WHInter: A Working set algorithm for High-dimensional sparse second order Interaction models.

Marine Le Morvan [1 2 3] **Jean-Philippe Vert** [1 2 3 4]

Abstract

Learning sparse linear models with two-way interactions is desirable in many application domains such as genomics. ℓ_1-regularised linear models are popular to estimate sparse models, yet standard implementations fail to address specifically the quadratic explosion of candidate two-way interactions in high dimensions, and typically do not scale to genetic data with hundreds of thousands of features. Here we present WHInter, a working set algorithm to solve large ℓ_1-regularised problems with two-way interactions for binary design matrices. The novelty of WHInter stems from a new bound to efficiently identify working sets while avoiding to scan all features, and on fast computations inspired from solutions to the maximum inner product search problem. We apply WHInter to simulated and real genetic data and show that it is more scalable and two orders of magnitude faster than the state of the art.

1. Introduction

In application domains where the number of features exceeds the number of available samples, sparsity-inducing regularisers have a long history of success. Genomic prediction of complex phenotypes, biomedical imaging, astronomy or finance are a few examples. In particular the least squares with ℓ_1 regularisation, known as the LASSO (Tibshirani, 1996), has been extensively studied. It enjoys desirable statistical properties, since the number of samples required for exact support recovery of a sparse model scales as the logarithm of the number of features, under some assumptions (Wainwright, 2009). It also enjoys practical advantages, notably the interpretability of the learned models and the availability of fast solvers.

Indeed, a lot of research effort has been devoted to accelerating solvers for sparsity constrained problems in high dimension. A central idea is to exploit the sparsity of the solution to develop algorithms that do not spend too much time on optimising coefficients that will end up being 0. For example, safe screening rules identify features which are guaranteed to be inactive at the optimum so that their corresponding coefficients can be safely zeroed and set aside from the pool of coefficients to update (El Ghaoui et al., 2012; Xiang et al., 2011; Xiang & Ramadge, 2012; Fercoq et al., 2015; Wang et al., 2013; Raj et al., 2016). Dynamic screening rules (Bonnefoy et al., 2015) such as the GAP safe rules (Fercoq et al., 2015) are particularly useful since more and more coefficients can be safely zeroed while the solver approaches the optimal solution. In spite of this, safe rules tend to be conservative, thereby limiting the potential speed-up. To remedy this drawback, new working set heuristics have been proposed. Working set algorithms enjoy great success in practice, as exemplified by the popular GLMNET package (Friedman et al., 2010). They iteratively solve subproblems, either problems restricted to a subset of features in the primal or to a subset of constraints in the dual, until convergence. Working set methods allow to focus coefficient updates on a set of features which can be significantly smaller than that yielded by safe rules. However this comes at a cost, that of checking the optimality conditions for all features at each iteration. BLITZ (Johnson & Guestrin, 2015) is a recently proposed working set algorithm that has been shown to have state-of-the-art performance for ℓ_1 regularised problems. Interestingly, the choice of the working sets in BLITZ can be seen as an aggressive use of the GAP safe rules (as noted in Massias et al., 2017) where the size of the working set is chosen to maximise the progress towards convergence. BLITZ can therefore be combined with the GAP safe rules (or the FLEX constraint elimination according to Johnson et al. terminology) at no cost. A direct comparison between BLITZ and the GAP safe rules by Ndiaye et al. (2017) illustrates the effectiveness of the working set approach. Further developments have also focused on coordinate descent (CD) to avoid wasteful

[1]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, 75006 Paris, France [2]Institut Curie, PSL Research University, 75005 Paris, France [3]INSERM, U900, 75005 Paris, France [4]Ecole Normale Supérieure, Department of Mathematics and Applications, 75005 Paris, France. Correspondence to: Jean-Philippe Vert <jean-philippe.vert@mines-paristech.fr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

coordinate updates, which represent most of the time spent by the solver (Fujiwara et al., 2016; Johnson & Guestrin, 2017).

The problem of fitting sparse linear models with two-way interactions has also attracted attention during the past decade. By two-way interactions we mean the entry-wise multiplication between two features; this is for example important in genomics to detect possible epistasis between genes. Surprisingly, very few of these works have links with the aforementioned literature. A majority of them focus on the design of sparsity-inducing penalties which enforce heredity assumptions and apply to moderate-dimensional settings ($p < 1,000$) (Radchenko & James, 2010; Bien et al., 2013; Lim & Hastie, 2015; Haris et al., 2016). Heredity assumptions state that an interaction can be included in the model only if one or both of its corresponding main effects are included. We note however that `glinternet` (Lim & Hastie, 2015) was applied to higher dimensional problems and in particular to a dataset with roughly $p = 27,000$ main effects, although the size of the learned model is not specified and the running time for the experiment is not reported by the authors. Interestingly, `glinternet` uses an active set strategy. Comparatively few works have been devoted to learning sparse regression models with interactions when the number of interactions is higher. Most of them are heuristics which start by selecting main effects and then incorporate interactions generated under the heredity constraint in a possibly iterative fashion. The simplest form of such heuristics consists in fitting a sparse linear model with the main effects only, and then fitting a second sparse linear model on all previously selected main effects and their interactions. This has been used in practice for example by Wu et al. (2009). Iterative refinements have been proposed where the LASSO is fit several times, and each time the set of candidate interactions considered is updated either by subsets, with the interactions between the K most relevant main effects selected at the previous fit (Bickel et al., 2010), or in a greedy fashion, where new interactions are included in the model as soon as a new main effect enters the LASSO path (Shah, 2016). In a similar vein, Hao & Zhang (2014) is based on a greedy model selection procedure instead of several LASSO fits. While these heuristics can deal with higher-dimensional problems than previous methods and enjoy some desirable statistical properties, they do not provide exact solutions and do not enjoy statistical properties as strong as those of the LASSO estimator.

An interesting link between the literature on interactions and that of solver acceleration with sparsity inducing norms has been made recently by Nakagawa et al. (2016). In the case where variables are binary or with values in $[0, 1]$, they propose an approach called Safe Pattern Pruning (SPP) which is able to provide the optimal solution of the LASSO with two-way interactions (or possibly higher-order interactions) for fairly high-dimensional problems, with no heredity constraint. Typically, for a problem with 1,000 samples and 10,000 main effects where two-way interactions are considered, SPP can provide solutions for a grid of regularisation parameters within one or two hours on a laptop with one core. SPP relies on the recently developed GAP safe screening rules. More precisely, the authors propose a safe pattern pruning criterion that can safely discard subsets of interactions from the model to speed up convergence. The performance of SPP is however hindered by several factors. One of them is that safe screening rules can be quite conservative even in the sequential setting. This property is inherited and amplified by the SPP criterion which can lead to heavy computations.

Inspired by SPP and the acceleration of solvers for sparsity constrained problems we propose a scalable algorithm, WHInter, to compute the optimal solution of ℓ_1-regularised linear problems with two-way interactions. WHInter is a working set method that efficiently delineates working sets among all interactions and main effects thanks to two contributions. First, we introduce a cheap and effective bound to rule out subsets of interactions that are guaranteed to be outside of the working set. Second, the identification of the working set among the remaining features is cast as a variant of the Maximum Inner Product Search (MIPS) problem to alleviate the corresponding computational load. We find that WHInter is up to two orders of magnitude faster than SPP. For example, a problem with roughly 700 samples and 100,000 main effects can be solved for a grid of regularisation parameters in half an hour on a laptop with one core compared to more than 30 hours with SPP. This improvement in the scalability opens up new horizons in several application fields. The rest of the paper is organised as follows. In Section 2, we present useful knowledge and notations used throughout the paper. In Section 3 we describe in details our algorithm and our main contributions. In Section 4, we evaluate WHInter on simulated datasets and finally in Section 5, we report results on a toxicogenomics prediction task.

2. Preliminaries

2.1. Setting and notations

For any integer $d \in \mathbb{N}$, we note $[\![d]\!] = \{1, \dots, d\}$ and $\mathbf{1}_d \in \mathbb{R}^d$ the d-dimensional vector of 1's. For any vector $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_d) \in \mathbb{R}^d$, we note $\| \mathbf{u} \|_1 = \sum_{i=1}^d | \mathbf{u}_i |$, $\| \mathbf{u} \|_2 = \left(\sum_{i=1}^d \mathbf{u}_i^2 \right)^{1/2}$, $\text{supp}(\mathbf{u}) = \{ i \in [\![d]\!] : \mathbf{u}_i \neq 0 \}$ and $\| \mathbf{u} \|_0 = | \text{supp}(\mathbf{u}) |$. For any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, $\mathbf{u} \odot \mathbf{v}$ is the vector of entry-wise products, i.e., $(\mathbf{u} \odot \mathbf{v})_i := \mathbf{u}_i \mathbf{v}_i$ for $i = 1, \dots, d$. For any matrix $\mathbf{M}$, we denote by $\mathbf{M}_{i,j}$ its (i, j)-th entry, $\mathbf{M}_j$ its j-th column and by $\boldsymbol{m}_i$ its i-th row. For any $\mathbf{u} \in \mathbb{R}^d$ and $\mathcal{I} \subset [\![d]\!]$, $\mathbf{u}_\mathcal{I} = (\mathbf{u}_i)_{i \in \mathcal{I}}$,

and similarly, if $\mathbf{M}$ is a matrix with d columns, $\mathbf{M}_{\mathcal{I}}$ is the sub-matrix with $|\mathcal{I}|$ columns $\mathbf{M}_{\mathcal{I}} = (\mathbf{M}_i)_{i\in\mathcal{I}}$.

Throughout the text we consider a design matrix $\mathbf{X} \in \{0,1\}^{n\times p}$ corresponding to n samples and p binary features, together with a response vector $\mathbf{y} \in \mathbb{R}^n$. We define an expanded design matrix $\mathbf{Z} \in \{0,1\}^{n\times D}$, with $D = p(p+1)/2$, which contains all p features from $\mathbf{X}$ plus the $p(p-1)/2$ interaction features. For clarity purposes, we define a symmetric indexing function $\tau : [\![p]\!]^2 \mapsto [\![D]\!]$ that uniquely assigns to every main effect and interaction an index in the expanded matrix $\mathbf{Z}$ such that $\mathbf{Z}_{\tau(j,k)} = \mathbf{Z}_{\tau(k,j)} := \mathbf{X}_j \odot \mathbf{X}_k$. In particular $\mathbf{Z}_{\tau(i,i)} = \mathbf{X}_i \odot \mathbf{X}_i = \mathbf{X}_i$ represents the i^{th} main effect. Since $\mathbf{X}$ is a binary matrix, the interaction feature $\mathbf{X}_j \odot \mathbf{X}_k$ corresponds to a logical AND between features $\mathbf{X}_i$ and $\mathbf{X}_j$. We organise the main effects and interactions in a simple tree as depicted in Figure 1 so as to reflect the property that $\forall (j,k) \in [\![p]\!]^2, \mathbf{Z}_{\tau(j,k)} \leq \mathbf{X}_j$ and $\mathbf{Z}_{\tau(j,k)} \leq \mathbf{X}_k$. In the sequel, the set composed of a main effect and its interactions with all other main effects will be referred to as *a branch* and for any $j \in [\![p]\!]$, we note branch(j) = $\{\tau(j,k) : k \in [\![p]\!]\}$.

We consider the convex optimisation problem:

$$\min_{(\mathbf{w},b)\in\mathbb{R}^D\times\mathbb{R}} P_{\mathbf{Z},\lambda}(\mathbf{w},b)\,, \tag{1}$$

with

$$\begin{aligned} P_{\mathbf{Z},\lambda}(\mathbf{w},b) &= F\left(\mathbf{Z}\mathbf{w} + b\mathbf{1}_n\right) + \lambda \left\|\mathbf{w}\right\|_1 \\ &= \sum_{i=1}^{n} f_i\left(\mathbf{z}_i\mathbf{w} + b\right) + \lambda \left\|\mathbf{w}\right\|_1 \,, \end{aligned}$$

where $\lambda > 0$ is a regularisation parameter and, for any $i \in [\![n]\!]$, $f_i : \mathbb{R} \mapsto [-\infty, +\infty]$ is a loss function parametrised by $\mathbf{y}_i$ and assumed to be convex and differentiable. Table 1 provides examples of classical loss functions in classification and regression. A dual formulation of (1) reads:

$$\max_{\boldsymbol{\theta}\in\Delta_{\boldsymbol{Z},\lambda}} D(\boldsymbol{\theta}) := -\sum_{i=1}^{n} f_i^*\left(-\boldsymbol{\theta}_i\right)\,, \tag{2}$$

where

$$\Delta_{\boldsymbol{Z},\lambda} = \left\{\boldsymbol{\theta} \in \mathbb{R}^n : \left|\mathbf{Z}^\top\boldsymbol{\theta}\right| \leq \lambda \mathbf{1}_D\,, \mathbf{1}_n^\top\boldsymbol{\theta} = 0\right\}\,, \tag{3}$$

and where f_i^* is the Fenchel-Legendre transform of the loss f_i, i.e., the function $f_i^* : \mathbb{R} \mapsto [-\infty, +\infty]$ defined by $f_i^*(u) = \sup_{v\in\mathbb{R}} uv - f_i(v)$. For the derivation of the dual problem, we refer the reader to Johnson & Guestrin (2015, Appendix E). The constraint $\mathbf{1}_n^\top\boldsymbol{\theta} = 0$ comes from the bias term $b\mathbf{1}_n$ in the primal problem (1). We denote by $(\mathbf{w}^*, b^*)$ and $\boldsymbol{\theta}^*$ a set of primal and dual optimal solutions to problems (1) and (2) respectively. Strong duality holds and therefore $(\mathbf{w}^*, b^*)$ and $\boldsymbol{\theta}^*$ satisfy Fermat's rules (Ndiaye et al., 2017):

$$\boldsymbol{\theta}^* = -\nabla F(\mathbf{Z}\mathbf{w}^* + b^*\mathbf{1}_n)\,, \tag{4}$$

and

$$\forall i \in [\![D]\!]\,, \quad \mathbf{Z}_i^\top\boldsymbol{\theta}^* \in \begin{cases} \{-\lambda, \lambda\} & \text{if } \mathbf{w}_i^* \neq 0\,, \\ [-\lambda, \lambda] & \text{if } \mathbf{w}_i^* = 0\,. \end{cases} \tag{5}$$

2.2. Basic working set algorithm

A general strategy to solve (1) is to follow a *working set* approach, as summarised in Algorithm 1. At each iteration, it solves (1) restricted to a small subset of features $\mathcal{W}$ called the working set. $\mathcal{W}$ is typically chosen as the set of features that violate the optimality condition (5) at the current iteration. In the sequel, we will call such features *violating features*. The algorithm converges when no violating feature remains, which occurs in a finite number of iterations as shown in Kowalski et al. (2011). When the number of interaction features runs into the billions, Algorithm 1 is not tractable since the delineation of the working set (line 3 in Alg. 1) requires $O(p^2n)$ operations at each iteration.

Algorithm 1 Working set algorithm

Input: $\mathbf{Z} \in \{0,1\}^{n\times D}, \mathbf{y} \in \mathbb{R}^n, \lambda > 0$
Output: $\mathbf{w}^*, b^*$
1: Set $\boldsymbol{\theta} \leftarrow -\nabla F(\mathbf{0}_n)$, $\mathcal{W} = \emptyset$.
2: **while** true **do**
3: $\quad \mathcal{W}' = \left\{i \in [\![D]\!] : \left|\mathbf{Z}_i^\top\boldsymbol{\theta}\right| \geq \lambda\right\}$
4: $\quad$ **if** $\max_{i\in\mathcal{W}'} \left|\mathbf{Z}_i^\top\boldsymbol{\theta}\right| \leq \lambda$ **then** Break **else** $\mathcal{W} \leftarrow \mathcal{W}'$
5: $\quad \mathbf{w}_{\mathcal{W}}^*, b^* \leftarrow \underset{\mathbf{w}_{\mathcal{W}}, b}{\operatorname{argmin}} P_{\mathbf{Z}_{\mathcal{W}},\lambda}(\mathbf{w}_{\mathcal{W}}, b)$
6: $\quad \boldsymbol{\theta} \leftarrow -\nabla F(\mathbf{Z}_{\mathcal{W}}\mathbf{w}_{\mathcal{W}}^* + b^*\mathbf{1}_n)$.
7: **end while**

3. The WHInter algorithm

3.1. Overview

WHInter is a working set algorithm that follows the general scheme of Algorithm 1 but implements an efficient strategy to delineate the working set among all main effects and interactions. It is described in Algorithm 2. The identification of the working set (line 3 in Algorithm 1) corresponds to lines 11-18 in Algorithm 2. Instead of scanning through all features to build the working set, WHInter first identifies branches that are guaranteed to contain no violating feature. These branches are identified via the evaluation of a *branch bound* $\eta(\mathbf{X}_j, \boldsymbol{\Theta}_j^{ref}, \boldsymbol{\theta}, \mathbf{m}_j^{ref})$ (line 13), which is presented in Section 3.2 together with the parameters it takes as input. The branch bound is cheap to evaluate since it solely depends on main effects and not on their numerous

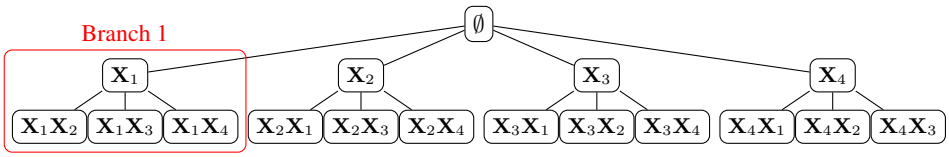

Figure 1 – Organisation of the main effects and interactions in a tree, depicted for 4 main effects.

Table 1 – Summary of useful functions for the LASSO and logistic regression: loss function f_i, its derivative f_i', its Fenchel-Legendre transform f_i^*.

	$f_i(u)$	$f_i'(u)$	$f_i^*(u)$
LASSO	$\frac{1}{2}(\mathbf{y}_i - u)^2$	$u - \mathbf{y}_i$	$\frac{1}{2}(\mathbf{y}_i + u)^2 - \frac{1}{2}\mathbf{y}_i^2$
Logistic regr.	$\log(1+\exp(-\mathbf{y}_i u))$	$-\frac{u}{\mathbf{y}_i}\log(-\frac{u}{\mathbf{y}_i}) + (1+\frac{u}{\mathbf{y}_i})\log(1+\frac{u}{\mathbf{y}_i})$	$\frac{-\mathbf{y}_i}{1+\exp(\mathbf{y}_i u)}$

interactions. Moreover, it is designed to efficiently rule out branches thanks to the exploitation of the shared structure among features in a branch, as well as the correlation among dual variables for two sufficiently close points in the optimisation path. In cases where a branch cannot be ruled out, features in the branch are considered one by one to build the working set, which is very computationally expensive. In order to reduce this cost, we cast the problem as a variant of the Maximum Inner Product Search (MIPS) problem, which is described in Section 3.3. If no violating feature is identified then the algorithm has converged. Otherwise, a new candidate solution is obtained by solving problem (1) restricted to the features in the working set (line 20), and the process is repeated until no violating feature remains. While any solver can be used to solve the restricted problem, we implemented in WHInter a coordinate descent approach with safe pruning.

3.2. The Branch bound η

As WHInter iterates, it produces candidate solutions $(\mathbf{w}^*, b^*)$ and corresponding dual variables $\boldsymbol{\theta}$ (lines 20 and 21 of Algorithm 2). For two sufficiently close iterations, or for two problems with sufficiently close regularisation parameters, the candidate solutions are likely to be *close* to one another, as well as the corresponding dual variables provided that the function F does not vary too quickly. WHInter exploits this intuition to speed up the identification of the working set from an iteration to another or from one problem to another. The following results relate the criteria used to identify the working set for two distinct dual variables (line 3 of Algorithm 1).

Lemma 3.1. *For any* $\mathbf{X} \in \{0,1\}^{n\times p}$, $\mathbf{v} \in \mathbb{R}_+^n$, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^n$, $j \in [\![p]\!]$, $\mathcal{I} \subset [\![p]\!]$ *and* $\alpha \in \mathbb{R}$, *the following holds:*

$$\begin{aligned} &\max_{k\in\mathcal{I}} \left| \boldsymbol{\theta}_2^\top (\mathbf{v} \odot \mathbf{X}_k) \right| \\ &\leq |\,\alpha\,| \max_{k\in\mathcal{I}} \left| \boldsymbol{\theta}_1^\top (\mathbf{v} \odot \mathbf{X}_k) \right| + \zeta(\boldsymbol{\theta}_2 - \alpha\boldsymbol{\theta}_1, \mathbf{v})\,, \end{aligned} \tag{6}$$

where $\forall (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^n \times \mathbb{R}_+^n$,

$$\zeta(\mathbf{u}, \mathbf{v}) = \max\left(\sum_{i:\mathbf{u}_i>0} \mathbf{u}_i\mathbf{v}_i, -\sum_{i:\mathbf{u}_i<0} \mathbf{u}_i\mathbf{v}_i \right).$$

The proof of Lemma 3.1 is provided in Appendix A. It is based on the decomposition $\boldsymbol{\theta}_2 = \alpha\boldsymbol{\theta}_1 + (\boldsymbol{\theta}_2 - \alpha\boldsymbol{\theta}_1)$, and exploits the tree structure among features in a branch. To exploit Lemma 3.1 in WHInter, we define for $\alpha \in \mathbb{R}$ and for all $(\mathbf{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, m) \in \mathbb{R}_+^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ the function:

$$\eta_\alpha(\mathbf{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, m) = |\,\alpha\,|\, m + \zeta(\boldsymbol{\theta}_2 - \alpha\boldsymbol{\theta}_1, \mathbf{v})\,, \tag{7}$$

and we maintain an active set $\mathcal{W} \subset [\![D]\!]$, a matrix $\boldsymbol{\Theta}^{ref} \in \mathbb{R}^{n\times p}$ that contains *reference dual variables* $\boldsymbol{\Theta}_j^{ref} \in \mathbb{R}^n$ for each branch $j \in [\![p]\!]$, and the vector $\mathbf{m}^{ref} \in \mathbb{R}^p$ defined by:

$$\forall j \in [\![p]\!]\,, \quad \mathbf{m}_j^{ref} = \max_{k\in[\![p]\!]:\tau(j,k)\notin\mathcal{W}} \left| \mathbf{Z}_{\tau(j,k)}^\top \boldsymbol{\Theta}_j^{ref} \right|. \tag{8}$$

We now state our main theorem which allows to identify branches that are guaranteed to not contain any violating feature (line 13 of Algorithm 2):

Theorem 3.1 (Branch pruning)**.** *For any* $\boldsymbol{\Theta}^{ref} \in \mathbb{R}^{n\times p}$, $\mathcal{W} \subset [\![p]\!]$, $j \in [\![p]\!]$, *let* $\mathbf{m}_j^{ref} \in \mathbb{R}_+$ *be given by (8). Then for any* $\boldsymbol{\theta} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$ *and* $\lambda > 0$, *if*

$$\eta_\alpha\left(\mathbf{X}_j, \boldsymbol{\Theta}_j^{ref}, \boldsymbol{\theta}, \mathbf{m}_j^{ref}\right) < \lambda\,, \tag{9}$$

then any feature from branch j *that belongs to the working set* $\left\{i \in [\![D]\!] : \left|\mathbf{Z}_i^\top \boldsymbol{\theta}\right| \geq \lambda\right\}$ *is already in* $\mathcal{W}$. *This holds in particular if*

$$\eta_{min} := \min_{\alpha\in\mathbb{R}} \eta_\alpha\left(\mathbf{X}_j, \boldsymbol{\Theta}_j^{ref}, \boldsymbol{\theta}, \mathbf{m}_j^{ref}\right) < \lambda\,. \tag{10}$$

Algorithm 2 WHInter

Input: $\mathbf{X} \in \{0,1\}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, $\lambda_1 > \cdots > \lambda_T$.
Output: $(\mathcal{W}, \mathbf{w}^*_{\mathcal{W}}, b^*)_t$ for each λ_t

```
    # Initialisation
 1: θ ← −∇F(0_n)
 2: for j in [[p]] do
 3:    Θ_j^ref ← θ
 4: end for
 5: W, m^ref ← update_W(X, θ, [[p]], λ_1, ∅)
 6: for t = 1 to T do
      # Pre-Solve
 7:    w*_W, b* ← argmin_{w_W, b} P_{Z_W, λ_t}(w_W, b)
 8:    θ ← −∇F(Z_W w*_W + b* 1_n).
 9:    W, m^ref ← clean_W(W, λ_t, θ, Θ^ref, m^ref)
10:    while true do
         # Branch pruning
11:       V ← ∅
12:       for j in [[p]] do
13:          if η(X_j, Θ_j^ref, θ, m_j^ref) ≥ λ_t then
14:             V ← V ∪ {j}
15:             Θ_j^ref ← θ
16:          end if
17:       end for
         # Identify the working set
18:       W', m_V^ref ← update_W(X, θ, V, λ_t, W)
19:       if max_{i∈W'} |Z_i^⊤ θ| ≤ λ then
20:          Break
21:       else
22:          W ← W'
23:       end if
         # Solve subproblem
24:       w*_W, b* ← argmin_{w_W, b} P_{Z_W, λ_t}(w_W, b)
25:       θ ← −∇F(Z_W w*_W + b* 1_n).
26:       W, m^ref ← clean_W(W, λ_t, θ, Θ^ref, m^ref)
27:    end while
28:    (W, w*_W, b*)_k ← (W, w*_W, b*)
29: end for

30: function clean_W(W, λ, θ, Θ^ref, m^ref)
31:    for i in W do
32:       if |Z_i^⊤ θ| < λ then
33:          Remove {i} from W
34:          for b in branch(i) do
35:             if m_b^ref < |Z_i^⊤ Θ_b^ref| then
36:                m_b^ref ← |Z_i^⊤ Θ_b^ref|
37:    return W, m^ref
```

Proof. Take $\mathcal{I} = \{k \in [\![p]\!] : \tau(j,k) \notin \mathcal{W}\}$, $\mathbf{v} = \boldsymbol{X}_j$, $\boldsymbol{\theta}_1 = \boldsymbol{\Theta}_j^{ref}$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}$ in Lemma 3.1. Then if (9) holds, we deduce from (6) that

$$\max_{k \in [\![p]\!] : \tau(j,k) \notin \mathcal{W}} \left| \mathbf{Z}_{\tau(j,k)}^\top \boldsymbol{\theta} \right| < \lambda \,.$$

This shows that there is no feature i in branch j such that $\left|\mathbf{Z}_i^\top \boldsymbol{\theta}\right| \geq \lambda$ *and* i is not already in $\mathcal{W}$. The fact that for fixed arguments, the function $\alpha \to \eta_\alpha$ has a minimum $\alpha^* \in \mathbb{R}$ is shown in Appendix B, along with an algorithm to compute it in $O\left(\| \mathbf{X}_j \|_0 \ln \| \mathbf{X}_j \|_0\right)$ operations . Since the statement is true for any α, it is *a fortiori* true for α^*. □

Theorem 3.1 provides criteria (9) and (10) that can be computed for each branch j, and which if satisfied allow to skip the search for violating variables in the branch. Importantly, the features that are already in the working set $\mathcal{W}$ are not taken into account to compute the criterion for a given branch. This subtlety allows to rule out branches even if they already contain features that were previously incorporated in the working set. Note that the reference dual variable for branch j, i.e, $\boldsymbol{\Theta}_j^{ref}$, is kept unchanged as long as branch j is pruned, and is otherwise updated to the latest dual variable (line 15 of Algorithm 2). As $\mathbf{m}_j^{ref}$ depends on the reference dual variable instead of the current one, it is solely reevaluated each time the reference residual is updated (line 18 of Algorithm 2) or when a feature from branch j leaves the working set (line 22 of Algorithm 2) .

Criterion (10) is the most stringent one, and therefore the most efficient one to prune branches, but it takes $O\left(\| \mathbf{X}_j \|_0 \ln \| \mathbf{X}_j \|_0\right)$ operations to compute. In order to balance computational complexity of the bound with its efficacy to prune branches, criterion (9) can be used as an alternative for a specific value of α. One simple choice is to just take $\alpha = 1$, which leads to the criterion

$$\eta_1\left(\mathbf{X}_j, \boldsymbol{\Theta}_j^{ref}, \boldsymbol{\theta}, \mathbf{m}_j^{ref}\right) = \mathbf{m}_j^{ref} + \zeta\left(\boldsymbol{\theta} - \boldsymbol{\Theta}_j^{ref}, \mathbf{X}_j\right) < \lambda \,. \tag{11}$$

Alternatively, a simple heuristic to expect a more efficient pruning is to choose an α that minimises $\| \left(\boldsymbol{\theta} - \alpha \boldsymbol{\Theta}_j^{ref}\right) \odot \mathbf{X}_j \|_2$, i.e,

$$\alpha_{\ell_2} = \frac{\boldsymbol{\theta}^\top \left(\boldsymbol{\Theta}_j^{ref} \odot \mathbf{X}_j\right)}{\| \boldsymbol{\Theta}_j^{ref} \odot \mathbf{X}_j \|_2^2} \,. \tag{12}$$

$\eta_{\alpha_{\ell_2}}$ is expected to be more effective than η_1 since it is reasonable to expect that $\zeta\left(\boldsymbol{\theta} - \alpha_{\ell_2} \boldsymbol{\Theta}_j^{ref}, \mathbf{X}_j\right)$ is smaller than $\zeta\left(\boldsymbol{\theta} - \boldsymbol{\Theta}_j^{ref}, \mathbf{X}_j\right)$. Overall, computing $\alpha = \alpha_{\ell_2}$ as in (12) is an $O(\| \mathbf{X}_j \|_0)$ operation. Since computing $\zeta(\boldsymbol{\theta} - \alpha \boldsymbol{\Theta}_j^{ref}, \mathbf{X}_j)$ for a fixed α is also a $O(\| \mathbf{X}_j \|_0)$ computation, the total cost of identifying branch j as violated is $O(\| \mathbf{X}_j \|_0)$ for criterion (9) with $\alpha = 1$ or $\alpha = \alpha_{\ell_2}$,

compared to $O\left(\| \mathbf{X}_j \|_0 \ln \| \mathbf{X}_j \|_0\right)$ for criterion (10). In Algorithm 2, the notation η refers to a user-defined function among $\eta_1, \eta_{\alpha_{\ell_2}}$ or η_{min}.

3.3. Updating the working set

When some branches $\mathcal{V} \subset [\![p]\!]$ cannot be pruned, the simultaneous updates of the working set $\mathcal{W}$ and of $\mathbf{m}_{\mathcal{V}}^{ref}$ requires scanning through all features in the branches $\mathcal{V}$ (lines 5 and 18 in Algorithm 2). In what follows we discuss strategies to make these updates efficient. For that purpose, let us first notice that:

$$\begin{aligned}\forall j,k \in [\![p]\!]\ , \left|\mathbf{Z}_{\tau(j,k)}^\top \boldsymbol{\theta}\right| &= \left|\left(\mathbf{X}_j \odot \mathbf{X}_k\right)^\top \boldsymbol{\theta}\right| \\ &= \left|\left(\mathbf{X}_j \odot \boldsymbol{\theta}\right)^\top \mathbf{X}_k\right| \\ &= \left|\boldsymbol{Q}_j^\top \mathbf{X}_k\right|\ ,\end{aligned}$$

where for any $j \in [\![p]\!]\ , \boldsymbol{Q}_j = \boldsymbol{X}_j \odot \boldsymbol{\theta}$. This allows us to write the updates of $\mathcal{W}$ and $\boldsymbol{m}_{\mathcal{V}}^{ref}$ as:

$$\begin{cases}\mathcal{W}' = \mathcal{W} \cup \left\{\tau(j,k) : j \in \mathcal{V}, k \in [\![p]\!]\ , \left|\boldsymbol{Q}_j^\top \boldsymbol{X}_k\right| \geq \lambda\right\}\ , \\ \boldsymbol{m}_j^{ref} = \max\limits_{k:\left|\boldsymbol{Q}_j^\top \boldsymbol{X}_k\right| < \lambda} \left|\boldsymbol{Q}_j^\top \boldsymbol{X}_k\right|, \forall j \in \mathcal{V}\,.\end{cases} \tag{13}$$

This highlights the fact that the updates of the working set $\mathcal{W}$ and of $\boldsymbol{m}_{\mathcal{V}}^{ref}$ can be cast as particular variants of the Maximum Inner Product Search (MIPS) problem. MIPS aims at finding a vector in a database of probes which maximises the inner product with a given query vector. If we consider $\mathbf{X}$ as a set of probes, and $\boldsymbol{Q}_j$ as a query, then (13) is a variant of MIPS where (i) the set of probe vectors satisfies some constraints and is not known upfront and (ii) the problem is a maximum *absolute* inner product search. The update of $\mathcal{W}$ involves what is sometimes referred to as *above-λ*-MIPS problems where again, maximum *absolute* inner products are considered. The interest of casting these updates as variants of MIPS problems is to exploit the ideas developed in the literature for solving these problems efficiently. Teflioudi & Gemulla (2016) and Fontoura et al. (2011) give good overviews of MIPS solvers developed for recommender systems and information retrieval applications respectively. In both cases, the proposed methods rely on two main ideas: (i) adequate indexing techniques or data structures and (ii) pruning criteria which allow to not compute all inner products entirely. Since none of these methods can directly be applied to problem (13) because of its specificities, we propose an appropriate algorithm based on a simple inverted index approach, which we will refer to as IL (standing for Inverted Lists), and which exploits the sparsity of the problem. Another option would be to leverage pruning techniques. We detail such an attempt in Appendix C. However, since our preliminary results with the pruning technique were not conclusive compared to IL on the simulated and real data, we only focus on the inverted index approach below. IL is detailed in Algorithm 3.

Algorithm 3 `update_W`

Input: $\mathbf{X} \in \{0,1\}^{n \times p}$, $\boldsymbol{\theta} \in \mathbb{R}^n$, $\mathcal{V} \subset [\![p]\!]$, $\lambda \in \mathbb{R}$, $\mathcal{W} \subset [\![D]\!]$
Output: $\mathcal{W}$, $\mathbf{m}^{ref}$

1: **for** $j \in \mathcal{V}$ **do**
2: $\quad \boldsymbol{m}_j^{ref} = 0$
3: $\quad$ Set $\mathbf{a}_k = 0$ for all $k \in [\![p]\!]$
4: $\quad$ **for each** i in supp$(\mathbf{X}_j)$ **do**
5: $\quad\quad$ **for each** k in supp$(\boldsymbol{x}_i)$ **do**
6: $\quad\quad\quad \mathbf{a}_k = \mathbf{a}_k + \boldsymbol{\theta}_i$
7: $\quad\quad$ **end for**
8: $\quad$ **end for**
9: $\quad$ **for each** k s.t. $\mathbf{a}_k \neq 0$ **do**
10: $\quad\quad$ **if** $\mathbf{m}_j^{ref} < |\,\mathbf{a}_k\,| < \lambda$ **then** set $\mathbf{m}_j^{ref} = |\,\mathbf{a}_k\,|$
11: $\quad\quad$ **if** $|\,\mathbf{a}_k\,| \geq \lambda$ and $\tau(j,k) \notin \mathcal{W}$ **then** add $\tau(j,k)$ to $\mathcal{W}$
12: $\quad$ **end for**
13: **end for**
14: **return** $\mathcal{W}, \boldsymbol{m}^{ref}$

The inverted lists consist of n lists, one for each dimension, where each list supp$(\mathbf{x}_i)$ records the indices of the features in $\mathbf{X}$ which have a non-zero element for the i^{th} dimension. These inverted lists can be computed once for all when WHInter starts and be reused for all MIPS problems, and therefore building the inverted lists requires a negligible additional computational cost. Algorithm (3) computes inner product following a *term-at-a-time* (TAAT) scheme (Fontoura et al., 2011), i.e, the inner products are accumulated simultaneously across probes and the contribution of the i^{th} dimension to the inner products is entirely processed before moving to the next one.

4. Simulation study

We first test the performances of WHInter on synthetic LASSO datasets. We assess the performances of the different branch pruning bounds presented in 3.2, i.e, η_{min}, η_1 and $\eta_{\alpha_{\ell_2}}$, and further compare WHInter to a working set method that uses the bound $\zeta(\boldsymbol{\theta}, \mathbf{X}_j)$ instead of η_α, but is otherwise equivalent to WHInter. We refer to this method as $\zeta + IL$. It is expected to prune less branches than WHInter but does not require to maintain $\boldsymbol{m}^{ref}$. We also compare WHInter to SPP (Nakagawa et al., 2016) and BLITZ (Johnson & Guestrin, 2015). In our experiments, we use a slightly modified, more efficient version of the code provided by the authors of SPP (cf Appendix D). As for BLITZ, since the method is not tailored for interaction problems, we first compute the matrix $\boldsymbol{Z}$ which is fed as input to BLITZ. For this reason we could not solve problems when p is too large (e.g.,

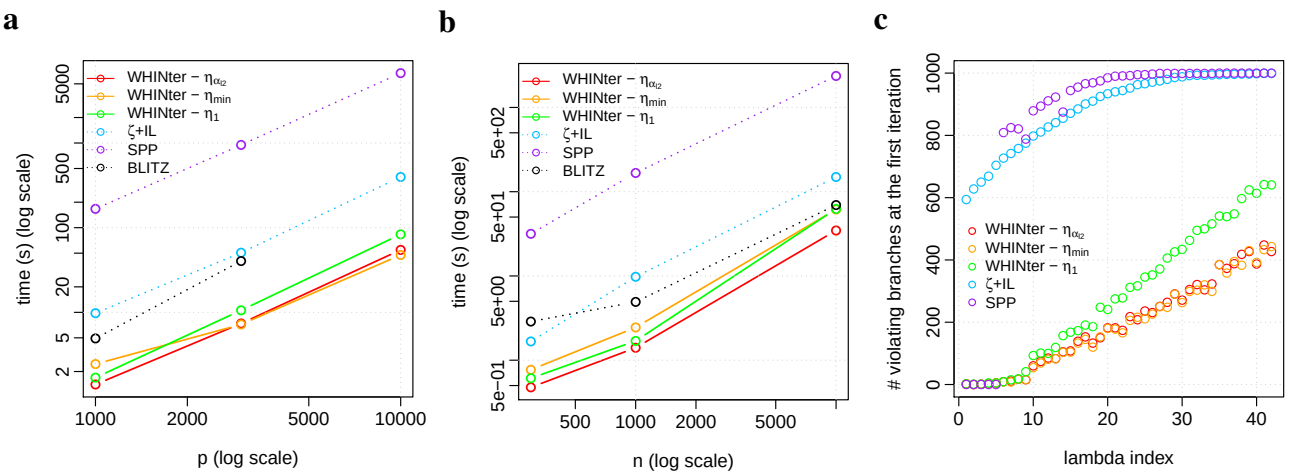

Figure 2 – Performance comparison on simulated data for an entire regularisation path. Comparison of WHInter with three branch pruning criteria $\eta \in \{\eta_{\alpha_2}, \eta_{min}, \eta_1\}$ to $\zeta + IL$, SPP and BLITZ. (a) Time in seconds for $n = 1,000$ fixed and p varied. (b) Time in seconds for $p = 1,000$ fixed and n varied. (c) Number of branches that could *not* pruned at the first iteration, as a function of λ, for $n = p = 1,000$.

$p = 10,000$ in the simulations) since, even in sparse format, storing $\boldsymbol{Z}$ requires too much memory. Importantly, the performances reported for BLITZ do not include the time required to compute $\boldsymbol{Z}$ from $\boldsymbol{X}$, which clearly advantages BLITZ compared to the other methods.

We simulate five datasets $\mathbf{X} \in \{0,1\}^{n \times p}$ with varying number of features and samples: three datasets with $p = 1,000$ fixed and $n \in \{300, 1,000, 10,000\}$, and two more with $n = 1,000$ fixed and $p \in \{3,000, 10,000\}$. The features are drawn from a Bernoulli distribution with parameter $q \in [0.1, 0.5]$ itself drawn from a uniform distribution $\mathcal{U}_{[0.1,0.5]}$. We then randomly pick a set $\mathcal{S}$ of 100 features among the main effects and interactions and compute the response as $\mathbf{y} = \mathbf{Z}_{\mathcal{S}}\mathbf{w}^*_{\mathcal{S}}$ where $\mathbf{w}^*_{\mathcal{S}} \sim \mathcal{N}(\mathbf{0}_{|\mathcal{S}|}, I_{|\mathcal{S}|})$. In all experiments, the LASSO is solved for a sequence $(\lambda_t)_{t \in [\![T]\!]}$, $T = 100$, logarithmically spaced between λ_{max} and $\max(0.01\lambda_{max}, \lambda')$ where λ_{max} is the largest value of λ for which at least one feature is selected, and λ' is the first λ_i for which 150 features or more are selected in the model. For all methods, the time to compute λ_{max} is included in the total time required to solve the regularisation path. In WHInter, λ_{max} can easily be deduced from the initialisation of $\mathbf{m}^{ref}$ since $\lambda_{max} = \max_{j \in [\![p]\!]} \boldsymbol{m}^{ref}_j$. All algorithms are implemented in C++ and compiled with the `-O3` optimisation flag. The experiments are run on a 64-bit machine with Intel Core i7 Processor 2.5 GHz, 16GB of memory and 6MB of cache.

Results are shown in Figure 2. For $n = 1,000$ (Figure 2a), LASSO solutions are computed for 42, 32 and 28 values of λ for $p = 1,000, p = 3,000$ and $p = 10,000$ respectively. In these cases smaller values of λ result in model sizes exceeding 150 features. For the remaining settings where $p = 1,000$ and $n = 300$ or $n = 10,000$ (Figure 2b), LASSO solutions are computed for 34 and all 100 values of λ between λ_{max} and $0.01\lambda_{max}$, respectively. All methods returned the exact same support for all values of λ.

In all settings, WHInter is the fastest method. Its better performance compared to $\zeta + IL$ highlights the benefit of using reference dual variables even if it implies to maintain $\boldsymbol{m}^{ref}$. The results also show the importance of α, since WHInter with η_{ℓ_2} is always better ($\times 1.2$ to $\times 1.8$) than WHInter with η_1 for example. Figure 2c confirms that the choice of α has an impact on the pruning efficiency and consequently on the performance. It shows, however, that on this experiment η_{min} does not allow to prune many more branches than η_{ℓ_2}. This explains why η_{ℓ_2} tends to outperform η_{min}, notably for large n, since the higher computational complexity of η_{min} does not sufficiently enhance the pruning. We also notice that SPP is the slowest algorithm, and in particular $\zeta + IL$ is $\times 17$ faster than SPP on average. This speed-up is mostly explained by the fact that $\zeta + IL$ relies on inverted lists to update the working set while SPP identifies the safe set naively. Overall, WHInter offers a signifiant speed-up of two orders of magnitude or more compared to its safe screening counterpart.

5. Results on real world data

We now illustrate the performance of the different algorithms on a real-world problem, where we want to predict the cytotoxic response of 884 lymphoblastoid cell lines split into a train ($n = 620$) and a test ($n = 264$) set, and characterized by about 1.2×10^6 single nucleotide polymorphisms (SNP) that represent their genotypes. The data was released as part of the Dialogue on Reverse Engineering Assessment and Methods 8 (DREAM 8) toxicogenetics challenge (Ed-

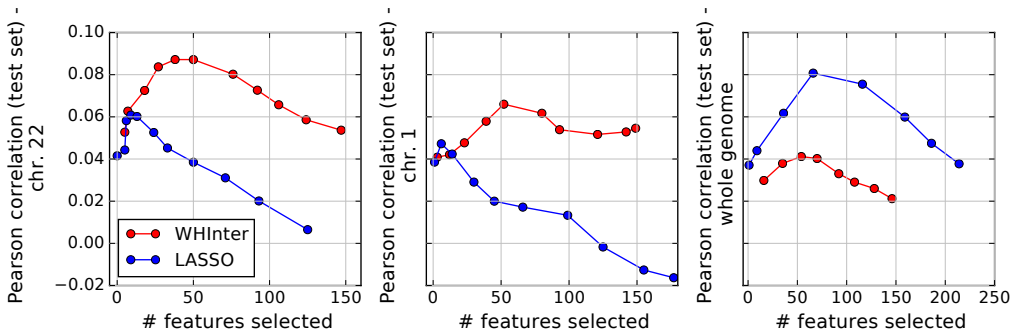

Figure 3 – Predictive performance on the test set. The y-axis reports the pearson correlation between the true and predicted response. The x-axis reports the number of selected features for the sequence of regularisation parameters tested.

uati et al., 2015). We encode the SNP data as a binary matrix were 1 stand for the presence of a minor allele on one or both copies of the chromosomes. As preprocessing we removed SNP with less than 5% of 1's and corrected the data for population structure as in Price et al. (2006). To focus on problems of increasing scales, we first considered the SNPs of the smallest chromosome only (chr. 22), then of the largest only (chr. 1) and finally of all chromosomes together. This leads to train matrices with $n = 620$ and $p = 18,168$ SNPs for chromosome 22, $p = 89,027$ SNPs for chromosome 1 and $p = 1,166,836$ SNPs for the whole genome.We consider a sequence of regularisation parameter λ logarithmically spaced between λ_{max} and $0.01\lambda_{max}$, and by default stop computations as soon as 150 features or more are selected. This occurs after the 12^{th}, the 11^{th} and the 9^{th} value of λ for chromosome 22, chromosome 1 and all chromosomes respectively. The time required to compute the regularisation paths are shown in Fig. 4. The relative performances of the methods are the same as for the simulations. $\eta_{\alpha_{\ell_2}}$ provides a $\times 1.4$ (resp. $\times 1.8$) speed up compared to using η_1 for chromosome 22 (resp. chr. 1). and compared to SPP, there is a $\times 81$ (resp. $\times 73$) speed up for chromosome 22 (resp chr. 1). In the case of the whole genome, we only ran WHInter with $\eta_{\alpha_{\ell_2}}$ which takes two days and a half. While this can seem a lot, we recall that this corresponds to a problem with roughly 680 *billion* features. We did not run other methods on the whole genome since most of them are expected to take too long.

Out of curiosity, we also obtained preliminary results concerning the predictive performance of WHInter compared to a LASSO with no interactions on such high-dimensional problems. The results, presented in Figure 3 , suggest that interactions are relevant predictors for this data. For the chromosomes 1 and 22 taken independently, the predictive accuracy of WHInter is better than that of the simple LASSO for almost every value of λ. By contrast, for the whole genome, the LASSO clearly performs better, which may underline statistical issues due to the huge number of variables in this case (Donoho & Tanner, 2009).

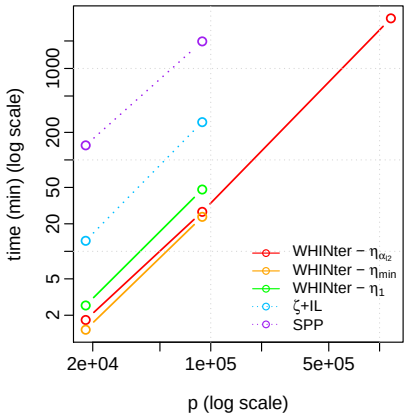

Figure 4 – Performance comparison on SNPs data for an entire regularisation path. The y-axis reports the total time (in minutes) required to compute the LASSO path for chromosome 22 (around 20,000 SNPs), chromosome 1 (around 90,000 SNPs) and the whole genome (around 1.2 million SNPs).

6. Discussion

We presented WHInter, a working set algorithm designed to solve large scale ℓ_1-penalised linear problems with interaction terms. WHInter implements a new branch pruning bound to efficiently delineate the working set among the many possible interaction variables, and a variant of MIPS solver that provides a further speed up. We showed that WHInter is up to two orders of magnitudes faster than competing approaches. While we presented WHInter for binary data, it could also be used for data rescaled in $[0, 1]$, provided that an appropriate solver is picked for the MIPS problems. As for future work, one could exploit the recent works on approximate MIPS (Shrivastava & Li, 2014; Teflioudi & Gemulla, 2016) to obtain an additional speed up for the computationally intensive updates, and possibly rely on recent post selection-inference (Suzumura et al., 2017) frameworks to characterise the approximate solution obtained.

Acknowledgements

We thank our anonymous reviewers for their useful comments as well as Nino Shervashidze for thoughtful discussions.

References

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. hierarchical selection of variables in sparse high-dimensional regression. In *Borrow. strength theory powering Appl. Festschrift Lawrence D. Brown*, pp. 56–69. Institute of Mathematical Statistics, 2010.

Bien, J., Taylor, J., and Tibshirani, R. A lasso for hierarchical interactions. *Ann. Stat.*, 41(3):1111–1141, 2013.

Bonnefoy, A., Emiya, V., Ralaivola, L., and Gribonval, R. Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso. *IEEE Trans. Signal Process.*, 63(19):5121–5132, 2015.

Donoho, D. L. and Tanner, J. Observed universality of phase transition in high-dimenisonal geometry, with applications for modern data analysis and signal processing. *Philos. Trans. R. Soc. London A Math. Phys. Eng. Sci.*, 367(1906):4273–4293, 2009.

Eduati, F., Mangravite, L. M., Wang, T., Tang, H., Bare, J. C., Huang, R., Norman, T., Kellen, M., Menden, M. P., Yang, J., Zhan, X., Zhong, R., Xiao, G., Xia, M., Abdo, N., Kosyk, O., Friend, S., Dearry, A., Simeonov, A., Tice, R. R., Rusyn, I., Wright, F. A., Stolovitzky, G., Xie, Y., Saez-Rodriguez, J., Aittokallio, T., Alaimo, S., Amadoz, A., Ammad-ud din, M., Azencott, C. A., Bacardit, J., Barron, P., Bernard, E., Beyer, A., Bin, S., van Bömmel, A., Borgwardt, K., Brys, A. M., Caffrey, B., Chang, J., Chang, J., Chheda, H., Christodoulou, E. G., Clément-Ziza, M., Cohen, T., Cowherd, M., Demeyer, S., Dopazo, J., Elhard, J. D., Falcao, A. O., Ferro, A., Friedenberg, D. A., Giugno, R., Gong, Y., Gorospe, J. W., Granville, C. A., Grimm, D., Heinig, M., Hernansaiz, R. D., Hintsanen, P., Hochreiter, S., Huang, L. C., Huska, M., Jaiswal, A., Jiao, Y., Kaski, S., Kaur, I., Ali Khan, S., Klambauer, G., Krasnogor, N., Kuhn, M., Bartosz Kursa, M., Kutum, R., Lazzarini, N., Lee, I., Leung, M. K., Khong Lim, W., Liu, C., Llinares López, F., Mammana, A., Mayr, A., Michoel, T., Mongiovì, M., Moore, J. D., Mpindi, J. P., Narasimhan, R., Opiyo, S. O., Pandey, G., Peabody, A. L., Perner, J., Poso, A., Pulvirenti, A., Rawlik, K., Reinhardt, S., Riffle, C. G., Ruderfer, D., Sander, A. J., Savage, R. S., Scornet, E., Sebastian-Leon, P., Sharan, R., Johann Simon-Gabriel, C., Stoven, V., Sun, J., Tang, J., Teixeira, A. L., Tenesa, A., Vert, J. P., Vingron, M., Walter, T., Wennerberg, K., Whalen, S., Wisniewska, Z., Wu, Y., Xu, H., Zhang, S., Zhao, J., Jim Zheng, W., and Ziwei, D. Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotechnol.*, 33(9):933–940, 2015.

El Ghaoui, L., Viallon, V., and Rabbani, T. Safe feature elimination in sparse supervised learning. *Pacific J. Optim.*, 8 (4):667–698, 2012.

Fercoq, O., Gramfort, A., and Salmon, J. Mind the Duality Gap: Safer Rules for the Lasso. In *Proc. 32nd Int. Conf. Mach. Learn.*, pp. 333–342, 2015.

Fontoura, M., Josifovski, V., Liu, J., Venkatesan, S., Zhu, X., and Zien, J. Y. Evaluation Strategies for Top-k Queries over Memory-Resident Inverted Indexes. *Proc. VLDB Endow.*, 4(12):1213–1224, 2011.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

Fujiwara, Y., Ida, Y., Shiokawa, H., and Iwamura, S. Fast Lasso Algorithm via Selective Coordinate Descent. In *Proc. 30th Conf. Artif. Intell.*, pp. 1561–1567, 2016.

Hao, N. and Zhang, H. H. Interaction Screening for Ultra-High Dimensional Data. *J. Am. Stat. Assoc.*, 109(507): 1285–1301, 2014.

Haris, A., Witten, D., and Simon, N. Convex Modeling of Interactions With Strong Heredity. *J. Comput. Graph. Stat.*, 25(4):981–1004, 2016.

Johnson, T. and Guestrin, C. Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization. In *Proc. 32nd Int. Conf. Mach. Learn.*, pp. 1171–1179, 2015.

Johnson, T. B. and Guestrin, C. StingyCD: Safely Avoiding Wasteful Updates in Coordinate Descent. In *Proc. 34th Int. Conf. Mach. Learn.*, pp. 1752–1760, 2017.

Kowalski, M., Weiss, P., Gramfort, A., and Anthoine, S. Accelerating ISTA with an active set strategy. In *OPT 2011 4th Int. Work. Optim. Mach. Learn.*, pp. 7, 2011.

Lim, M. and Hastie, T. Learning Interactions via Hierarchical Group-Lasso Regularization. *J. Comput. Graph. Stat.*, 24(3):627–654, 2015.

Massias, M., Gramfort, A., and Salmon, J. From safe screening rules to working sets for faster Lasso-type solvers. *ArXiv e-prints*, 2017.

Nakagawa, K., Suzumura, S., Karasuyama, M., Tsuda, K., and Takeuchi, I. Safe Pattern Pruning: An Efficient Approach for Predictive Pattern Mining. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1785–1794, 2016.

Ndiaye, E., Fercoq, O., Gramfort, A., and Salmon, J. Gap Safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res.*, 18(128):1–33, 2017.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, 2006.

Radchenko, P. and James, G. Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions. *J. Am. Stat. Assoc.*, 105(492):1541–1553, 2010.

Raj, A., Olbrich, J., Gärtner, B., Schölkopf, B., and Jaggi, M. Screening Rules for Convex Problems. *ArXiv e-prints*, 2016.

Shah, R. D. Modelling interactions in high-dimensional data with backtracking. *J. Mach. Learn. Res.*, 17(207): 1–31, 2016.

Shrivastava, A. and Li, P. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Adv. Neural Inf. Process. Syst.*, pp. 2321–2329, 2014.

Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K., and Takeuchi, I. Selective Inference for Sparse High-Order Interaction Models. In *Proc. 34th Int. Conf. Mach. Learn.*, volume 70, pp. 3338–3347, 2017.

Teflioudi, C. and Gemulla, R. Exact and Approximate Maximum Inner Product Search with LEMP. *ACM Trans. Database Syst.*, 42(1):5:1—-5:49, 2016.

Tibshirani, R. Regression Selection and Shrinkage via the Lasso. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, 58(1): 267–288, 1996.

Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202, 2009.

Wang, J., Zhou, J., Wonka, P., and Ye, J. Lasso screening rules via dual polytope projection. In *Adv. Neural Inf. Process. Syst.*, pp. 1070–1078, 2013.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

Xiang, Z., Xu, H., and Ramadge, P. Learning sparse representations of high dimensional data on large scale dictionaries. In *Adv. Neural Inf. Process. Syst.*, pp. 900–908, 2011.

Xiang, Z. J. and Ramadge, P. J. Fast lasso screening tests based on correlations. In *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2137–2140, 2012.