\mu_i)^2 \sum_{j=1}^{m-1} \mu_j^2$$

Since $\sum_j \mu_j^2 < +\infty$, we have $\sum_{j=m}^n \mu_j^2 \to 0$ as $m, n \to +\infty$. Moreover, since $\sum_{i=m}^n \mu_i \to +\infty$ as $n \to +\infty$, we have

$$\prod_{i=m}^n (1-\mu_i)^2 \leq \prod_{i=m}^n \exp(-2\mu_i) = \exp\left(-2\sum_{i=m}^n \mu_i\right) \to 0$$

as $n \to +\infty$.

Finally, fix $j \in [d]$ and $n \in \mathbb{N}$. For $m \leq n$, define $Y_m \overset{\text{def}}{=} \sum_{j=1}^m \eta_j^{(n)}(f_j - (\mathbb{E}[f_1])_j)$. Then, $\{Y_m\}_{m\leq n}$ is a Martingale and $|Y_m - Y_{m-1}| \leq 2\eta_m^{(n)} B$. By Azuma-Hoeffding inequality, given any $v > 0$,

$$\mathbb{P}\left(|Y_n| \geq v\right) \leq 2\exp\left(-\frac{v^2}{8B^2 \sum_{j=1}^n (\eta_j^{(n)})^2}\right)$$

and hence, by the union bound,

$$\mathbb{P}\left(\|\bar{f}_n - \mathbb{E}[f_1]\|_\infty \geq v\right) \leq 2d\exp\left(-\frac{v^2}{8B^2 \sum_{j=1}^n (\eta_j^{(n)})^2}\right) \to 0, \qquad n \to +\infty,$$

so $\bar{f}_n$ converges to $\mathbb{E}[f_1]$ in probability. To show that it converges almost surely, we make use of Lemma A.3: Let $Y_n \overset{\text{def}}{=} |\bar{f}_n - \mathbb{E}[f_1]|^2$. Note that

$$\begin{aligned} Y_n &= |(1-\mu_n)(\bar{f}_{n-1} - \mathbb{E}[f_1]) + \mu_n(f_n - \mathbb{E}[f_1])|^2 \\ &= (1-\mu_n)^2 Y_{n-1} + \mu_n^2 (f_n - \mathbb{E}[f_1])^2 + (1-\mu_n)\mu_n(\bar{f}_{n-1} - \mathbb{E}[f_1])(f_n - \mathbb{E}[f_1]). \end{aligned}$$

So, taking expectation with respect to $\{f_j\}_{j=1}^{n-1}$, it follows that

$$\begin{aligned} \mathbb{E}_{n-1}[Y_n] &= (1-\mu_n)^2 Y_{n-1} + \mu_n^2 \mathbb{E}_{n-1}[(f_n - \mathbb{E}[f_1])^2] \\ &\leq Y_{n-1} + 4B^2\mu_n^2. \end{aligned}$$

Since $\sum_n \mu_n^2 < +\infty$, it follows from Lemma A.3 that Y_n converges almost surely, and this converges almost surely to 0 since Y_n converges to 0 in probability. In particular, $\bar{f}_n$ converges to $\mathbb{E}[f_1]$ almost surely. □

Now we are ready to prove the result of Proposition 3.4.

Proof of Proposition 3.4. To prove (i), first note that for each β, with probability 1, $|\bar{P}^{(t)}(\beta) - P(\beta)| \to 0$ as $t \to +\infty$, by (iv) of Lemma A.4.

Note that $\{\bar{P}^{(t)}\}_t$ is equi-continuous: By the mean value theorem there exists $\{\xi_s\}_{s=1}^t \subset \mathcal{B}_R$, where we denote by $\mathcal{B}_R$ the ball of radius R with $R \overset{\text{def}}{=} \sup_{x\in\mathcal{X},\beta\in\mathcal{O}} |\langle x,\, \beta\rangle|$, such that

$$|\bar{P}^{(t)}(\beta) - \bar{P}^{(t)}(\beta')| = |\textstyle\sum_{s=1}^t \eta_s^{(t)} f'_{y_s}(\xi_s) x_s^\top (\beta - \beta')|$$

For some fixed point $\xi_0 \in \mathcal{D}$,

$$\begin{aligned}|\bar{P}^{(t)}(\beta) - \bar{P}^{(t)}(\beta')| &\le |\textstyle\sum_{s=1}^t \eta_s^{(t)} (f'_{y_s}(\xi_s) - f'_{y_s}(\xi_0)) x_s^\top (\beta - \beta')| \\ &\quad + |\textstyle\sum_{s=1}^t \eta_s^{(t)} f'_{y_s}(\xi_0) x_s^\top (\beta - \beta')| \\ &\le L \sum_{s=1}^t \eta_s^{(t)} |\xi_s - \xi_0| \|x_s\| \|\beta - \beta'\| \\ &\quad + |\big(\textstyle\sum_{s=1}^t \eta_s^{(t)} f'_{y_s}(\xi_0) x_s\big)^\top (\beta - \beta')|\end{aligned}$$

where we have used that fact that f'_{y_s} is L-Lipschitz. By boundedness of $\mathcal{B}_R$ and $\mathcal{X}$, there exists B such that

$$L \sum_{s=1}^t \eta_s^{(t)} |\xi_s - \xi_0| \|x_s\| \|\beta - \beta'\| \le B \|\beta - \beta'\| \sum_{s=1}^t \eta_s^{(t)} = B \|\beta - \beta'\|.$$

By Lemma A.4 (iv), we know that with probability 1,

$$\sum_{s=1}^t \eta_s^{(t)} f'_{y_s}(\xi_0) x_s \to \mathbb{E}[f'_y(\xi_0) x], \qquad t \to +\infty.$$

Therefore, $\| \sum_{s=1}^t \eta_s^{(t)} f'_{y_s}(\xi_0) x_s\|$ is uniformly bounded, and there exists a constant $C > 0$ such that, with probability 1,

$$|\bar{P}^{(t)}(\beta) - \bar{P}^{(t)}(\beta')| \le C \|\beta - \beta'\|$$

Hence, by Arzela-Ascoli (see [24, Theorem 11.3.2] or [1]), it follows that $\bar{P}^{(t)}$ converges uniformly to P on compact sets.

To prove (ii), let $\beta^\star$ be a minimiser of P, then

$$\begin{aligned}P(\beta^{(t),*}) - P(\beta^\star) &= P(\beta^{(t),*}) - \bar{P}^{(t)}(\beta^\star) + \bar{P}^{(t)}(\beta^\star) - P(\beta^\star) \\ &\le P(\beta^{(t),*}) - \bar{P}^{(t)}(\beta^{(t),*}) + \bar{P}^{(t)}(\beta^\star) - P(\beta^\star)\end{aligned}$$

By (i), the right hand side converges to zero as $t \to +\infty$. Therefore, $\{\beta^{(t),*}\}$ is a minimising sequence. By compactness of the sublevel sets of Ω, there exists a convergent subsequence, and the limit of this subsequence is a minimiser since Ω is lower-semicontinuous and $f(\cdot, y_i)$ is continuous.

To prove (iii), denote $Z^{(t)}(\beta) \overset{\text{def}}{=} -\frac{1}{\lambda} \sum_{s=1}^t \eta_s^{(t)} f'(x_s^\top \beta, y_s) x_s$. Note that given $\beta^{(t)} \in \operatorname{argmin}_\beta \bar{P}^{(t)}(\beta)$, $Z^{*,(t)} = Z^{(t)}(\beta^{(t)})$. By the triangle inequality,

$$\begin{aligned}\|Z^{(t)}(\beta^{(t)}) - \mathbb{E}[-\tfrac{1}{\lambda} f'(x^\top \beta^\star, y) x]\|_\infty &\le \|Z^{(t)}(\beta^{(t)}) - Z^{(t)}(\beta^*)\|_\infty \\ &\quad + \|Z^{(t)}(\beta^*) - \mathbb{E}[-\tfrac{1}{\lambda} f'(x^\top \beta^\star, y) x]\|_\infty\end{aligned}$$

By Lemma A.4 (iv), we have that $\|Z^{(t)}(\beta^*) - \mathbb{E}[-\frac{1}{\lambda}f'(x^\top\beta^\star, y)x]\|_\infty \to 0$ as $t \to +\infty$. To bound the first term on the RHS, let $\theta_s^{(t)} = f'(x_s^\top\beta^{(t)}, y_s)$ and $\theta_s^* = f'(x_s^\top\beta^\star, y_s)$. Note that by strong convexity of $\bar{D}^{(t)}$ (c.f. the proof of Lemma 3.6), we have

$$\frac{1}{2L}\sum_{s=1}^{t}\eta_s^{(t)}|\theta_s^{(t)} - \theta_s^*|^2 \le \bar{D}^{(t)}(\theta^{(t)}) - \bar{D}^{(t)}(\theta^*) \le \bar{P}^{(t)}(\beta^\star) - \bar{D}^{(t)}(\beta^\star),$$

which again converges to zero by Lemma A.4 (iv) and optimality of $\beta^\star$.

□

Proof of Proposition 3.5. Let $\mathcal{F}_t$ be the σ-algebra generated by $\{(x_s, y_s)\}_{s\le t}$. Observe that

$$\begin{aligned}-\frac{1}{\lambda}\sum_{s=1}^{t}\eta_s^{(t)}\theta_s x_s - Z^\star &= \frac{-1}{\lambda}\sum_{s=1}^{t}\eta_s^{(t)}(f'_{y_s}(x_s^\top\beta_s)x_s - \mathbb{E}[f'_{y_s}(x_s^\top\beta_s)x_s|\mathcal{F}_{s-1}])\\ &\quad + \sum_{s=1}^{t}\eta_s^{(t)}\big(\tfrac{-1}{\lambda}\mathbb{E}[f'_{y_s}(x_s^\top\beta_s)x_s|\mathcal{F}_{s-1}] - Z^\star\big)\\ &= \frac{1}{\lambda}\sum_{s=1}^{t}\eta_s^{(t)}(f'_{y_s}(x_s^\top\beta_s)x_s - \mathbb{E}[f'_{y_s}(x_s^\top\beta_s)x_s|\mathcal{F}_{s-1}])\\ &\quad + \sum_{s=1}^{t}\eta_s^{(t)}\big(\tfrac{-1}{\lambda}\mathbb{E}[f'_{y}(x^\top\beta_s)x] - Z^\star\big).\end{aligned}$$

Therefore,

$$\|-\tfrac{1}{\lambda}\textstyle\sum_{s=1}^{t}\eta_s^{(t)}\theta_s x_s - Z^\star\|_\infty \le \tfrac{1}{\lambda}\|\textstyle\sum_{s=1}^{t}\eta_s^{(t)}z_s\|_\infty + \displaystyle\frac{CL}{\lambda}\sum_{s=1}^{t}\eta_s^{(t)}\|\beta_s - \beta^\star\|$$

where $z_s \overset{\text{def}}{=} f'_{y_s}(x_s^\top\beta_s)x_s - \mathbb{E}[f'_{y_s}(x_s^\top\beta_s)x_s|\mathcal{F}_{s-1}]$, and the constant $C > 0$ comes from the assumption that $x \in \mathcal{X}$ which is compact.

Fix $k \in [n]$, and define

$$Y_n^k \overset{\text{def}}{=} \sum_{s=1}^{n}\eta_s^{(t)}(z_s)_k$$

This is a martingale with bounded difference $|Y_n^k - Y_{n-1}^k| = \eta_n^{(t)}|(z_n)_k| \le 2B\eta_n^{(t)}$. By Azuma-Hoeffding inequality,

$$\mathbb{P}\big(|Y_t^k| \ge v\big) \le 2\exp\Big(-\frac{v^2}{8B^2\sum_{s=1}^{t}(\eta_s^{(t)})^2}\Big)$$

Therefore, by the union bound,

$$\mathbb{P}\big(\max_k |Y_t^k| \ge v\big) \le 2n\exp\Big(-\frac{v^2}{8B^2\sum_{s=1}^{t}(\eta_s^{(t)})^2}\Big).$$

The RHS converges to 0 as $t \to +\infty$ by (iii) of Lemma A.4.

Finally, $\sum_{s=1}^{t}\eta_s^{(t)}\|\beta_s-\beta^\star\|\to 0$ provided that $\|\beta_s-\beta^\star\|\to 0$: For any $m\le t$:

$$\sum_{s=1}^{t}\eta_s^{(T)}\|\beta_s-\beta^\star\| = \sum_{s=1}^{m}\eta_s^{(t)}\|\beta_s-\beta^\star\| + \sum_{s=m+1}^{t}\eta_s^{(t)}\|\beta_s-\beta^\star\|.$$

For any $\epsilon>0$, there exists m such that the second term on the RHS is at most $\epsilon/2$ since $\sum_s \eta_s^{(t)}=1$ and $\|\beta_s-\beta^\star\|\to 0$ as s increases. For the first term on the RHS, since $\|\beta_s-\beta^\star\|$ is uniformly bounded and for any fixed m, $\sum_{s=1}^{m}\eta_s^{(t)}\to 0$ as $t\to+\infty$, it is clear that for any $\epsilon>0$, there exists t_0 such that for all $t\ge t_0$, $\sum_{s=1}^{t}\eta_s^{(t)}\|\beta_s-\beta^\star\|\to 0$. It therefore follows that $\bar Z_t - Z^\star$ converges to 0 in probability. To conclude almost sure convergence, we apply Lemma A.3. Let $Y_t=|\bar Z_t - Z^\star|^2$, then,

$$\begin{aligned}\mathbb{E}_{t-1}[Y_t] = (1-\mu_t)Y_{t-1} + \mu_t^2\mathbb{E}_{t-1}\big(-\tfrac{1}{\lambda}\theta_t x_t - Z^\star\big)^2\\ + (1-\mu_t)\mu_t Y_{t-1}\mathbb{E}_{t-1}\big(-\tfrac{1}{\lambda}\theta_t x_t - Z^\star\big).\end{aligned}$$

Recall that $Z^\star = -\frac{1}{\lambda}\mathbb{E}_{(x,y)}[f'(x^\top\beta^\star,y)x]$ and $\|\beta_s\|$ is uniformly bounded, so there exists $B,B'>0$ such that

$$\begin{aligned}\mathbb{E}_{t-1}[Y_t] &\le (1-\mu_t)^2Y_{t-1} + \mu_t^2 B\\ &\quad + (1-\mu_t)\mu_t Y_{t-1}\frac{1}{\lambda}\mathbb{E}_{(x,y)}[|(f'(x^\top\beta_t,y) - f'(x^\top\beta^\star,y))x|]\\ &\le (1-\mu_t)^2Y_{t-1} + \mu_t^2 B + (1-\mu_t)\mu_t Y_{t-1}\|\beta_t-\beta^\star\|\\ &= (1-\mu_t)(1-\mu_t+\mu_t\|\beta_t-\beta^\star\|)Y_{t-1} + \mu_t^2 B.\end{aligned}$$

Since $\beta_t\to\beta^\star$, for t sufficiently large, $1-\mu_t+\mu_t\|\beta_t-\beta^\star\|\in(0,1)$, so we can apply Lemma A.3 to conclude that Y_t converges almost surely to 0.

For (ii), let $\beta^\star$ be a minimizer of $\mathcal{P}$. We can establish in the same way as above that

$$\sum_{s=1}^{t}\eta_s f_{y_s}^*(\theta_s)\to \mathbb{E}[f_y^*(f_y'(x^\top\beta^\star))], \qquad t\to+\infty,$$

and

$$|\bar P^{(t)}(\bar\beta_t) - \mathcal{P}(\beta^\star)| \le |\bar P^{(t)}(\bar\beta_t) - \mathcal{P}(\bar\beta_t)| + |\mathcal{P}(\bar\beta_t) - \mathcal{P}(\beta^\star)|$$

which converges to 0 as $t\to+\infty$ since we have uniform convergence of $\bar P^{(t)}$ to $\mathcal{P}$ by Proposition 3.4 and $\bar\beta_t\to\beta^\star$. Finally, since $-\frac{1}{\lambda}\sum_{s=1}^{t}\eta_s^{(t)}\theta_s x_s\to Z^\star$, and Ω^* is lower semicontinuous, we have that

$$\liminf_{t\to+\infty}\Omega^*\big(-\frac{1}{\lambda}\sum_{s=1}^{t}\eta_s^{(t)}\theta_s x_s\big)\ge \Omega^*(Z^\star).$$

Therefore,

$$\lim_{t\to+\infty}\bar P^{(t)}(\bar\beta_t) - \bar D^{(t)}((\theta_s)_{s\le t}) \le \mathcal{P}(\beta^\star) - \mathbb{E}[f_y^*(f_y'(x^\top\beta^\star))] - \Omega^*(Z^\star) = 0,$$

and we conclude the proof. □

A.4 Proofs for Section 3.2

Proof of Lemma 3.6. Since f_{y_s} is L-Lipschitz smooth, it follows that $f^*_{y_s}$ is $1/L$- strongly convex,

$$\begin{aligned}\frac{1}{2L}|\theta_s - \theta_s^{(t),*}|^2 &\le -f^*_{y_s}(\theta_s^{(t),*}) + f^*_{y_s}(\theta_s) - (f^*_{y_s})'(\theta_s^{(t),*})(\theta_s - \theta_s^{(t),*}) \\ &\le -f^*_{y_s}(\theta_s^{(t),*}) + f^*_{y_s}(\theta_s) - (f^*_{y_s})'(\theta_s^{(t),*})(a_t\theta_s - \theta_s^{(t),*}) \\ &\quad - (1-a_t)(f^*_{y_s})'(\theta_s^{(t),*})\theta_s\end{aligned} \tag{A.2}$$

where $a_t = \min(1, 1/\Omega^D(\bar{Z}_t))$ is such that $a_t(\theta_s)_{s\le t} \in \mathcal{K}_{\lambda,\eta^{(t)}}$, the dual constraint set defined in (2.4). Since $\theta^{(t),*}$ is a dual optimal point, by Fermat's rule

$$-\sum_{s\le t} \eta_s^{(t)} (f^*_{y_s})'(\theta_s^{(t),*})(a_t\theta_s - \theta_s^{(t),*}) \le 0.$$

Therefore, multiplying (A.2) by $\eta_s^{(t)}$ and summing from $s = 1, \dots, t$, we obtain

$$\begin{aligned}\frac{1}{2L}\sum_{s=1}^t \eta_s^{(t)}|\theta_s - \theta_s^{(t),*}|^2 \le &\sum_{s=1}^t \eta_s^{(t)}\left(-f^*_{y_s}(\theta_s^{(t),*}) + f^*_{y_s}(\theta_s)\right) \\ &- (1-a_t)\sum_{s=1}^t \eta_s^{(t)}(f^*_{y_s})'(\theta_s^{(t),*})\theta_s.\end{aligned} \tag{A.3}$$

Note that by optimality of $\theta^{(t),*}$, $\sum_{s=1}^t \eta_s^{(t)}\left(-f^*_{y_s}(\theta_s^{(t),*})\right) \le \bar{P}^{(t)}(\beta)$ for all $\beta \in \mathbb{R}^d$. Moreover, letting $\beta^{(t),*} \in \operatorname{argmin}_\beta \bar{P}^{(t)}(\beta)$, we have $\theta_s^{(t),*} = f'_{y_s}(x_s^\top \beta^{(t),*})$, and using the fact that $(f^*_{y_s})' \circ f'_{y_s} = \mathrm{Id}$, we have

$$\sum_{s=1}^t \eta_s^{(t)}(f^*_{y_s})'(\theta_s^{(t),*})\theta_s = \sum_{s=1}^t \eta_s^{(t)}(f^*_{y_s})'(f'_{y_s}(x_s^\top\beta^{(t),*}))\theta_s = (\sum_{s=1}^t \eta_s^{(t)}\theta_s x_s)^\top \beta^{(t),*}.$$

By optimality of $\beta^{(t),*}$, we have

$$\Omega(\beta^{(t),*}) \le \frac{1}{\lambda}\sum_{s=1}^t \eta_s^{(t)} f_{y_s}(0).$$

Plugging these estimates back into (A.3) yields, for any $\beta \in \mathbb{R}^n$,

$$\frac{1}{2L}\sum_{s=1}^t \eta_s^{(t)}|\theta_s - \theta_s^{(t),*}|^2 \le \mathrm{Gap}_t(\beta) + \sum_{s=1}^t \eta_s^{(t)} f_{y_s}(0)\ (\Omega^D(\bar{Z}_t) - 1)_+ \tag{A.4}$$

since $(1-a_t)\Omega^D(\bar{Z}_t) = (\Omega^D(\bar{Z}_t) - 1)_+$. Finally, by the Cauchy-Schwarz inequality,

$$\begin{aligned}&\Omega_g^D\left(\tfrac{1}{\lambda}\textstyle\sum_{s=1}^t \eta_s^{(t)}(\theta_s x_s - \theta_s^{(t),*}x_s)\right) \\ &\quad= \frac{1}{\lambda}\sup_{\Omega_g(z)\le 1} \langle \textstyle\sum_{s=1}^t \eta_s^{(t)}(\theta_s x_s - \theta_t^{(t),*}x_s),\, z\rangle \\ &\quad\le \frac{1}{\lambda}\sqrt{\textstyle\sum_{s=1}^t \eta_s^{(t)}|\theta_s - \theta_s^{(t),*}|^2}\sup_{\Omega_g(z)\le 1}\sqrt{\textstyle\sum_{s=1}^t \eta_s^{(t)}|\langle x_s,\, z\rangle|^2} \\ &\quad\le \frac{1}{\lambda}\sqrt{\textstyle\sum_{s=1}^t \eta_s^{(t)}|\theta_s - \theta_s^{(t),*}|^2}\sqrt{\textstyle\sum_{s=1}^t \eta_s^{(t)}\Omega_g^D(x_s)^2}\end{aligned}$$

and the result follows by combining this with (A.4). □

References

[1] D. WK Andrews. Generic uniform convergence. *Econometric theory*, 8(2):241–257, 1992.

[2] Runxue Bao, Bin Gu, and Heng Huang. Fast oscar and owl regression via safe screening rules. In *International Conference on Machine Learning*, pages 653–663. PMLR, 2020.

[3] D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[4] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. *IEEE Transactions on Signal Processing*, 63(19):5121–5132, 2015.

[5] L. Bottou and Y. L. Cun. Large scale online learning. In *Advances in neural information processing systems*, pages 217–224, 2004.

[6] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[7] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[8] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

[9] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

[10] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[11] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[12] J. Kiefer, J. Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[13] Quentin Klopfenstein, Quentin Bertrand, Alexandre Gramfort, Joseph Salmon, and Samuel Vaiter. Model identification and local linear convergence of coordinate descent. *arXiv preprint arXiv:2010.11825*, 2020.

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

[15] S. Lee and S. J. Wright. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(Jun):1705–1744, 2012.

[16] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002.

[17] A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, 2016.

[18] J. Liang, J. Fadili, and G. Peyré. Activity identification and local linear convergence of Forward–Backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017.

[19] J. Liu, Z. Zhao, J. Wang, and J. Ye. Safe screening with variational inequalities and its application to lasso. In *International Conference on Machine Learning*, pages 289–297, 2014.

[20] J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pages 2283–2291, 2013.

[21] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *Advances in neural information processing systems*, pages 811–819, 2015.

[22] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparse-group lasso. In *Advances in Neural Information Processing Systems*, pages 388–396, 2016.

[23] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparsity enforcing penalties. *The Journal of Machine Learning Research*, 18(1):4671–4703, 2017.

[24] C. Poon, J. Liang, and C.-B. Schoenlieb. Local convergence properties of saga/prox-svrg and acceleration. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4124–4132. PMLR, 2018.

[25] Suhasini Subba Rao. A course in time series analysis.

[26] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[27] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[28] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

[29] Y. Sun and F. Bach. Safe screening for the generalized conditional gradient method. *arXiv preprint arXiv:2002.09718*, 2020.

[30] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[31] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

[32] Samuel Vaiter, Mohammad Golbabaee, Jalal Fadili, and Gabriel Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3):230–287, 2015.

[33] J. Wang, J. Zhou, J. Liu, P. Wonka, and J. Ye. A safe screening rule for sparse logistic regression. In *Advances in neural information processing systems*, pages 1053–1061, 2014.

[34] J. Wang, J. Zhou, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. In *Advances in neural information processing systems*, pages 1070–1078, 2013.

[35] Z. J. Xiang, Y. Wang, and P. J. Ramadge. Screening tests for lasso problems. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):1008–1027, 2016.

[36] Z. J Xiang, H. Xu, and P. J. Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. In *Advances in neural information processing systems*, pages 900–908, 2011.

[37] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

[38] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.