Learning Interpretable Models using Soft Integrity Constraints

Khaled Belahcene, Nataliya Sokolovska, Yann Chevaleyre, Jean-Daniel Zucker

Learning Interpretable Models using Soft Integrity Constraints

Khaled Belahcene[1], **Nataliya Sokolovska**[1], **Yann Chevaleyre**[2], **Jean-Daniel Zucker**[3]
[1]NutriOmics, INSERM, Sorbonne University, Paris, France
[2]LAMSADE, Dauphine University, PSL Research University, UMR CNRS 7243, Paris, France, Paris, France
[3]UMI 209 UMMISCO, IRD, Bondy, France

Abstract

Integer models are of particular interest for applications where predictive models are supposed not only to be accurate but also interpretable to human experts. We introduce a novel penalty term called *Facets* whose primary goal is to favour integer weights. Our theoretical results illustrate the behaviour of the proposed penalty term: for small enough weights, the Facets matches the L_1 penalty norm, and as the weights grow, it approaches the L_2 regularizer. We provide the proximal operator associated with the proposed penalty term, so that the regularized empirical risk minimizer can be computed efficiently. We also introduce the *Strongly Convex Facets*, and discuss its theoretical properties. Our numerical results show that while achieving the state-of-the-art accuracy, optimisation of a loss function penalized by the proposed Facets penalty term leads to a model with a significant number of integer weights.

1 Introduction

The goal of supervised learning is to estimate a model from observations which generalises as accurately as possible to unseen data. We are interested in interpretable models, and we focus on linear models. Linear models whose weights are 1) *sparse*; 2) *small*; and 3) *integers* are even more preferable for human experts, since these models are easier to interprete.

Traditionally, a machine learning algorithm is cast as an optimisation problem. In a classification task, one would aim to maximize directly the *accuracy* of the model, however, the corresponding loss function, the 0-1 loss, is not convex and its minimization is intractable for real-world applications. Therefore, a widely used approach is to relax the optimization problem with a surrogate loss, chosen to be convex (or even better: strongly convex, or smooth), and to bound the 0-1 loss from above. Such an upper bound obtained on the surrogate loss provides some guarantees on the accuracy.

In the supervised learning scenario, learning models with small parameters or weights, and also sparse models, is already known to be beneficial, since the compact models overfit less. Shrinking parameters of a model is often addressed through *regularization* where the objective function, subject to minimization, consists of two terms, namely, of a loss term enforcing accuracy, and of a penalty term which is responsible for sparsity and for the parameters magnitude. A number of penalty functions have been proposed. The most known are probably Tikhonov regularization (Hastie, Tibshirani, and Friedman 2009) shrinking parameters towards zero, and Lasso regularization (Tibshirani 1996) setting a controlled (by a hyperparameter) number of weights exactly to zero. A number of penalty terms including various norms and their combinations have been proposed in the past decade (Hastie, Tibshirani, and Wainwright 2015).

Our aim is to introduce an efficient method to learn *compact integer linear models*. Several prior works addressed this question, and among them we would like to mention the following results. (Golovin et al. 2013) aim to find a model that is both sparse and integral, mostly for memory-saving reasons, and they use a randomized rounding scheme at each step of an online gradient descent to achieve the goal. (Chevaleyre, Koriche, and Zucker 2013) challenge to estimate a model with very small integers, i.e., either in $\{0, 1\}$ or in $\{-1, 0, 1\}$. Their motivation is to increase interpretability of machine learning models, and they use either a randomized or a so-called greedy rounding scheme at the end of the optimization process. In (Chevaleyre, Koriche, and Zucker 2013), the focus is on the hinge loss, and, therefore, on large-margin classifiers. In the SLIM, proposed by (Ustun and Rudin 2016), an integer model that jointly optimizes sparsity and accuracy by formulating and solving an Integer Linear Program is introduced and successfully tested on several benchmarks.

The state-of-the-art methods mentioned above achieve full integrity, however, they either rely on a serious tinkering of an algorithm such as in (Golovin et al. 2013), or a post-processing phase (Chevaleyre, Koriche, and Zucker 2013), or come with a drastic computational burden (Ustun and Rudin 2016).

Our contribution to interpretable models learning is multifold:

- We introduce a *novel penalty term*, called *Facets*, which favours models with small integers;
- We consider theoretical properties and optimisation issues of the *Facets* and *Strongly Convex Facets* penalty terms;

Submitted.

note that the introduced penalty term does not compromise the convexity of the objective function;

- Finally, we illustrate that the proposed method achieves the state-of-the-art results on real-world data.

The paper is organised as follows. Section 2 is devoted to notations we use in the paper. We introduce the Facets penalty term in Section 3. Section 4 is dedicated to theoretical results and properties of the Facets regularization. We discuss the optimisation issues in Section 5. In Section 6 we demonstrate our numerical results. Concluding remarks and perspectives close the paper.

2 Preliminaries

We are in the context of supervised learning where a training method has access to n observations and their labels. In this section, we introduce some notions we use throughout the paper.

Models. A linear model is a vector $\mathbf{w} \in \mathbb{R}^m$. The *integrity* of a model is the proportion of coefficients $w_j, j \in \{1, \dots, m\}$ that are integers. The *magnitude* of a model is an upper bound of a norm of $\mathbf{w}$ (for an arbitrary norm).

Penalty functions. A *penalty function* is a convex, non-negative function $\Omega : \mathbb{R}^m \to \mathbb{R}$ which is added to an objective function for the following reasons:

- to avoid overfitting of an objective function ℓ;
- to ensure *parsimony* of the model, e.g. to promote *sparsity* via the L_1 penalty term, and/or to control coefficients magnitude via the L_2 regularization;
- in this contribution, our particular goal is to enforce integrity of a model via a penalty term.

Regularized objective. Given two convex functions ℓ and Ω, and a positive real number λ, the *λ-regularized objective* is the function $\ell + \lambda\Omega$. In the context of the Lagrangian theory, this formulation can be seen as the *soft* formulation of the *hard* constrained problem $\min_{\mathbf{w}:\Omega(\mathbf{w})\leq k} \ell(\mathbf{w})$, with a latent correspondence between parameters λ and k. It can also be considered as a convex surrogate objective for the bi-objective minimization problem $\min(\ell, \Omega)$. From this viewpoint, λ is the *price* regulating the trade-off between ℓ and Ω. Throughout this paper, ℓ is fixed (e.g. the *Ordinal Least Squares* loss for regression, or the *log-loss* for classification), and we denote $\mathbf{w}^\star_{\lambda\Omega}$ the unique model minimizing the regularized objective $\ell + \lambda\Omega$.

Level sets. Given a function $\phi : \mathbb{R}^m \to \mathbb{R}$, and a real number k, we denote $\mathcal{B}^\phi_k := \{\mathbf{w} \in \mathbb{R}^m : \phi(\mathbf{w}) \leq k\}$ the *level set* of ϕ for value k. Thus, $\mathcal{B}^{\Omega^{L_1}}_k$ is the closed ball for the L_1 norm centered on the origin of radius k, and $\mathcal{B}^{\Omega^{L_2}}_k$ is the closed ball for the L_2 norm centered on the origin of radius k[2].

Proximal operators. Given a penalty function Ω, a positive real number μ and a model $\mathbf{w}$, the function $\mathbb{R}^m \to \mathbb{R}, \mathbf{v} \mapsto \mu\Omega(\mathbf{v}) + \frac{1}{2}\|\mathbf{v} - \mathbf{w}\|_2^2$ is strictly convex and therefore has a unique minimizer, allowing to define the *proximal operator* of the function Ω:

$$\text{Prox}_{\mu\Omega} : \mathbb{R}^m \to \mathbb{R}^m, \mathbf{w} \mapsto \underset{\mathbf{v}\in\mathbb{R}^m}{\arg\min} \frac{1}{2}\|\mathbf{v} - \mathbf{w}\|_2^2 + \mu\Omega(\mathbf{v}). \tag{1}$$

When Ω is separable, i.e. $\Omega : \mathbf{w} \mapsto \sum_{j=1}^m \Omega_j(w_j)$, computing its proximal operator is equivalent to finding the intersection between the graphical representation of the subgradient $\partial\Omega_j$ and the line $y = (w - x)/\mu$ in the 2-dimensional space. The proximal operators of some widely-used penalty functions can be found in the literature, e.g., in (Bach et al. 2012; Bauschke and Combettes 2017), usually with a focus on norms and sparsity-inducing functions. In particular:

- the L_2 penalty term leads to *shrinkage*: $\text{Prox}_{\mu\Omega^{L_2}} : \mathbf{w} \mapsto \frac{\mathbf{w}}{1+\mu}$;
- the L_1 penalization leads to *soft thresholding*: $\text{Prox}_{\mu\Omega^{L_1}_{1D}} : w \mapsto \text{sign}(x)(|x| - \mu)_+$.

3 The Facets Penalty Term

In this section, we introduce the *Facets* regularizer and discuss its properties.

3.1 The Facets Function and its Subgradient

Without loss of generality, suppose $m = 1$, and we consider a 1-dimensional problem. In order to obtain integer coefficients, we can use the integer indicator function $\mathbb{1}_\mathbb{Z} : w \mapsto 1$, if $w \in \mathbb{Z}$, and 0 otherwise. Unfortunately, this function is far from being convex, and its optimisation is not straightforward. We propose rather to consider the following penalty functions.

Definition 1. *Let $\boldsymbol{\alpha} = (\alpha_i)_{i\in\mathbb{N}}$ be a sequence of strictly positive integers. The $\boldsymbol{\alpha}$-Facets penalty in the one-dimensional case is defined as*

$$\Omega^{\boldsymbol{\alpha}-Facets}_{1D} : w \mapsto \sum_{i=0}^{\infty} \alpha_i \max(0, |w| - i).$$

This penalty may seem arbitrary, but it is not. In fact, we can prove that in the 1D case, it is the only penalty satisfying a few natural properties:

Proposition 1. *(Characterization of the $\boldsymbol{\alpha}$-Facets penalty). A one-dimensional penalty function Ω_{1D} satisfies the following properties* if and only if *it is a $\boldsymbol{\alpha}$-Facets penalty for some sequence $\boldsymbol{\alpha}$ of strictly positive integers.*

1. ***Nullity.*** $\Omega_{1D}(0) = 0$.
2. ***Even penalty.*** $\Omega_{1D}(w) = \Omega_{1D}(-w)$ *for all* $w \in \mathbb{R}$.
3. ***Integrality.*** *If the objective function is linear, then adding our penalty always yields integer weights. More precisely, define $F(\delta) = \arg\min_{w\in\mathbb{R}} \ell_\delta(w) + \lambda\Omega_{1D}(w)$ where ℓ_δ is the linear objective function $w \mapsto \delta w$. Let $D = \{\delta \in \mathbb{R} : card(F(\delta)) = 1\}$ be the set of all values $\delta \in \mathbb{R}$ on which the solution to the minization problem is unique. Then, the image of D under F is $\mathbb{Z}$.*

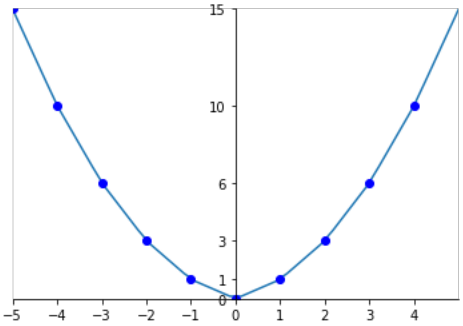

Figure 1: Graphical representation of Ω_{1D}^{Facets}.

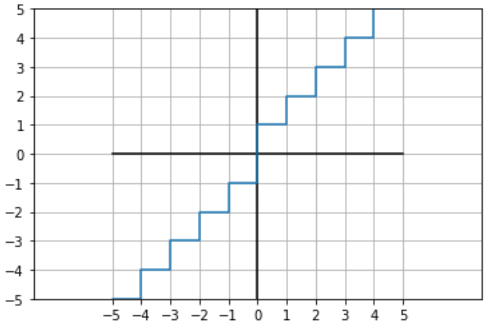

Figure 2: Subgradient of Ω_{1D}^{Facets}.

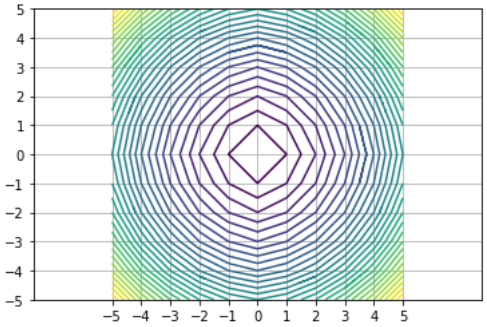

Figure 3: Level sets of Ω^{Facets}.

Building upon these 1D penalty function, we now define the multidimensional penalty as follows:

$$\Omega^{\boldsymbol{\alpha}-Facets} : \mathbf{w} \mapsto \sum_{j=1}^{m} \Omega_{1D}^{\boldsymbol{\alpha}-Facets}(w_j).$$

The choice of the $\boldsymbol{\alpha}$ sequence has a large impact on the results. Intuitively, if for some $i, j \in \mathbb{N}$ we have $\alpha_i > \alpha_j$, the $\Omega^{\boldsymbol{\alpha}-Facets}$ penalty will favor integer i over integer j. Thus, in order not to favour any integer, we will choose $\boldsymbol{\alpha} = (1, 1, \ldots)$. For the sake of clarity of notation, we will omit the symbol $\boldsymbol{\alpha}$ in the penalties from now on.

Proposition 2 (Properties of the Facets Penalty)**.**

1. $\Omega_{1D}^{Facets} : w \mapsto \int_0^{|w|} \lceil x \rceil dx$,
2. *The subgradient of* Ω_{1D}^{Facets} *is odd. For* $w \in [0, +\infty)$*, it is given by*

$$\partial\Omega_{1D}^{Facets}(w) = \begin{cases} \{\lceil w \rceil\}, & \text{if } w \in (0, +\infty) \setminus \mathbb{N}; \\ [w, w+1], & \text{if } w \in \mathbb{N}^\star; \\ [-1, 1], & \text{if } w = 0. \end{cases} \tag{2}$$

The partial subgradient of Ω^{Facets} *wrt coordinate* $j \in \{1, \ldots, m\}$ *is* $\partial\Omega_j^{Facets}(\mathbf{w}) = \partial\Omega_{1D}^{Facets}(w_j)$.

3. Ω^{Facets} *can be computed in closed form:*

$$\Omega^{Facets}(\mathbf{w}) = \sum_{j=1}^{m} \frac{\lfloor w_j \rfloor (\lfloor w_j \rfloor + 1)}{2} + (\lfloor w_j \rfloor + 1)(w_j - \lfloor w_j \rfloor). \tag{3}$$

Figure 1 illustrates the function Ω_{1D}^{Facets}, and Figure 2 depicts its subgradient $\partial\Omega_{1D}^{Facets}$.

4 Properties of the Facets-Regularized Optimal Solution

Here we provide some properties of the model $\mathbf{w}^\star_{\lambda\Omega^{Facets}}$ obtained by minimizing the regularized risk. We discuss its low magnitude, high integrity, and its ability to correctly represent a learning set, and generalize beyond it. The intuition behind the theoretical properties is provided by the level sets of Ω^{Facets}, depicted on Figure 3. Indeed, the *regularized* problem $\min_{\mathbf{w}\in\mathbb{R}^m} \ell(\mathbf{w}) + \lambda\Omega(\mathbf{w})$ and the *constrained* problem $\min_{\mathbf{w}\in\mathcal{B}_k^\Omega} \ell(\mathbf{w})$ are tightly related, with a latent correspondence between the parameters λ and k. Therefore, the shape of the level sets $\mathcal{B}_k^\Omega$ tells a lot about the properties of the minimizer. Precisely, in the case of Ω^{Facets}:

- The level sets have different shapes for different k: while the innermost sets (small k) are squares, namely, a L_1 ball, the outer sets (bigger k) are increasingly refined approximations of a circle, of a L_2 ball. This behavior is a consequence of the *inhomogeneity* of the Facets penalty, and we propose to leverage it via *scaling* which we consider further in the paper.
- The level sets are polyhedra — *facets*. The facets, to be precise, the angles cause that a number of weights are integers, similarly to the L_1 norm which sets a number of parameters to zero due to the Karush–Kuhn–Tucker conditions satisfied by the minimizer.

4.1 Weights Magnitude

The following result states that the models magnitude can be arbitrarily controlled by a hyperparameter.

Claim 1. $\|\mathbf{w}^\star_{\lambda\Omega^{Facets}}\| \xrightarrow{\lambda\to+\infty} 0$.

The Facets term adds a penalty that is stronger than the L_1 and the squared L_2 norms of the weight vector: $\forall \mathbf{w} \in \mathbb{R}^m$, $\Omega^{Facets}(\mathbf{w}) \geq \Omega^{L_1}(\mathbf{w})$, with equality if and only if $\|\mathbf{w}\|_\infty \leq 1$, and $\forall \mathbf{w} \in \mathbb{R}^m$, $\Omega^{Facets}(\mathbf{w}) \geq \Omega^{L_2}(\mathbf{w})$, with equality if and only if $\mathbf{w} = 0$. This leads to the following inclusions for level sets: $\forall k \geq 0, \mathcal{B}_k^{Facets} \subseteq \mathcal{B}_k^{L_1}$ and $\mathcal{B}_k^{Facets} \subseteq \mathcal{B}_k^{L_2}$. Moreover, elementary calculus yields that,

for all models $\mathbf{w} \in \mathbb{R}^m$:

$$\frac{\|\mathbf{w}\|_1 + \|\mathbf{w}\|_2^2}{2} \leq \Omega^{\textit{Facets}}(\mathbf{w}) \leq \frac{\|\mathbf{w}\|_1 + \|\mathbf{w}\|_2^2 + \frac{m}{4}}{2}. \quad (4)$$

4.2 PAC Setting

We prove that the Facets penalty adds *regularity* to the learning process, so that the estimated model is guaranteed to improve as the number of iterations increases.

Claim 2. *The risk and margin bounds of model* $\mathbf{w}^\star_{\lambda\Omega^{\textit{Facets}}}$ *are at least as good as those of* $\mathbf{w}^\star_{\lambda\Omega^{L_1}}$ *and* $\mathbf{w}^\star_{\lambda\Omega^{L_2}}$.

Recall the definition of the Rademacher complexity of a function class $\mathcal{F}$:

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(x_i)\epsilon_i\right], \quad (5)$$

where the ϵ_i are random variables that take values in $\{-1, +1\}$ with equal probability.

As a consequence of the results on the parameters magnitude, for a given radius $k \geq 0$, the Rademacher complexity of linear predictors with small magnitude weight vectors $\mathcal{B}_k^{L_1}, \mathcal{B}_k^{L_2}, \mathcal{B}_k^{\textit{Facets}}$ satisfy:

$$\mathcal{R}_n(\mathcal{B}_k^{\textit{Facets}}) \leq \mathcal{R}_n(\mathcal{B}_k^{L_1}) \leq X_\infty W\sqrt{2\log(2m)}n^{-1/2}, \quad (6)$$

$$\mathcal{R}_n(\mathcal{B}_k^{\textit{Facets}}) \leq \mathcal{R}_n(\mathcal{B}_k^{L_2}) \leq X_2 W n^{-1/2}. \quad (7)$$

This leads to risk and margin bounds similar to those provided in (Kakade, Sridharan, and Tewari 2009).

4.3 Integrity

Define a solution of the regularized problem $\mathbf{w}^R_{\Omega,\lambda} = \arg\min_{\mathbf{w}\in\mathbb{R}^m} \ell(\mathbf{w}) + \lambda\Omega(\mathbf{w})$ (where superscript R stands for *regularized*). It is strongly tied to that of the constrained problem $\mathbf{w}^C_{\Omega,k} = \arg\min_{\mathbf{w}\in\mathcal{B}_k^\Omega} \ell(\mathbf{w})$ (superscript C stands for *constrained*).

What the Facets offers on top of shrinkage and selection, is *integrity*: as depicted by Figure 3, the minimizers in $\mathbf{w}^C_{\Omega^{\textit{Facets}},k}$ are likely to be found at one of the many vertices of the polyhedron $\mathcal{B}_k^{\textit{Facets}}$, where some coefficients w_j are integers.

We expect the optimal parameter $\mathbf{w}^R_{\Omega^{\textit{Facets}},\lambda}$ to achieve a good level of *integrity*, with a magnitude and an accuracy comparable to one achieved by the state-of-the-art $\mathbf{w}^R_{\Omega^{L_2},\lambda}$ of ridge regression. We consider $\mathbf{w}^R$ further in the paper, and denote it $\mathbf{w}$ for simplicity.

Scaling. Given a positive real number γ, γ-*scaling* is an operator that transforms a penalty function Ω into a scaled penalty function $\textit{scaled}_\gamma(\Omega)$ such that:

$$\textit{scaled}_\gamma(\Omega) : \mathbf{w} \mapsto \Omega\Big(\frac{\mathbf{w}}{\gamma}\Big). \quad (8)$$

The hyperparameter γ can be interpreted as the unit length of the parameter scale. Accordingly, it seems relevant to revise our notion of *integrity*, in order to account for the target $\gamma\mathbb{Z}$ scale:

$$\gamma - \textit{integrity} : \mathbb{R}^m \to [0,1], \mathbf{w} \mapsto \frac{|\{j : w_j \in \gamma\mathbb{Z}\}|}{m}. \quad (9)$$

Note that the majority of penalty functions proposed in the literature are *absolutely homogeneous* (and often *norms*), so that for any positive real a, $\Omega(a\mathbf{w}) = |a|\Omega(\mathbf{w})$. In such a case, λ-pricing and γ-scaling are redundant, as $\lambda\ \textit{scaled}_\gamma(\Omega) \equiv \frac{\lambda}{\gamma}\Omega$. Conversely, as $\Omega^{\textit{Facets}}$ is deliberately inhomogeneous, the two hyperparameters λ and γ should enable to select separately the size, or *strength* of the level set, governed by the shrinkage effect of the penalization, and its *shape*, and the number of facets. We assume that the scaling allows us to increase accuracy and simplicity of the model.

5 Efficient Minimization of the Facets-Regularized Risk

In this section, we discuss the optimisation issues of the Facets penalty term.

Regularized risk minimizers are theoretical objects that we cannot compute directly, but rather try to approximate through an optimization algorithm, that yields a sequence $\langle\mathbf{w}^t\rangle_{t\in\mathbb{T}}$ of iterates. While magnitude and accuracy are convex and continuous properties of the parameter $\mathbf{w}$, this is not the case for integrity. Therefore, even in the case of a fully integral limit $\mathbf{w}^\star_{\lambda\Omega^{\textit{Facets}}} \in \mathbb{Z}^m$, it is quite possible that the iterates have low, or even zero, integrity. Therefore, it is of utmost importance to select carefully the algorithm performing the optimization.

We consider the *operator splitting* approach, widely used for non-smooth optimization and already known to favor sparsity under sparsity-inducing regularization. We give a brief overview of the *Proximal Gradient Descent* algorithm, and we give a closed-form expression of the proximity operator of the *Facets* penalty allowing its efficient implementation. We also introduce Strongly Convex Facets that add *elasticity*, similarly to the Elastic Net penalty (Zou and Hastie 2005), facilitating both the theoretical analysis of the algorithm and its performance.

5.1 Proximal Gradient Descent

Proximal algorithms (sometimes called *operator splitting* methods) (Moreau 1965; Parikh, Boyd, and others 2014) were developed to minimize an objective function $\ell + \Omega$, where ℓ is a smooth differentiable function with Lipschitz-continuous gradient, while Ω is a non-differentiable function. *Iterative Shrinkage-Thresholding Algorithm* (ISTA), introduced by (Daubechies, Defrise, and De Mol 2004; Beck and Teboulle 2009), which is a *Proximal Gradient Descent* algorithm, is a two-step fixed-point scheme *à la* Picard. It is based on the assumption that, even though the function Ω might be non-differentiable, the optimization problem defining its proximity operator can be solved efficiently. At each time step $t \in \mathbb{T}$ of ISTA, given a step size $\tau^t > 0$:

1. the smooth function ℓ is linearized around $\mathbf{w}^t$ so $\ell(\mathbf{w}) \approx \ell(\mathbf{w}^t) + (\mathbf{w} - \mathbf{w}^t)\cdot\nabla\ell(\mathbf{w}^t)$, and optimized by a *forward*

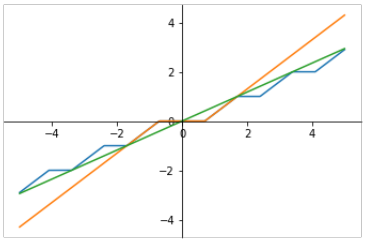

Figure 4: Proximity operators of penalty functions: $\Omega_{1D}^{L_1}$ (in orange), $\Omega_{1D}^{L_2}$ (in green), and Ω_{1D}^{Facets} (in blue).

gradient step:

$$\mathbf{w}^{t+\frac{1}{2}} \leftarrow \mathbf{w}^t - \tau \nabla \ell(\mathbf{w}); \tag{10}$$

2. the non-smooth Ω is augmented by a proximal regularization term proportional to $\|\mathbf{w} - \mathbf{w}^t\|^2$, in order to i) keep the update close to the previous point, where the linear approximation of ℓ is reasonable; ii) to ensure that the regularization term is strictly convex and smooth; and iii) to ensure the descent of $\mathbf{w}^t$ towards the minimizer of $\ell + \Omega$. The optimization of this term is done via a *backward* (implicit) *proximal step*:

$$\mathbf{w}^{t+1} \leftarrow \text{Prox}_{\tau\Omega}(\mathbf{w}^{t+\frac{1}{2}}). \tag{11}$$

5.2 Proximal Operator of the Facets Penalty

Fortunately, the proximal operator of Ω^{Facets} can also be efficiently computed in a closed form.

Proposition 3. *For all* $\mu \in [0, +\infty[$, *for all* $\mathbf{w} \in \mathbb{R}^m$, $Prox_{\mu\Omega^{Facets}}(\mathbf{w}) = (sign(w_1) \cdot v_1, \ldots, sign(w_m) \cdot v_m)$, $j \in \{1, \ldots, m\}$:

$$v_j = \left\lfloor \frac{|w_j|}{\mu + 1} \right\rfloor + \left(|w_j| - (\mu + 1) \left\lfloor \frac{|w_j|}{\mu + 1} \right\rfloor - \mu \right)_+ . \tag{12}$$

The proof is provided in the supplementary material.

Figure 4 compares the proximity operators of Ω^{Facets}, Ω^{L_1}, and Ω^{L_2} [1]. The curve representing $\text{Prox}_{\mu\Omega_{1D}^{Facets}}$ follows the general trend given by $\text{Prox}_{\mu\Omega_{1D}^{L_2}}$, which is a straight line with slope $\frac{1}{1+\mu}$, but instead of a constant slope, it displays a plateau of width μ followed by a 45 degrees slope where $\Delta x = \Delta y = 1$, what is identical to the behavior of $\text{Prox}_{\mu\Omega_{1D}^{L_1}}$ between 0 and $1 + \mu$.

5.3 Strongly Convex Facets

The Facets penalty is neither strongly, nor strictly convex, as a result of its locally constant subgradient. This is a disadvantage, since it provokes a number of optimisation problems, such as absence of unique solution, procedural regularity violations, slow convergence rate, etc.

[1] Interestingly, the same functions and diagrams appear in (Hastie, Tibshirani, and Friedman 2009), without any reference to proximity operators. Soft thresholding and shrinkage appear as the modification of a regression problem penalized by ordinary least squares, when adding respectively L_1 and L_2 penalization, when the observation matrix is orthogonal.

In order to enforce the strong convexity of the penalty, we tweak the subgradient of the Facets penalty by adding a separable correcting term Ω^{corr}, so that, for $w \geq 0$,

$$\partial\Omega_{1D}^{\text{corr}}(w) = (w - \lfloor w \rfloor) = \begin{cases} 0, & \text{if } w \in \mathbb{Z}; \\ w + 1 - \lceil w \rceil, & \text{otherwise.} \end{cases} \tag{13}$$

Consequently,

$$\Omega^{\text{corr}} = \Omega^{L_2} + \Omega^{L_1} - \Omega^{Facets}. \tag{14}$$

Hence, for $0 < \epsilon < 1$, the Strongly Convex Facets (SCF) function defined by

$$\Omega^{SCF_\epsilon} := \Omega^{Facets} + \epsilon\Omega^{\text{corr}} = (1 - \epsilon)\Omega^{Facets} + \epsilon(\Omega^{L_1} + \Omega^{L_2}) \tag{15}$$

is symmetric, null at $\mathbf{0}$, and ϵ-strongly convex.

The proximity operator of this modified penalty can be efficiently computed as follows:

$$\left|\text{Prox}_{\mu\Omega_{1D}^{SCF_\epsilon}}(w)\right| = \lfloor a \rfloor + \min\left(\frac{1+\mu}{1+\mu\epsilon}(a - \lfloor a \rfloor), 1\right),$$
$$\text{with } a = \left(\frac{|w| - \mu}{1 + \mu}\right)_+ \tag{16}$$

The correcting term modifies the proximity operator of the *Facets* penalty in the following manner: the width of the plateaus (except the one around zero) is shortened by a length $\mu\epsilon$, while the width of the slopes is increased by $\mu\epsilon$, and the resulting operator is now conveniently $(1+\mu\epsilon)^{-1} <$ 1 Lipschitz continuous[2].

Strict convexity entails the uniqueness of the minimizer of the regularized objective. In turn, this property provides resilience to potential correlations between features. Consider a situation where features j and j' are *clones*, i.e. for all data points i, $x_{i,j} = x_{i,j'}$. In this case, obviously, the loss function is blind to trade-offs between w_j and $w_{j'}$, and, because it is piecewise linear, so might be the Facets loss (even though this behavior tends to be localized, as opposed to the issues encountered by the Lasso). This behavior is not desired, as it could result in a violation of *procedural regularity* (Kroll et al. 2016). The principle of *equal treatment of equals* imposes that the features j and j' receive equal attention, so that the j and j' coordinates of $\mathbf{w}_\Omega^\star$ are equal, but this is unlikely to happen if the penalty Ω is not strictly convex (as it is the case for Ω^{Facets} or Ω^{L_1}). While cloning might be considered as an extreme situation, maybe resulting from an adversarial behavior, the issue of having a non-unique minimizer might arise as soon as the observation matrix is not full column rank, i.e. when some features are (strongly) correlated.

Scaling. We already mentioned the idea of *scaling* the Ω^{Facets} penalty term to obtain a meaningful additional hyperparameter. Fortunately, scaling interacts smoothly with

[2] This is indeed a particular case of a more general result, found in e.g. (Bauschke and Combettes 2017), tying strongly convex regularization and shrinkage: a α-strongly convex function has a α-strongly monotone subgradient, and, therefore, its proximity operator is Lipschitz continuous with constant $(\mu\alpha + 1)^{-1} \in]0, 1[$.

proximal calculus (see e.g. (Bauschke and Combettes 2017), proposition 24.8):

$$\text{Prox}_{\mu\, scaled_\gamma(\Omega)} = \gamma\, scaled_\gamma(\text{Prox}_{\frac{\mu}{\gamma^2}\Omega}). \quad (17)$$

In the cases of Ω^{Facets} and Ω^{SCF_ϵ}, γ-scaling simultaneously divides the length of the plateau by γ, and multiplies both the width and height of the slope by γ.

5.4 Computational Efficiency

Strong convexity leads to computational benefits.

Claim 3. *When applied to the Strongly Convex Facets regularizer, the proximal gradient algorithm enjoys linear convergence, i.e.*

$$\|w^t - w^\star_{\lambda\Omega^{SCF_\epsilon}}| \leq (\lambda\tau\epsilon + 1)^{-t}\|w^0 - w^\star_{\lambda\Omega^{SCF_\epsilon}}\|. \quad (18)$$

A precise (and convoluted) demonstration can be found in (Bauschke and Combettes 2017), example 28.12. It can be briefly summarized as follows:

- the forward gradient step $w \mapsto w - \tau\nabla\ell(w)$ is non-expansive when $\nabla\ell$ is $2/\tau$-Lipschitz continuous;
- the backward proximal step $\text{Prox}_{\lambda\tau\Omega^{\text{EF}}_\epsilon}$ is a $(\lambda\tau\epsilon + 1)^{-1}$-contraction.

Linear convergence follows from the Banach-Picard fixed-point theorem applied to the forward-backward operator consisting in alternating these two steps.

This fast convergence should be compared to the much more modest performance achieved by PGD/ISTA in the general case, which is $O(t^{-1})$ (or $O(t^{-2})$ for the Nesterov-accelerated version FISTA) (Beck and Teboulle 2009).

We discuss possible acceleration scenarios in the supplementary material.

5.5 Regularization Path

Strong convexity of the penalty leads to a proper optimization problem[3]. We can therefore define the *regularization path* RP as the function mapping hyperparameters to the (unique) minimizer of the regularized objective:

$$RP_\epsilon : (\lambda, \gamma) \mapsto \mathbf{w}^\star_{\lambda\, scaled_\gamma(\Omega^{SCF_\epsilon})}. \quad (19)$$

Claim 4. *The hyperparameters provide smooth control over the selected model, as RP_ϵ is continuous over $]0, +\infty[\times]0, +\infty[$.*

We provide the proof in the supplementary material.

6 Experiments: a Case Study

We are particularly interested in medical applications, and we consider the Heart Disease data set (downloadable from the UCI Machine Learning repository). The data set contains information about 303 patients and 75 features, however, only 14 features are really used in all previous studies on this data.

[3]Contrast this clean-cut situation with the convoluted discussion about 'having a single solution when the columns of the observation matrix are in *general position*' surrounding the Lasso regularization.

We implemented the proposed approach in Python, and the implementation will be publicly available as soon as the paper is de-anonymised.

To fix the hyperparameters, we apply an extensive grid search over (λ, γ). It is the trade off between λ and γ which is important, let $\alpha := \lambda/\gamma$, and let $\beta := \lambda/\gamma^2$. We perform the 10-fold cross validation, and plot the mean values. Figure 5 illustrates accuracy as a function of the shape and strength of the Facets regularizer. The first remark is that we achieve the state-of-the-art performance on the data set. Figure 6 shows the integrity (in blue) and sparsity (in orange) for the corresponding models. It is easy to see that optimal (in generalizing accuracy) models reach also the highest sparsity and integrity.

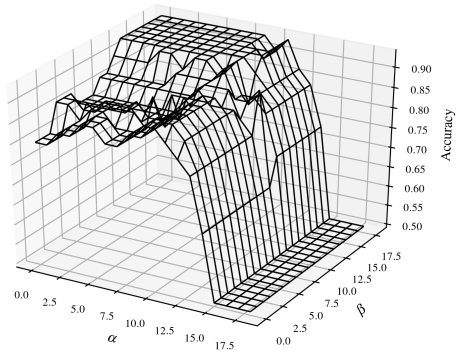

Figure 5: Heart Disease data. Accuracy as a function of (λ, γ), $\alpha = \lambda/\gamma, \beta = \lambda/\gamma^2$.

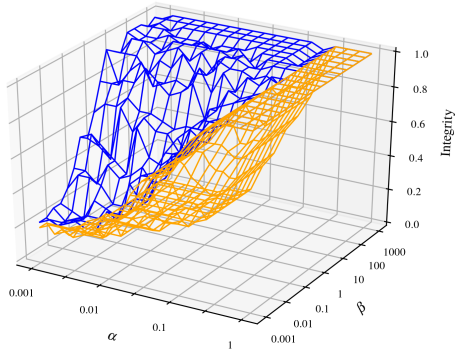

Figure 6: Heart Disease data. Integrity in blue, sparsity in orange. Note that integrity is always bigger than sparsity, and often strictly bigger (especially if α is small).

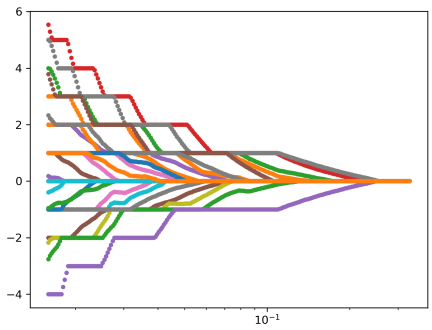

Figure 7: Heart Disease data. Scatter plot of $\frac{\mathbf{w}_j^\star}{\gamma}$ as a function of α. Different colors correspond to different values of j, $\beta \approx 0.3$.

Regularization paths. As argued above, we are not using the hyperparameters (λ, γ), but $\alpha = \lambda/\gamma$ for a fixed value of τ. The scaled values of $\frac{\mathbf{w}^\star}{\gamma}$ along the regularization path are represented on Figure 7. It displays two striking properties:

Integrity: the many plateaus corresponding to values of $\mathbf{w}_j^\star \in \gamma\mathbb{Z}$ testify the success of our attempt.

Continuity: the scatterplot looks a lot like a curve, with a smooth evolution of the coefficients $\mathbf{w}_j^\star$ when α varies. This continuous aspect testifies a good functioning of the hyperparameters as *knobs* permitting to steer the optimization process.

7 Discussion: Why Small Integer Weights?

Models with small integer weights have several clear advantages. We mention some of them below.

- Accuracy and prevention of overfitting. The importance to control the magnitude of the parameters is well explained in (Kakade, Sridharan, and Tewari 2009), through the upper bound on the Rademacher complexity of the hypothesis class. In the same vein, favoring (small) integers prevents a learning algorithm from unnecessarily fine-grained solutions.
- Reduced memory footprint. Small integers take little RAM. This is the motivation behind the Google's results described in (Golovin et al. 2013; McMahan et al. 2013). The aim is to learn simple prediction models that can be replicated on highly distributed systems, and that require very little unitary bandwidth to process billions of requests.
- Procedural regularity and user empowerment. Sparse linear models with small integers can be easily used to make quick predictions by human experts, without computers. Such models are transparent for users, and can be efficiently used in criminalistics (Rudin, Wang, and Coker 2019), and medicine (Ustun and Rudin 2016).
- Sparsity and interpretability. Favoring integrity can be seen as an instance of structured risk minimization (Vapnik 1990). This intuition is made more explicit in (Belahcene et al. 2019), where the positive integer weights of a linear model are interpreted as a number of repetitions of premises of a *ceteris paribus* reasoning, similarly to the coefficients mentioned by Benjamin Franklin in his *Moral Algebra*. Integrity is a requirement for interpretability, while magnitude is a proxy for simplicity.
- Explainable AI. There exists theoretical and practical importance to be able to explain power indices, such as the Shapley's index, in order to interpret the importance of a feature. To illustrate this issue, consider a linear model with three features taking values in $\{0, 1\}$, with the corresponding weights $w_1 = w_2 = 0.49$, and $w_3 = 0.02$, and an intercept equal to -0.5. One could conclude that features 1 and 2 are far more important than feature 3. In a game-theoretic approach, one considers various combinations of features. It then becomes clear that this model is equivalent to the decision rule "at least two features present". While magnitude alone does not help (consider dividing the weights by 100), nor integrity (consider multiplying the weights by 100), their cumulative effect could lead to a model with weights $w_1 = w_2 = w_3 = 1$, and an intercept of -2, that faithfully reflects the respective influence of each feature.
- Knowledge discovery. Very small integers can be directly interpreted, such as $0/1$ – presence/absence, or $1/-1/0$ – friend/foe/neutral, and to reveal biologically relevant relationships in complex ecosystems.

8 Conclusion

We proposed a novel principled method to learn a model with integer weights via soft constraints. We introduced a new penalty term called Facets. Our main theoretical results provide some theoretical foundations of our approach.

The main claim of our contribution is that the novel Facets penalization can be used to efficiently learn sparse linear models with small integer weights.

The numerical experiments – the case study of a real-world medical application – illustrates practical efficiency of the proposed method. Currently we challenge to accelerate and to increase the stability of the optimisation procedure. Another important research direction is to apply our novel methodology to real hospital data, and to construct real medical scores which can be integrated into clinical routines.

Supplemental Material

Proposition 1 (Characterization of the $\boldsymbol{\alpha}$-Facets penalty). A one-dimensional penalty function Ω_{1D} satisfies the following properties *if and only if* it is a $\boldsymbol{\alpha}$-Facets penalty for some sequence $\boldsymbol{\alpha}$ of strictly positive integers.

1. **Nullity.** $\Omega_{1D}(0) = 0$.
2. **Even penalty.** $\Omega_{1D}(w) = \Omega_{1D}(-w)$ for all $w \in \mathbb{R}$.
3. **Integrality.** If the objective function is linear, then adding our penalty always yields integer weights. More precisely, define $F(\delta) = \arg\min_{w \in \mathbb{R}} \ell_\delta(w) + \lambda\Omega_{1D}(w)$ where ℓ_δ is the linear objective function $w \mapsto \delta w$. Let $D = \{\delta \in \mathbb{R} : card\,(F(\delta)) = 1\}$ be the set of all values $\delta \in \mathbb{R}$ on which the solution to the minization problem is unique. Then, the image of D under F is $\mathbb{Z}$.

Proof. (sketch) The *if* part of the proof is straighforward, the reader can check that $\boldsymbol{\alpha}$-Facets penalties satisfy these conditions. Let us focus on the *only-if* part. For the sake of clarity, let $\Omega : \mathbb{R} \to \mathbb{R}$ be our 1D penalty function, which is, be definition, convex and non-negative.Assume Ω satisfies the three properties stated in the proposition.

Let us first show that Ω is piece-wise linear. Assume that Ω is twice differentiable on an open interval $]a, b[$, and $\Omega''(x) > 0$ for $x \in]a, b[$. Let $z \in]a, b[$. Then, by Taylor's theorem on Ω', for any $y \in]a, b[$ we have: $\Omega'(y) = \Omega'(z) + \Omega''(z)(y - z) + (y - z)o(1)$. Thus, $\Omega'(y) - \Omega'(z) = (y - z)\,(\Omega''(z) + o(1))$. Because the $o(\cdot)$ term tends to zero, there exists $\hat{y} \in]a, b[$ with $|\hat{y} - z| < 1$ such that $\Omega'(\hat{y}) - \Omega'(z) \neq 0$. Define $\ell_{\hat{y}}(w) = -w\lambda\Omega'(\hat{y})$ and $\ell_z(w) = -w\lambda\Omega'(z)$. Let $f_{\hat{y}}(w) = \ell_{\hat{y}}(w) + \lambda\Omega(w)$ and $f_z(w) = \ell_z(w) + \lambda\Omega(w)$. Clearly, $f'_{\hat{y}}(\hat{y}) = 0$ and $f'_z(z) = 0$. Because $\Omega''(x) > 0$ for $x \in]a, b[$, $\hat{y} = \arg\min_w f_{\hat{y}}(w)$ and $z = \arg\min_w f_z(w)$, and these minimizers are unique. But because $|\hat{y} - z| < 1$, at least one of these minimizers is not an integer, which contradicts the *integrity* property. Thus, on all open intervals, either Ω is non twice differentiable, either $\Omega''(x) = 0$. This caracterizes piecewise linear functions.

Next, let us show that the discontinuities of Ω' occur at each integer. If Ω' is discontinuous at x then there exists δ such that $w \mapsto \ell_\delta(w) + \lambda\Omega(w)$ is minimized at x and this minimizer is unique. So the set of discontinuities of Ω' is exactly the set of unique minimizers of $\ell_\delta(w) + \lambda\Omega(w)$. Thus, this set of discontinuities is $\mathbb{Z}$.

Finally, it is easy to show that any piecewise linear even convex fonction Ω, null at zero, such that its set of discontinuities is precisely $\mathbb{Z}$ can be written as $\Omega(w) = \sum_{i=0}^{\infty} \alpha_i \max(0, |w| - i)$. □

Proposition 3 For all $\mu \in [0, +\infty[$, for all $\mathbf{w} \in \mathbb{R}^m$, $\text{Prox}_{\mu\Omega^{Facets}}(\mathbf{w}) = (\text{sign}(w_1) \cdot v_1, \ldots, \text{sign}(w_m) \cdot v_m)$, $j \in \{1, \ldots, m\}$:

$$v_j = \left\lfloor \frac{|w_j|}{\mu + 1} \right\rfloor + \left(|w_j| - (\mu + 1) \left\lfloor \frac{|w_j|}{\mu + 1} \right\rfloor - \mu \right)_+ . \quad (20)$$

Proof. First, as Ω^{Facets} is separable, so is its proximity operator, and we only need to solve a $\mathbb{R} \to \mathbb{R}$ optimization problem. Second, as Ω_{1D}^{Facets} is even, its proximal operator is odd. Third, as, for any nonnegative x, $\Omega_i^{Facets}(x + 1) = \Omega_{1D}^{Facets}(x) + 1$, we have that $y = \text{Prox}_{\mu\Omega_{1D}^{Facets}}(x) \iff y + 1 + \mu = \text{Prox}_{\mu\Omega_{1D}^{Facets}}(x + 1)$, so the curve representing $\text{Prox}_{\mu\Omega_{1D}^{Facets}}$ in the half-plane $x \geq 0$ is invariant by translation of vector $(1 + \mu, 1)$. Finally, it is straightforward to check that, for $x \in [0, 1 + \mu[$, $\text{Prox}_{\mu\Omega_{1D}^{Facets}}$ is the *soft thresholding* operator $x \mapsto (x - \mu)_+$. □

Claim 4 The hyperparameters provide smooth control over the selected model, as RP_ϵ is continuous over $]0, +\infty[\times]0, +\infty[$.

Proof. Our argument relies on the fixed-point scheme described by equations (10) and (11). For any positive real numbers $\underline{\lambda}, \overline{\gamma}$, the function $F : \mathbb{R}^m \times [\underline{\lambda}, +\infty[\times]0, \overline{\gamma}] \to \mathbb{R}^m, (w^t, \lambda, \gamma) \mapsto \mathbf{w}^{t+1}$ is both

- continuous in (w^t, λ, γ), as the gradient step (eq. 10) is continuous, since ℓ is smooth; and the proximal step (eq. 11) is continuous because of the specific form of $\text{Prox}_{\mu\Omega_{\lambda,\gamma}}$;
- uniformly k-Lipschitz w.r.t. $\mathbf{w^t}$, independently of the values of γ and λ, with $k = 1/(1 + (\underline{\lambda}\epsilon)/\overline{\gamma}^2) < 1$.

Therefore, the limit $\mathbf{w}^\star$ of the fixed-point scheme depends continuously on the parameters (λ, γ). □

Acceleration of the Optimisation Procedure. An optimal step size is essential to accelerate the optimisation procedure, and a number of schemes are used to find an optimal sequence of τ^t governing the proximity term:

- an optimal choice is to let τ^t be the Hessian matrix of the objective function at $\mathbf{w}^t$. In this case, the function has to be twice differentiable. Although this choice leads to a faster convergence of the proximal algorithm, it also demands much more computations at each iteration.
- on the other side of the spectrum, it is possible to let τ^t be a scalar constant, with convergence guarantees for the case where it is bigger than a Lipschitz constant of the gradient of the objective function. Exactly this option is implemented in the ISTA algorithm.
- τ^t can be chosen scalar, but regularly updated. (Beck and Teboulle 2009) propose an adaptive strategy consisting in choosing τ^t just big enough to ensure that the update from $\mathbf{w}^t$ to $\mathbf{w}^{t+1}$ is indeed a descent step. (McMahan and Streeter 2010) also propose a learning rate which is an update per coordinate, decreasing in magnitude.

The ISTA is not a very fast algorithm, it requires to compute the full gradient of the objective function at each time step, and its convergence towards the minimizer of the function is in t^{-1}. Two types of acceleration techniques are widely used:

Stochastic gradient: Instead of computing the full gradient of the function, an unbiased estimation of it, e.g. by line sampling, *Online Gradient Descent*, yielding the FTRL-Proximal algorithm (McMahan et al. 2013; McMahan 2017), or column sampling *à la* Coordinate Descent can be used.

The Nesterov's trick: Instead of updating the parameters directly by the proximity operator of the gradient step, interpret this new value as a direction only, and find an update along this direction (Nesterov 2013). This leads to a t^{-2} rate of convergence. However, as integrity is not a convex property of the parameters, it might suffer from the interpolation step.

References

Bach, F. R.; Jenatton, R.; Mairal, J.; and Obozinski, G. 2012. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning* 4(1):1–106.

Bauschke, H. H., and Combettes, P. L. 2017. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer.

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2(1):183–202.

Belahcene, K.; Labreuche, C.; Maudet, N.; Mousseau, V.; and Ouerdane, W. 2019. Comparing options with argument schemes powered by cancellation. In *IJCAI*.

Chevaleyre, Y.; Koriche, F.; and Zucker, J. 2013. Rounding methods for discrete linear classification. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 651–659.

Daubechies, I.; Defrise, M.; and De Mol, C. 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 57(11):1413–1457.

Golovin, D.; Sculley, D.; McMahan, H. B.; and Young, M. 2013. Large-scale learning with less RAM via randomization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 325–333.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

Hastie, T.; Tibshirani, R.; and Wainwright, M. 2015. *Statistical learning with sparsity: the Lasso and generalisations*. CRC Press.

Kakade, S. M.; Sridharan, K.; and Tewari, A. 2009. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, 793–800.

Kroll, J. A.; Barocas, S.; Felten, E. W.; Reidenberg, J. R.; Robinson, D. G.; and Yu, H. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165:633.

McMahan, H. B., and Streeter, M. J. 2010. Adaptive bound optimization for online convex optimization. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, 244–256.

McMahan, H. B.; Holt, G.; Sculley, D.; Young, M.; Ebner, D.; Grady, J.; Nie, L.; Phillips, T.; Davydov, E.; Golovin, D.; Chikkerur, S.; Liu, D.; Wattenberg, M.; Hrafnkelsson, A. M.; Boulos, T.; and Kubica, J. 2013. Ad click prediction: a view from the trenches. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, 1222–1230.

McMahan, H. B. 2017. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research* 18:90:1–90:50.

Moreau, J.-J. 1965. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France* 93:273–299.

Nesterov, Y. 2013. Gradient methods for minimizing composite functions. *Mathematical Programming* 140(1):125–161.

Parikh, N.; Boyd, S.; et al. 2014. Proximal algorithms. *Foundations and Trends® in Optimization* 1(3):127–239.

Rudin, C.; Wang, C.; and Coker, B. 2019. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*.

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

Ustun, B., and Rudin, C. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102(3):349–391.

Vapnik, V. 1990. *The nature of statistical learning theory*. Springer.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67(2):301–320.