Learning Interpretable Models using Soft Integrity Constraints

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

Proceedings of Machine Learning Research 129:529–544, 2020 ACML 2020

Learning Interpretable Models using Soft Integrity Constraints

Khaled Belahcene KHALED.BELAHCENE@HDS.UTC.FR
Heudiasyc, UMR 7253, UTC, Compiègne, France

Nataliya Sokolovska NATALIYA.SOKOLOVSKA@SORBONNE-UNIVERSITE.FR
NutriOmics, UMR S 1269, INSERM, Sorbonne University, Paris, France

Yann Chevaleyre YANN.CHEVALEYRE@LAMSADE.DAUPHINE.FR
LAMSADE, Dauphine University, PSL Research University, UMR CNRS 7243, Paris, France

Jean-Daniel Zucker JEAN-DANIEL.ZUCKER@IRD.FR
UMI 209 UMMISCO, IRD/SU, Bondy; NutriOmics, UMR S 1269, INSERM, SU, Paris, France

Editors: Sinno Jialin Pan and Masashi Sugiyama

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 regulariser. We provide the proximal operator associated with the proposed penalty term, so that the regularized empirical risk minimiser 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 penalised by the proposed Facets penalty term leads to a model with a significant number of integer weights.

Keywords: Regularisation, interpretable models, integer models

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 maximise directly the *accuracy* of the model, however, the corresponding loss function, the 0-1 loss, is not convex and its minimisation is intractable for real-world applications. Therefore, a widely used approach is to relax the optimisation 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 *regularisation* where the objective function, subject to minimisation, 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 et al., 2009) shrinking parameters towards zero, and Lasso regularization (Tibshirani, 1996) setting a controlled (by a hyper-parameter) 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 et al., 2015).

Our contribution to interpretable models learning is multi-fold:

- 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; 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 problem of learning interpretable models from data, and specifically *compact linear integer models*, has gathered recent interest. We relate our work to three particular approaches:

- Chevaleyre et al. (2013) consider the problem of maximising accuracy subject to hard integrity and magnitude constraints, and solve it through rounding schemes. We believe our soft constraints approach to be more apt at adapting to data and find more attractive trade-offs between accuracy and magnitude of a fully integral model.
- Golovin et al. (2013), where the focus is on the reduction of memory usage and the processing of massive amounts of data. Their approach focuses on online learning and is based on randomised rounding and counting, with a diminishing learning rate and discretisation grid size. They compute a no-regret model stored as a tuple of integers, with a magnitude increasing as the square root of the number of examples. However, they are not interested in interpretability, and their model cannot be considered integral, because of the per-coordinate grid size.
- Ustun and Rudin (2016) describe a loss function expressing a linear trade-off between sparsity, magnitude and accuracy, over all integral models, then defer the NP-complete exact minimisation to a MILP solver. In comparison, we propose an actual algorithm based on convex optimisation to approximately solve a similar problem. We make use of the huge computational gain to remain agnostic considering the admissible magnitude vs accuracy trade-off and better explore its Pareto front.

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* regularisation. 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 $w_j, j \in \{1, \ldots, m\}$ 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 nonnegative function $\Omega : \mathbb{R}^m \to \mathbb{R}$ which is added to an objective function for the following reasons: (1) to avoid overfitting of an objective function ℓ; (2) 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 regularisation; (3) in this contribution, our particular goal is also to enforce integrity of a model via a penalty term.

Regularised objective. Given two convex functions ℓ and Ω, and a positive real number λ, the λ-*regularised 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 minimisation 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 *Ordinary Least Squares* loss for regression, or the *log-loss* for classification), and we denote $\mathbf{w}^\star_{\lambda\Omega}$ the unique model minimising the regularised 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 centred on the origin of radius k, and $\mathcal{B}^{\Omega^{L_2}}_k$ is the closed ball for the L_2 norm centred 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:

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

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.

3. The Facets Penalty Term

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

Ideally, in order to obtain integer weights of small L_2 norm, the optimisation problem we should solve is $\arg\min_{w\in\mathbb{Z}^m} \ell(w) + \lambda \|w\|_2^2$. This constrained optimization problem can be equivalently written in unconstrained form:

$$\arg\min_{w\in\mathbb{R}^m} \ell(w) + \lambda\Omega^{smallint}(w),$$

where $\Omega^{smallint}(w) = \|w\|_2^2 + \mathbb{1}_C(w)$, and where $\mathbb{1}_C(w) = 0$, if $w \in C$, and $\mathbb{1}_C(w) = \infty$, otherwise; $C = \mathbb{Z}$, is the domain where the parameters take their values. Unfortunately, this function is not convex, and classical learning problems where this function appears are computationally intractable. To make this problem tractable, we will build a convex approximation of $\Omega^{smallint}$. A standard way of building a convex approximation of a given function is to take its *lower convex envelope (LCE),* defined as the supremum over all convex functions that lie under that function.

For the sake of simplicity, we will first consider the 1-dimensional case ($m = 1$). It turns out that the 1-dimensional *LCE* of the above non-convex penalty belongs to a wider class of functions of independent interest, which we call α-Facets penalties:

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*

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

We can now show how this penalty relates to the non-convex $\Omega^{smallint}$ function.

Proposition 2 *The $\boldsymbol{\alpha}$-Facets penalty function with $\boldsymbol{\alpha} = (1, 1, \ldots)$, which we will now refer to as Ω_{1D}^{Facets}, is the* LCE *of $\Omega^{smallint}$ in 1 dimension.*

In the rest of this paper, we will build upon the Ω_{1D}^{Facets} penalty. However, as mentioned earlier, the whole family of $\boldsymbol{\alpha}$-Facets penalties is of high interest: in fact, the following proposition states that any reasonable convex penalty enforcing integrity must be a $\boldsymbol{\alpha}$-Facets penalty.

Proposition 3 (Characterisation 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 straightforward, 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 minimisers are unique. But because $|\hat{y} - z| < 1$, at least one of these minimisers 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 caracterises 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 minimised at x and this minimiser is unique. So the set of discontinuities of Ω' is exactly the set of unique minimisers 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 $\Omega(w) = \sum_{i=0}^{\infty} \alpha_i \max(0, |w| - i)$. ■

Note that the above proposition provides us with guarantees in the case of linear loss functions. For general convex losses, no such integrality guarantee can be provided, because the optimisation problem is known to be NP-hard (Chevaleyre et al., 2013).

Let us now extend the Ω_{1D}^{Facets} to the multi-dimensional case:

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

Proposition 4 (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:*

$$\Omega_{1D}^{Facets}(w) = \frac{\lfloor w \rfloor(\lfloor w \rfloor + 1)}{2} + (\lfloor w \rfloor + 1)(w - \lfloor w \rfloor). \tag{3}$$

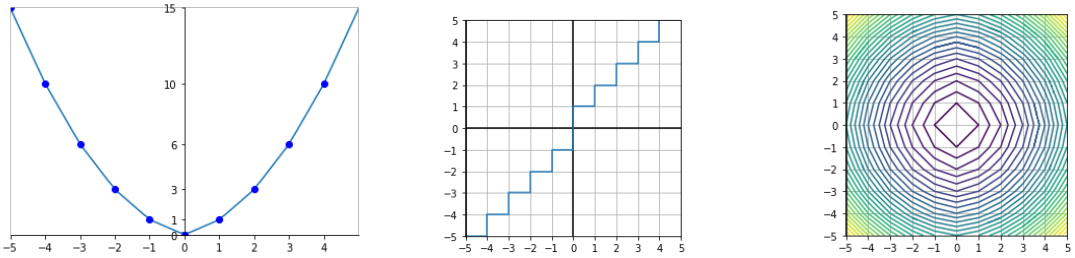

Figure 1: Graphical representations of Ω^{Facets}. On the left: penalty term Ω^{Facets}_{1D}, in the center: subgradient of Ω^{Facets}_{1D}, on the right: level sets of Ω^{Facets}_{2D}.

Figure 1 illustrates the function Ω^{Facets}_{1D}, and depicts its subgradient $\partial\Omega^{Facets}_{1D}$. Ω_{2D} is the penalty term for the 2-dimensional case.

4. Properties of the Facets-Regularised Optimal Solution

Here we provide some properties of the model $\mathbf{w}^{\star}_{\lambda\Omega^{Facets}}$ obtained by minimising the regularised risk. We discuss its low magnitude, high integrity, and its ability to correctly represent a learning set, and generalise beyond it. The intuition behind the theoretical properties is provided by the level sets of Ω^{Facets}, depicted on Figure 1 on the right.

Indeed, the *regularised* problem $\min_{\mathbf{w}\in\mathbb{R}^m} \ell(\mathbf{w}) + \lambda\Omega(\mathbf{w})$ and the *constrained* problem $\min_{\mathbf{w}\in\mathcal{B}^{\Omega}_{k}} \ell(\mathbf{w})$ are tightly related, with a latent correspondence between the parameters λ and k. Therefore, the shape of the level sets $\mathcal{B}^{\Omega}_{k}$ tells a lot about the properties of the minimiser. 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 behaviour 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*. Due to the Karush – Kuhn –Tucker (KKT) conditions, the minimisers of the constrained problems are likely to be found at one of the many vertices of the polyhedron, where some of the coordinates are integers – the same reason that explains why the L_1 norm induces sparsity.

4.1. Weights Magnitude

The following result states that the models magnitude can be arbitrarily controlled by a hyper-parameter.

Proposition 5 $\|\mathbf{w}^{\star}_{\lambda\Omega^{Facets}}\| \xrightarrow{\lambda\to+\infty} 0.$

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

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

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.

Proposition 6 *The risk and margin bounds of model* $\mathbf{w}^\star_{\lambda\Omega^{Facets}}$ *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 $\mathcal{F}$:

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

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 $\mathcal{B}_k^{L_1}$, $\mathcal{B}_k^{L_2}$, $\mathcal{B}_k^{Facets}$ satisfy:

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

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

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

4.3. Integrity

Scaling the discretization grid. We remark that, for a classification problem, the model is invariant by scaling, i.e. for all positive γ, the classifiers $\mathbf{x} \mapsto \text{sign}(\mathbf{w}\cdot\mathbf{x})$ and $\mathbf{x} \mapsto \text{sign}(\gamma\mathbf{w}\cdot\mathbf{x})$ are identical, even though they might be treated differently by the loss function ℓ and the penalty Ω. Therefore, we consider two alternative versions of the regularized objective:

$$\mathbf{w}^\star_{\lambda,\gamma} = \arg\min_{\mathbf{w}} \ell(\gamma\mathbf{w}) + \lambda\Omega(\mathbf{w}); \tag{8}$$

$$\widehat{\mathbf{w}}^\star_{\lambda,\gamma} = \arg\min_{\widehat{\mathbf{w}}} \ell(\widehat{\mathbf{w}}) + \lambda\Omega(\frac{1}{\gamma}\widehat{\mathbf{w}}). \tag{9}$$

Because of the KKT conditions, we expect the solution of problem 8 to have a high integrity, i.e. $\mathbf{w}^\star$ should have most of its coordinate in $\mathbb{Z}$, and $\widehat{\mathbf{w}}^\star$ should have most of its coordinates in $\gamma\mathbb{Z}$. Therefore, γ can be interpreted as the unit scale of the discretization grid of the model.

In our soft computing approach, we treat λ, γ as two distinct hyper-parameters, that we set by performing a cross-validated grid search. These two hyperparameters allow to

control separately the size and the shape of the level set of the penalty function; in turn, this should allow us to reach simultaneously models with high accuracy, high integrity and low magnitude. This contrasts to related works:

- in (Golovin et al., 2013), similar parameters are used and decay according to an adaptive, per-coordinate schedule;
- in (Ustun and Rudin, 2016), ℓ is the 0-1 loss and Ω the $\|\cdot\|_0$ semi-norm, both invariant to scaling, and λ is a parameter fixed beforehand to reflect the price of sparsity;
- usual penalties, e.g. L_1, L_2, *elastic net*, are absolutely homogenous, i.e. $\lambda\Omega(\frac{1}{\gamma}\widehat{\mathbf{w}}) = \frac{\lambda}{\gamma}\Omega(\widehat{\mathbf{w}})$. Thus, the parameters λ and γ are redundant.

5. Efficient Minimisation of the Facets-Regularised Risk

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

Regularised risk minimisers are theoretical objects that we cannot compute directly, but rather try to approximate through an optimisation 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 $\mathbf{w}^\star_{\lambda\Omega^{Facets}} \in \mathbb{Z}^m$, 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 optimisation.

We consider the *operator splitting* approach, widely used for non-smooth optimisation and already known to favour sparsity under sparsity-inducing regularisation. 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 et al., 2014) were developed to minimise 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 in (Daubechies et al., 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 optimisation 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 linearised 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 optimised by a *forward 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 regularisation 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 regularisation term is strictly convex and smooth; and iii) to ensure the descent of $\mathbf{w}^t$ towards the minimiser of $\ell+\Omega$. The optimisation 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 7 *For all* $\mu \in [0,+\infty[$, *for all* $\mathbf{w} \in \mathbb{R}^m$, $Prox_{\mu\Omega^{Facets}}(\mathbf{w}) = (sign(w_1)\cdot v_1, \ldots, sign(w_m)\cdot v_m)$, $j \in \{1,\ldots,m\}$:

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

Proof First, as Ω^{Facets} is separable, so is its proximity operator, and we only need to solve a $\mathbb{R} \to \mathbb{R}$ optimisation 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 $x \mapsto (x-\mu)_+$. ■

Figure 2 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 behaviour of $\text{Prox}_{\mu\Omega_{1D}^{L_1}}$ between 0 and $1+\mu$.

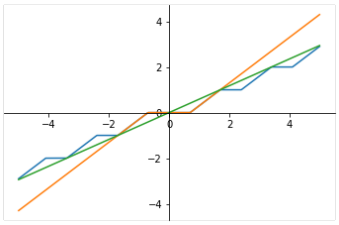

Figure 2: Proximity operators of penalty functions: $\Omega_{1D}^{L_1}$ (in orange), $\Omega_{1D}^{L_2}$ (in green), and Ω_{1D}^{Facets} (in blue).

5.3. Strongly Convex Facets

The Facets penalty is neither strongly, nor strictly convex, as a result of its locally constant

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

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.

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 $w \geq 0$,

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

Hence, $\Omega^{\text{corr}} = \Omega^{L_2} + \Omega^{L_1} - \Omega^{Facets}$ and, for $0 < \epsilon < 1$, the Strongly Convex Facets (SCF) function defined by

$$\Omega^{SCF_\epsilon} := \Omega^{Facets} + \epsilon\Omega^{\text{corr}} = (1-\epsilon)\Omega^{Facets} + \epsilon(\Omega^{L_1} + \Omega^{L_2}) \tag{14}$$

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

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

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

$$\text{with } a = \left(\frac{|w| - \mu}{1+\mu}\right)_+. \tag{15}$$

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 minimiser of the regularised objective. In turn, this property provides resilience to potential correlations between features.

Scaling. To efficiently solve the empirical risk minimisation problem 9, we need to compute the proximity operator of a scaled penalty $\Omega\left(\frac{\cdot}{\gamma}\right)$. Fortunately, scaling interacts smoothly with proximal calculus (see e.g. (Bauschke and Combettes, 2017), proposition 24.8):

$$\text{Prox}_{\mu\,\Omega\left(\frac{\cdot}{\gamma}\right)} = \gamma\ \text{Prox}_{\frac{\mu}{\gamma^2}\Omega}\left(\frac{\cdot}{\gamma}\right). \tag{16}$$

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.

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

Proposition 8 *When applied to the Strongly Convex Facets regulariser, the proximal gradient algorithm enjoys linear convergence, i.e.*

$$\|w^t - w^\star_{\lambda\Omega^{SCF_\epsilon}}| \leq (\lambda\tau\epsilon + 1)^{-t}\|w^0 - w^\star_{\lambda\Omega^{SCF_\epsilon}}\|. \tag{17}$$

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

5.5. Regularisation Path

Strong convexity of the penalty leads to a proper optimisation problem[3]. We can therefore define the *regularisation path* RP as the function mapping hyper-parameters to the (unique) minimiser of the regularised objective:

$$RP_\epsilon : (\lambda, \gamma) \mapsto \mathbf{w}^\star_{\lambda\, scaled_\gamma(\Omega^{SCF_\epsilon})}. \tag{18}$$

The hyperparameters provide smooth control over the selected model.

Proposition 9 *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). ■

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 regularisation.

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.

6. Experiments

We run the experiments on 8 publicly available data sets (all downloadable from the UCI Machine Learning Repository): Mammography, Bankruptcy, Breast Cancer, Haberman, Heart Disease, Mushrooms, Spam, and Adult. All the hyper-parameters, μ, ϵ, and all other hyper-parameters are chosen by cross-validation, ϵ is fixed to 0.01, so that the impact of the L_1 and L_2 penalty terms were minimal.

To fix the hyper-parameters, we apply an extensive grid search over (λ, γ). As we want to focus on the trade-off between λ and γ, we introduce $\alpha := \lambda/\gamma$, that should govern the size of the Facets regulariser and $\beta := \lambda/\gamma^2$ which should control its shape. Table 1 illustrates the results of these experiments. Performance is rather homogeneous over the grid — as long as the estimated model is not null, which happens as soon as regularisation is too strong — while integrity is not, e.g., for the Mammography data we retrieve 236 fully

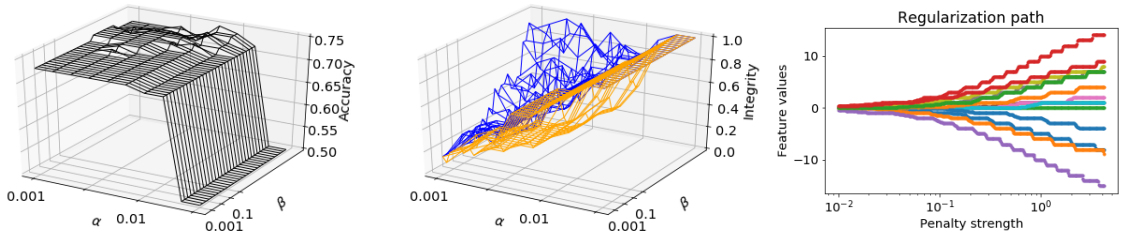

Figure 3: Mammography data set. Variations of the objectives as a function of the hyper-parameters (shape and strength of the Facets regulariser). On the left: accuracy, in the center: integrity, on the right: the regularisation path.

Data	*Chev.*	*Golo.*	*SLIM*	Lasso	Ridge	*Facets*
Adult	0.75	0.77	0.82	0.82	0.82	0.70
Bank	0.98	0.97	0.99	0.99	0.99	**0.99**
Breast	0.94	0.95	0.97	0.97	0.97	**0.97**
Haberman	0.75	0.71	0.71	0.58	0.63	**0.75**
Heart	0.68	0.67	0.83	0.85	0.85	**0.85**
Mammo	0.66	0.60	0.81	0.80	0.81	**0.80**
Mushooms	0.81	0.99	1	1	0.98	0.92
Spam	0.86	0.88	0.93	0.90	0.74	0.89

Table 1: Mean test accuracy from 10-fold CV. Integral models are in italics.

integral models, 23 achieve maximal accuracy (equal to the one achieved by non-integral models). These numbers can be found in Table 2. The obtained accuracy is in line with the state-of-the-art techniques, SLIM (Ustun and Rudin, 2016) for integral models, L_1, or L_2 regularisation, and much better than our implementation of the Chevaleyre's and Golovin's approaches. We report the results obtained on the seven other public datasets reported in (Ustun and Rudin, 2016) (Table 6 pp 26-27) in Table 1, and we highlight the ones where *Facets* obtains results within the error margin of the best accuracy. Table 1 seems to indicate that, in the case where *Facets* seems to fail to deliver an accurate model

Data	grid points	null	integral	optimal
Adult	2400	600	122	5
Bank	2400	300	522	14
Breast	2400	400	458	7
Haberman	2400	800	606	9
Heart	2400	500	101	7
Mammo	2400	500	236	23
Mushrooms	2400	400	34	1
Spam	2400	600	45	1

Table 2: Statistics for *Facets* grid search.

(i.e. Adult, Mushroom and to some extent Spambase), the grid search did not yield many integral or optimal models.

Figure 3 on the left, illustrates the accuracy on Mammography data set. Figure 3 in the center shows integrity (the mean values of 10-fold CV) obtained on Mammography, and on the same figure on the right we plot the values of the coefficients as a function of the parameters of the Mammography data set, for a 24×100 grid. We have noticed that the Facets running time is comparable to the one of the Lasso (it depends on the optimisation method used but we focused on the proximal gradient descent in our paper). The storage cost is more similar to the Elastic Net.

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 et al., 2009), through the upper bound on the Rademacher complexity of the hypothesis class. In the same vein, favouring (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 et al., 2019), and medicine (Ustun and Rudin, 2016).

- Sparsity and interpretability. Favouring integrity can be seen as an instance of structured risk minimisation (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 models with integer weights via soft constraints. We introduced a new penalty term called Facets. Our main theoretical results provide some theoretical foundations for our approach.

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

The numerical experiments – and a detailed study of variations of the accuracy and integrity, as well as the regularisation path, in a real-world medical application – illustrate 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.

Acknowledgments

This work was supported by the French National Research Agency (ANR JCJC *DiagnoLearn*).

References

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

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

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

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

Y. Chevaleyre, F. Koriche, and J.-D. Zucker. Rounding methods for discrete linear classification. In *ICML*, 2013.

I. Daubechies, M. Defrise, and C. De Mol. 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, 2004.

D. Golovin, D. Sculley, H. B. McMahan, and M. Young. Large-scale learning with less RAM via randomization. In *ICML*, 2013.

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pages 793–800, 2009.

H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. Mar Hrafnkelsson, T. Boulos, and J. Kubica. 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*, pages 1222–1230, 2013. doi: 10.1145/2487575.2488200.

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

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

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

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

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

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

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

B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016. doi: 10.1007/s10994-015-5528-6.

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

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