SCIENCES SORBONNE UNIVERSITÉ

En vue de l'obtention de l'Habilitation à Diriger des Recherches (HDR)
de Sorbonne Université

Mémoire présenté et soutenu le mercredi 08 mars 2023

Contributions to convex regularization and statistical learning with or without missing data

Claire Boyer

Composition du jury

Francis Bach	Directeur de recherche, INRIA Paris	Rapporteur
Emmanuel Candès	Professeur des Universités, Stanford University	Rapporteur
Albert Cohen	Professeur des Universités, Sorbonne Université	Examinateur
Gersende Fort	Directrice de recherche, CNRS Toulouse	Examinatrice
Sara van de Geer	Professeure des Universités, ETH Zurich	Rapportrice
Rémi Gribonval	Directeur de recherche, INRIA Lyon	Examinateur

Ecole doctorale : Ecole doctorale des Sciences Mathématiques de Paris Centre (ED386)
Unité de recherche : Laboratoire de Probabilités, Statistique et Modélisation (LPSM, UMR 8001)

Abstract

This dissertation covers the work conducted by the author mostly as "maîtresse de conférences" at the Laboratoire de Probabilités, Statistique & Modélisation (LPSM) of Sorbonne Université since 2016. During this period, the author strengthened her contributions to the compressed sensing community and investigated new fields of research that may be summarized under the following labels "powerful methods in supervised learning" and "missing data in machine learning". This report is not meant to present an exhaustive description of the results developed by the author, but rather a synthetic view of her main contributions. The interested reader may consult the cited articles for further details and the precise mathematical formalism of the topics presented there.

Remerciements

Si le geste est rituel, il n'en reste pas moins sincère. Je tiens à remercier chaleureusement mes rapporteurs et rapportrice, Francis Bach, Emmanuel Candès et Sara van de Geer, d'avoir accepté de relire ce manuscrit, alors qu'ils étaient déjà bien occupés à façonner la communauté internationale de la statistique et de l'apprentissage.

Francis Bach, sommité (étoile noire ?) de l'apprentissage, est une légende vivante au sens littéral du terme, du latin, 'legenda', ce qui doit être lu, ce qui subséquemment doit être retenu, qu'il s'agisse de sa science qui nous exalte avec une constance immarcescible, ou de sa camaraderie et la dérision qu'il insuffle aux moments conviviaux que peut offrir la recherche.

J'ai eu aussi la chance de croiser en conférences à plusieurs reprises Emmanuel Candès, dont je connaissais la plupart des papiers par cœur, et qui, tel un éclaireur, ouvre régulièrement des voies scientifiques en suscitant toujours l'engouement de la communauté et le mien. Merci aussi pour la vision avertie et plus globale dont il est capable, et son savoir-voir. And then I crossed paths with Sara van de Geer, warrior goddess of statistics, who commands respect by attacking problems with an ice axe, north face. She is one of those decisive encounters that give stars in the academic night which, if you do not always have the courage to identify with them, drive and enchant you. Thank you all 3 for the constant inspiration.

Albert Cohen est un pape de l'analyse harmonique, de la théorie de l'approximation et des équations aux dérivées partielles, qui m'avait déjà fait l'honneur de présider mon jury de thèse jadis. Je suis ravie de le compter encore aujourd'hui comme membre interpolateur des comités qui ponctuent ma vie académique. Je remercie également Gersende Fort, spécialiste et référence des algorithmes stochastiques, qui au-delà de l'excellence et de la rigueur de la science qu'elle nous transmet, est aussi de conseils avisés et bienveillants. Je suis enfin reconnaissante à Rémi Gribonval de prendre part à ce jury, dont la munificence n'a d'égal que son extrême qualité scientifique et qui donne l'exemple réjouissant de dynamique depuis le traitement du signal jusqu'à l'apprentissage.

Le lecteur le plus novice l'aura compris : le line-up de ce jury est à en faire pâlir et défaillir toute personne sur le point de défendre son habilitation. Avant que je ne rebrousse chemin, je m'empresse donc de remercier, outre mon jury, mes collègues, les trajectoires personnelles n'étant que résonances du collectif.

Merci au feu Laboratoire de Statistique Théorique et Appliquée -devenu Laboratoire de Probabilités, Statistique et Modélisation depuis- de m'avoir fait confiance en me recrutant, alors que ma coloration statistique était encore à étoffer, et malgré quelques aspérités formelles dont j'ai eu fait preuve (ante, in, post comité de sélection), bref alors que j'étais peu recommandable. Toute ma gratitude à Gérard Biau, qui a allègrement remplacé Ulysse à Ithaque, qui a revitalisé, à son insu, la diaspora audoise dans le 5ème arrondissement, qui revisite Glivenko-Cantelli à l'envi, et qui ne boude pas son plaisir à régulièrement brocher sur le tout. Merci à Arnaud Guyader, "deus"

auto-proclamé, pour ses encouragements intransigeants, son humeur railleuse, son verbe corrosif, bref avec qui on est rarement sorti de l'auberge.

Merci à Maxime Sangnier, avec qui nous avons partagé un beau syndrôme de l'imposteur (de l'ingénieur ?) immédiatement et irrémédiablement dès le 1er septembre 2016, qui est d'une rigueur rassurante, et à vrai dire l'Atlas du laboratoire. Merci à Maud Thomas pour son entrain, à Charlotte Dion qui, si elle n'est pas déjà cocaïnomane, devrait le considérer au vu du nombre de tâches qu'elle assure, à Anna Ben-Hamou pour son charme instruit et sa passion pour le bolidage, à Eddie Aamari pour "ses problèmes de reach", sa pugnacité dans la lecture des travaux de Misha Belkin, et le deal de triglycérides à base de griottes et pistaches, à Anna Bonnet pour sa mémoire d'outre-tombe et sa cordialité inaltérable, à Antoine Godichon-Baggioni pour sa complétion irrépressible de cases blanches, à Pierre Tarrago, pour sa présence aussi singulière qu'égayante, et au renouveau du laboratoire, Stéphane Robin et Sylvain Le Corff. Merci à Olivier Lopez, pour ses goûts douteux -tout du moins autrefois- en cinéma et en cocktails. Merci à Tabea Rebafka d'avoir partagé la direction du master de statistique, réjouissance que connaît à présent Etienne Roquain. J'ai une pensée également pour Louise Lamart, Corinne van Vlierberghe, et Valérie Juvé qui nous ont épaulés ou nous épaule encore avec abnégation, et pour Hugues Moretto qui doit travailler au milieu de cette engeance dont je fais partie. Les doctorantes et doctorants de l'équipe (et au-delà) animent joyeusement le quotidien du laboratoire, merci à vous de subir les vieux schnocks que nous sommes. Merci à tous les autres membres du LPSM que je n'aurais pas cités nommément, et qui restent toujours de bonne compagnie du dilicule au crépuscule.

Je suis également reconnaissante au Département de Mathématiques et Applications de l'Ecole Normale Supérieure de la rue d'Ulm, de m'avoir fait l'honneur d'en faire partie entre 2017 et 2020. J'ai eu la chance d'y côtoyer des étudiantes et étudiants éblouissants, et de partager le bureau de Stéphane Boucheron (puis plus rapidement celui de Stéphane Gaïffas, un Stéphane pouvant en cacher un autre), merci Stéphane -vieux vélibataire qui m'a depuis convertie- pour ta culture scientifique sans borne et ta générosité.

L'équipe MOKAPLAN à INRIA Paris, bastion du punk, s'il en est, m'a hébergée par intermittence. Merci particulièrement à Jean-David Benamou, Guillaume Carlier, Thomas Gallouët, Paul Pegon, et Irène Waldspurger pour les conversations toujours éclairantes.

Je dois beaucoup à mes collaborateurs et collaboratrice prodigieux, remèdes à toute impéritie. Qu'ils soient célébrés ici (ndlr, dans la suite du manuscrit aussi). Par ordre chronologique, merci à Yohann de Castro (pour sa clairvoyance, ses lumières dignes de la plus grande pyrotechnie, et ses espièglerie qui me feront toujours rire), à Joseph Salmon (mais non merci aux surnoms qu'il me donne), à Ben Adcock and Simon Brugiapaglia (for this delightful Pacific interlude), à Antonin Chambolle (dont la finesse d'esprit vous satellise, et s'il n'y a pas de bonne ou mauvaise situation, satellite d'Antonin Chambolle c'est déjà pas mal), à Vincent Duval (pour son goût trop prononcé pour la topologie, les diagrammes commutatifs et les chaînes), à Frédéric de Gournay (pour son amour des cimetières), à Pierre Weiss (pour ses "Weississitudes", son souci d'exigence, et tout ce qu'il m'a transmis), à Jonas Kahn (l'oracle), à Maximilian März (pour une collaboration deutsche qualität), à Julie Josse (pour avoir co-encadré avec moi une première thèse, et qui reste de gaieté inébranlable en recherche), à Aymeric Dieuleveut (pour sa finesse d'esprit, son amitié, et la profondeur des choses qu'il révèle en les regardant), à Boris Muzellec (auteur des plus beaux "barplots" jamais créés), à Marco Cuturi (danseur invétéré qui nous transporte optimalement) et à Erwan Scornet (botanomancien brillant).

Mon chemin a croisé celui d'étudiantes et étudiants remarquables : Aude Sportisse (première doctorante épatante, à qui j'espère avoir transmis autre chose que mon optimisme légendaire),

Ludovic Arnould (sur le point de la délivrance), Alexis Ayme (qui ne devrait pas suivre l'exemple de la gestion des vacances dont font preuve 2 de ses 3 encadrants), et Nathan Doumèche (pour qui Dudley ou Evans n'ont plus de secret ; de premiers pas très prometteurs donc). J'ai eu le plaisir de travailler avec Kimia Nadjahi, qui dompte aussi bien les processus gaussiens que les pogos. Sans oublier les stagiaires de master, Patrick Lutz, Théo Uscidda, Linus Bleistein, Paul Liautaud, Imke Mayer (à qui je souhaite de se réaliser pendant leur doctorat et après - si la dernière ne l'a pas déjà fait).

Ces remerciements s'éternisent[1], et je faillis déjà à rendre hommage à toutes les personnes que je croise, écoute en conférences, qui me nourrissent scientifiquement, et avec qui je ris. À celles et ceux qui se reconnaîtraient, ces moments comptent, merci.

Et puis, il y a tous les autres, qui restent étrangers au monde académique, et qui n'en sont pas moins essentiels. Merci aux frisées (pour l'envoi de fusées, leur allant, et leurs sourires qui fendent leurs visages en 2 et vos cœurs en 1000), au "groupe très sympathique" (qui fait tout autant mon bonheur que celui des propriétaires de maison ouvertes à la location), à la joyeuse bande du Sud (reine des petites finales, que je chéris autant que son café), aux Sarla(dirlada)dais (amateurs contrariés de Mona Chollet, de truffes ou de pigeons d'argile), à Sophie et Fanny (raisons pour lesquelles j'ai vu briller 3 soleils dans le ciel). Aux fantômes, des vivants et des morts, dont les souvenirs ont une aptitude astringente à la ténacité. À ma grand-mère qui récite du Hugo sur commande dès potron-minet. À mes parents pour leur patience, leur soutien et leur SAV indéfectible. À ma sœur (et la beauté) pour sa complicité et son hospitalité à l'égard d'enseignantes-chercheuses itinérantes. À Benjo, pour tout, pour rien, et avec qui je danse dans le noir.

Ce manuscrit n'a pas pour ambition d'embrasser une grande postérité (je reste à vrai dire aussi pauvre, bête et finie que je ne l'étais avant ce travail (amen)), mais j'espère que vous y lirez tout du moins le témoignage d'un bon souvenir.

[1] La bêtise consiste à vouloir conclure, avait écrit Flaubert.[2]

[2] Mesdames, l'asséner sur un ton docte vous permettra également d'éconduire avec panache tout prétendant lourd et empressé.

Contents

Introduction

This manuscript describes the research contributions and developments that I have carried out since obtaining my Ph.D.

Scientific background

After engineering studies, I conducted my PhD work (2012-2015) at the Toulouse Institute of Mathematics (IMT) under the supervision of Jérémie Bigot and Pierre Weiss. Motivated by applications in MRI imaging, the aim was to study a theory of compressed acquisition, area of applied mathematics opened by Emmanuel Candès and David Donoho, that would better capture the mechanisms at work in practice (block constrained acquisition in particular).

To celebrate my PhD, I spent two weeks in Saint-Flour, a country of cheese which, if it is a July vacationer, can only be oozing, but also and above all, a country of high-level mathematics. I followed with assiduity and pleasure the courses, and in particular the one given by Sara van de Geer that year. Her lecture actually inspired a whole article reported later, written with Yohann de Castro (Université d'Orsay) and Joseph Salmon (Télécom Paris), who I met during my evening revisions in Saint-Flour, and with whom I still collaborate for one, and probably again one day for the other.

For personal reasons, I had to give up the post-doctorate that was planned under the direction of Ben Adcock (Simon Fraser University), thus preventing me from singing Véronique Sanson at the top of my lungs on a Pacific port. So I took in extremis a full time position as an ATER position (teaching and research assistant position including about 180h of teaching) at INSA Toulouse - I would like to thank the Toulouse team who helped me in the emergency to find a solution and also for their benevolence which facilitated among other things the absorption of the teaching load. During the ATER(moiement), I therefore worked with the aforementioned Yohann de Castro and Joseph Salmon on the themes of super-resolution (Boyer et al., 2017), conjugating an approach studied by Sara van de Geer and a theme also dear to Emmanuel Candès.

Pushed in large part by the director of the Toulouse mathematics doctoral school, Jean-Michel Roquejoffre, I participated in my first MCF application campaign in 2016. I had the pleasure of joining the LSTA, which later became the LPSM, within the Pierre and Marie Curie University, which later became Sorbonne University. I found there a welcoming and stimulating team. I collaborated with Maxime Sangnier (Fouillen et al., 2022) and Antoine Godichon-Baggioni (Boyer and Godichon-Baggioni, 2020).

I was also able to visit Ben Adcock (Simon Fraser University) at last during a two-month research visit to Canada in 2017. We worked in collaboration with Simone Brugiapaglia (SFU), to obtain "optimal" reconstruction bounds for structured compressed acquisition (Adcock et al., 2021).

From 2017 to 2020, I was an associate member of the mathematics department of the Ecole Normale Supérieure, rue d'Ulm. I had the opportunity to interact with exceptionally talented students, among whom one will soon start his thesis under my co-supervision.

I also took advantage of the richness of the Parisian academic world and of my roots in Toulouse to pursue research projects related to inverse problems involving convex regularizations. I thus collaborated (again) with Yohann de Castro (Orsay), Vincent Duval (Inria Paris), Antonin Chambolle (Ecole Polytechnique), Frédéric de Gournay (University of Toulouse III), Pierre Weiss (CNRS, Toulouse), Jonas Kahn (CNRS, Toulouse) and Maximilian Marz (TU Berlin) concerning the papers Boyer et al. (2019b); März et al. (2020). On this occasion, we realized that Francis Bach had already asked himself similar questions, almost 10 years earlier, and had answered them exhaustively in his book Bach et al. (2013).

Through my teaching and my responsibilities (direction of the Master 2 of Statistics of SU), I have progressively oriented my research towards statistics and machine learning.

From adventure to adventure, I met Julie Josse (Ecole Polytechnique) with whom I co-supervised Aude Sportisse's thesis from 2018 to 2021. We worked on the statistical and methodological aspects of the missing data problem (Sportisse et al., 2020a,b,c; Descloux et al., 2022; Sportisse et al., 2021). Since 2020, I co-supervise with Erwan Scornet (Ecole Polytechnique) the thesis of Ludovic Arnould studying the connections between neural networks and tree ensemble methods (Arnould et al., 2021; Lutz et al., 2022; Arnould et al., 2022). Since 2021, Aymeric Dieuleveut (Ecole Polytechnique)[3], Erwan Scornet and myself are supervising Alexis Ayme's thesis in machine learning with missing data (Ayme et al., 2022). I also have had the opportunity to work with a postdoc, Kimia Nadjahi, on related temporal aspects since December 2021.

These few lines already announce that my research is a real collective experience and I really enjoy it this way. I sincerely thank my collaborators (I obviously include the students) for making research a joyful journey.

Mathematics of sparsity

In this prelude, I allow myself to set out the theoretical framework of the theory of compressed sensing in order to situate my thesis work in a concise manner. The objective will be twofold: to give context to the work presented in Chapter 2, but also to help the reader measure the path taken since then.

I therefore entered the research through the door of compressed sensing (CS), or how, from a limited number of measurements, one can be confident on the reconstruction of high or even infinite-dimensional objects.

The latter can be a vector in finite dimension (this is in general what compressed sensing refers to), or a matrix with missing entries (this is therefore called matrix completion), or a Radon measure (usually termed off-the-grid compressed sensing). To compensate for the lack of information on these objects, a strong structural a priori must be considered to disambiguate the reconstruction: the sparsity is invoked in the vector case, a low-rank prior in the matrix case, and an atomic structure (as a finite sum of Dirac masses) as for a Radon measure.

The CS problem, often encountered in practice, covers a wide range of applications such as medical, radar imaging or even finance. This is why research in this field has been very popular in the 2000s and 2010s, posing beautiful problems at the frontier of optimization, statistics, probabilities,

[3]After 4 mentions to Polytechnique, I guess that we can legitimately call it the X-factor.

and signal processing. The interested reader may consult Chen et al. (2001); Tibshirani (1996); Fuchs (2005); Candès and Tao (2006) for seminal papers and Bühlmann and Van De Geer (2011); Chafaï et al. (2012); Foucart and Rauhut (2013); Giraud (2021); Hastie et al. (2009); Vershynin (2018) for reviews and book chapters.

Canonical results of compressed sensing (CS)

CS formalism The standard CS problem can be stated as follows. Let $x_\star \in \mathbb{R}^d$ denote an s-sparse vector. The signal $x_\star$ is unknown but (noiseless) linear measurements y are performed such that

$$y = Ax_\star \in \mathbb{C}^m \qquad A = \begin{pmatrix} a_1^* \\ \vdots \\ a_m^* \end{pmatrix} \in \mathbb{C}^{m \times d} \tag{1}$$

for some sensing matrix A with $m \ll n$, denoting its rows $(a_i)_{1 \leq i \leq m}$. The goal is to reconstruct the signal $x_\star$ from y, by using the sparse prior in $x_\star$. Indeed, to promote the sparsity of the target, it is natural to consider a reconstruction via an ℓ^1-minimization problem, called the *basis pursuit*,

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \qquad \text{such that} \quad y = Ax. \tag{2}$$

The principle of *basis pursuit* is to choose, among the infinity of solutions of the underdetermined linear system (1), the one that minimizes the ℓ^1-norm, which has the good grace to be a convex surrogate of the pseudo ℓ^0-norm, counting the nonzero elements in the vector of interest. Hence, the basis pursuit is a convex problem that can be solved with efficient algorithms (e.g. simplex or Douglas-Rachford algorithms).

Note that we consider the case without noise, for simplicity, but the theoretical analysis extends to the case of noisy observations.

Key elements in CS Typical CS results ensure that $x_\star$ can be reconstructed solving (1), usually requiring measurements to be "incoherent", which means that although small in number compared to the dimension of the object to be reconstructed, they have the ability to capture the total energy of the signal with high probability. Therefore they are usually assumed to be random in theoretical analyses as it leverages concentration tools and allows to derive recovery guarantees with high probability. Historically, two types of measurements have been mainly considered:

(i) randomized Fourier measurements: The Fourier transform is often used in practice to model the physics of acquisition, e.g., as in MRI, although the sampling patterns are usually deterministic for practical purposes. Its randomization actually helps the mathematician to benefit from theoretical concentration results.

(ii) Gaussian measurements: each measurement vector is made of independent Gaussian entries. These types of sensors are more abstracted than acquisition models in practice, but remain suitable for theoretical purposes. They allow, in particular, characterizing phase transitions about the sampling rate of a convex program, i.e., the required number of measurements to ensure successful recovery via the convex program at stake. We will return to these notions when developing our work on phase transition when the signal is represented in redundant dictionaries; see Section 1.2.

An archetypal result in CS Typical results in compressed sensing focus on providing reconstruction guarantees via ℓ^1-minimization for *any* s-sparse vector, under conditions often expressed in terms of the number of measurements to observe.

Theorem 1. *Let $s \in \{1, \dots, d\}$ be a degree of sparsity, and let $A \in \mathbb{R}^{m \times d}$ be a Gaussian sensing matrix. Then any s-sparse vector x can be reconstructed from measurements $y = Ax$ via the basis pursuit (2), with probability $1 - \delta$, if*

$$m \gtrsim s \log\left(\frac{ed}{s}\right) + \log\left(\frac{2}{\delta}\right).$$

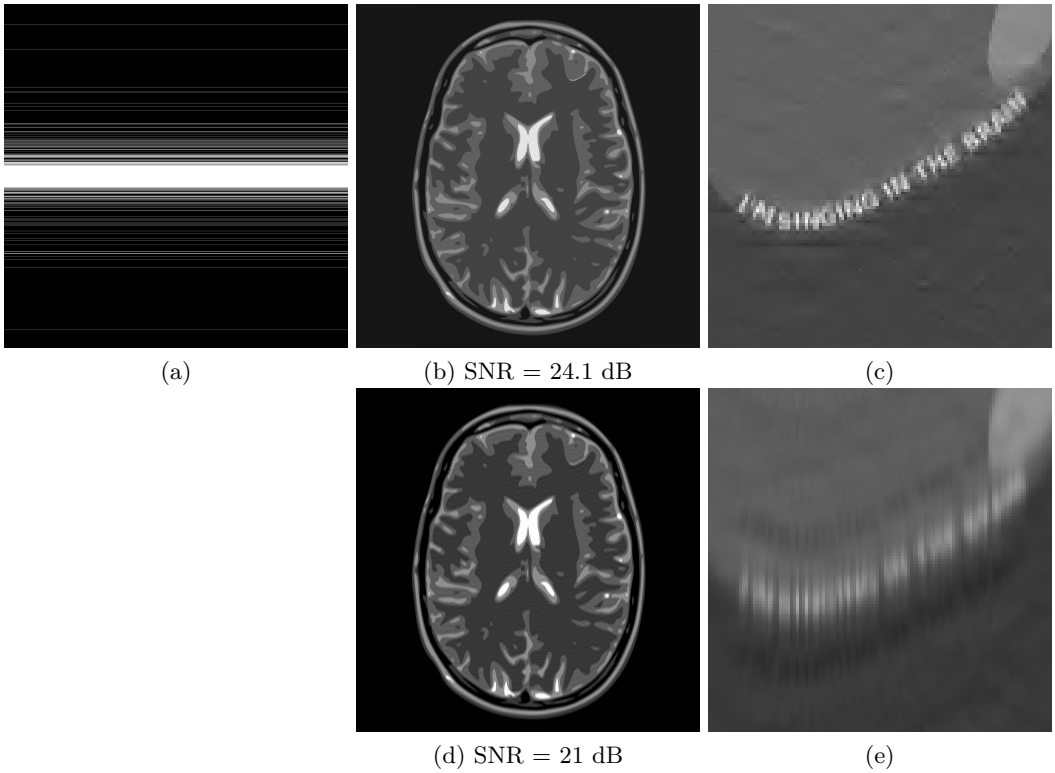
This result exhibits that exactly recovering a high-dimensional vector is possible using a nonlinear reconstruction method as soon as the number of (incoherent) measurements is of the order of the degree of sparsity of the given vector, the ambient dimension only intervening in a logarithmic factor; hence the name of compressed sensing. A noticeable fact about this result is its uniformity: it controls the probability of reconstruction of *any* s-sparse vector given a random Gaussian matrix. There exist an extension of this theorem in terms of robustness (to some noise in the observations) and stability (to the assumed sparse model). Under similar conditions, the reconstruction error typically reads

$$\|\hat{x} - x\|_2 \lesssim \underbrace{\eta}_{\text{robustness}} + \underbrace{\frac{\min_{z \ s-\text{sparse}} \|x - z\|_1}{\sqrt{s}}}_{\text{stability}}$$

where $\hat{x}$ is the solution of the BP given $y = Ax + \epsilon$ with a bounded noise $\|\epsilon\|_2 \leq \eta$.

PhD. contributions

The work conducted during my doctoral studies (Bigot et al., 2016; Boyer et al., 2019a) provides better theoretical guarantees of reconstructions when data acquisition is strongly constrained, which is often the case in applications. For instance, in MRI, Fourier measurements are typically sensed, not arbitrarily in the phase domain, but grouped by blocks, see for instance Figure 1 (a). And still, nonlinear reconstruction methods give good performances in such a case (see Figures 1 (b)(c) compared to (d)(e)). The most popular results at that time in the literature actually failed to explain this phenomenon, as they usually provided uniform guarantees, by quantifying the probability of exact recovery of *any* s-sparse vector. In particular, we have shown that it is impossible to uniformly reconstruct the whole class of s-sparse signals (which forces us to resort to other proof techniques based on dual certificates as in Candès and Plan (2011) instead of using the restricted isometry property (RIP)). Our results are therefore non-uniform, in the sense that they allow the reconstruction of a given signal, and that the reconstruction success depends on the structure of the signal, and more precisely on its support. We exemplify the theoretical contributions on the case of block-constrained acquisition in the Fourier basis, when the object to reconstruct can be sparsely represented in a (Haar) wavelet basis, matching to some extent the setting of MRI. From this thread of research, one can try to minimize the theoretical required number of measurements with respect to the sampling probability. This gives a target distribution from which samples should be drawn. However, in practice, using random sampling schemes is far from being efficient: we should prefer deterministic surrogates. Based on this observation, we have proposed new deterministic sampling techniques (Boyer et al., 2016), (i) mimicking a target sampling distribution advocated by the CS theory, (ii) complying with admissible sampling patterns (dictated by the sensing device), and (iii)



(a) (b) SNR = 24.1 dB (c)

(d) SNR = 21 dB (e)

Figure 1: An example of reconstruction of a 2048×2048 MR image from blocks of measurements. (a) Sampling pattern horizontal lines (13% of measurements). (b) Corresponding reconstruction by ℓ_1-minimization. (c) A zoom on a part of the reconstructed image. (d) Image obtained by using the pseudo-inverse transform. (e) A zoom of a part of this image.

covering the space efficiently. These strategies have been implemented on real MR scanners, during the Ph.D. thesis of Carole Lazarus (Lazarus et al., 2019a,b) at Neurospin (CEA Saclay), thanks to her pugnacity and that of Pierre Weiss (CNRS, Toulouse).

Overview of contributions

I present here a summary of my work conducted after my Ph.D. studies: they are organized in three thematic axes, instead of any chronological order.

Contributions in convex regularization In Chapter 1, I present results obtained in relation to convex regularization techniques. In Adcock et al. (2021), we propose improved sampling complexity bounds for stable and robust sparse recovery in compressed sensing. The unified analysis based on ℓ^1-minimization encompasses the case where (i) the measurements are block-structured samples to reflect the structured acquisition that is often encountered in applications; (ii) the signal

has an arbitrarily structured sparsity, so that the results directly depend on its support. Within this framework and under a random sign assumption, the number of measurements needed by ℓ^1-minimization can be shown to be of the same order than the one required by an oracle least-squares estimator. Moreover, these bounds can be minimized by adapting the variable density sampling to a given prior on the signal support and to the coherence of the measurements.

In März et al. (2020), we investigate the problem of signal recovery from undersampled noisy sub-Gaussian measurements under the assumption of a synthesis-based sparsity model. Solving the ℓ^1-synthesis basis pursuit allows us to simultaneously estimate a coefficient representation as well as the sought-for signal. However, due to linear dependencies within redundant dictionary atoms, it might be impossible to identify a specific representation vector, although the actual signal is still successfully recovered. We study both estimation problems from a nonuniform, signal-dependent perspective. By utilizing results from linear inverse problems and convex geometry, we identify the sampling rate describing the phase transition of both formulations: it involves the Gaussian width of a linearly transformed polyhedral descent cone. We propose a "tight" estimate of this quantity which can be evaluated on a computer.

Then, we leave the finite-dimensional setting to study sparse spike deconvolution over the space of complex-valued measures when the input measure is a finite sum of Dirac masses. In Boyer et al. (2017), we introduce a modified version of the Beurling Lasso (BLasso), a semidefinite program that we refer to as the Concomitant Beurling Lasso (CBLasso). This procedure estimates the target measure and the unknown noise level simultaneously. Contrary to previous estimators in the literature, the obtained theoretical results, including spikes localization and prediction bounds, hold for a tuning parameter that depends only on the sample size, so that it can be used for unknown noise-level problems.

Finally, we step back to consider very general optimization programs that include *any* convex regularizer (while the fidelity-to-data term does not necessarily need to be convex). In Boyer et al. (2019b), we characterize the structural properties of minimizers that can be expressed as convex combinations of a small number of atoms. These atoms are identified with the extreme points and elements of the extreme rays of the regularizer level sets. This very general analysis embraces many regularization problems that have been extensively studied in the past years. As a by-product, we characterize the minimizers of the total gradient variation, which was still an unresolved problem at that time.

Contributions in general supervised learning In Chapter 2, I give a quick overview of my work related to general learning methods. In supervised learning, predictors are usually learned by minimizing some (empirical) risk. To do so, we propose two general learning algorithms that, respectively, belong to the classes of boosting methods and stochastic optimization.

First, gradient boosting is a prediction method that iteratively combines weak learners to produce a complex and accurate model. From an optimization point of view, the learning procedure of gradient boosting mimics a gradient descent on a functional variable to minimize a risk. In Fouillen et al. (2022), we introduce a proximal boosting algorithm and its residual version, building upon the proximal point algorithm. Theoretical convergence rates are obtained; in particular, they do not require the differentiability of the objective function in the case of the residual algorithm. The relevancy of Nesterov's acceleration in such a setting is also questioned, as it introduces instabilities in the algorithm behaviour.

In Boyer and Godichon-Baggioni (2020), we define general (weighted averaged) stochastic Newton algorithms to perform a learning task when the data comes in a streaming fashion. The

implementation does not require the inversion of a Hessian estimate at each iteration, but leverages from the possibility to directly update the inverse of the Hessian matrix at each iteration in $O(d^2)$ operations, with d the ambient dimension. Asymptotic convergence results are derived without requiring the strong convexity of the considered risk, and they also shed some light on the fact that the choice of the averaging technique is not always innocuous.

Then, we explore tree-ensemble methods and neural networks, 2 classes of powerful machine learning predictors, and how one can benefit from the other. In Arnould et al. (2021), we provide a numerical and theoretical analyses of the deep forest algorithm, which is a meta neural network model in which each neuron consists of a (non-differentiable) random forest. In particular we study a toy model of a 2-layer tree network shown to enhance the performance of classical decision trees in a specific theoretical framework. In Lutz et al. (2022), we propose a new sparse initialization technique for (potentially deep) multilayer perceptrons (MLP): we first train a tree-based procedure to detect feature interactions and use the resulting information to initialize the network, which is subsequently trained via standard stochastic gradient strategies. This wise MLP initialization actually raises the performances of the resulting NN methods to that of gradient boosting on tabular data.

Finally, in Arnould et al. (2022), we study the performances of the popular random forest algorithm in interpolation regimes. Even if it is commonly admitted that very complex models, interpolating training data, will be generally poor at predicting unseen examples, this statistical wisdom has been recently challenged. Benign overfitting regimes have indeed been identified, especially in the case of parametric models: generalization capabilities may be preserved despite the model high complexity. While it is widely known that fully-grown decision trees interpolate and, in turn, have bad predictive performances, we show that (median) random forests can be consistent despite of interpolation.

Contributions in the problem of missing data In Chapter 3, I address the issue of missing data in machine learning, reconnecting with medical applications as well. Missing values are becoming more and more present as the size of datasets always increases, hindering standard statistical analyses.

A first attractive idea can be to fill in the missing entries in order to get a completed data set that can then be processed by any learning algorithm. In Sportisse et al. (2020c,b), we propose imputation techniques using a low-rank prior and an EM strategy, on the one hand, and a graphical approach, on the other. Both works are intended to deal with a particular complex type of missing values said to be missing not at random (MNAR). In Muzellec et al. (2020), we take advantage of optimal transport distances to define a loss function that is used as an imputation criterion. This approach, versatile enough to allow for building non-parametric or parametric imputers, exploits the idea that two batches extracted randomly from the same dataset should share the same distribution. OT-based methods are shown to match or out-perform state-of-the-art imputation methods, even for high percentages of missing values.

Then we turn to the problem of model estimation in linear regression for which we prefer to adapt the estimation method to missing values instead of resorting to a pre-processing imputation step. In Sportisse et al. (2020a), we study a debiased averaged stochastic gradient algorithm that handles missing features to perform linear regression in an online context. Under a mechanism of "missing completely at random (MCAR) data", we show that this algorithm achieves convergence rate of $O(1/n)$ at iteration n, the same as without missing values.

Note that being able to perform model estimation despite missing data does not help for predic-

tive purposes: the estimated model parameters cannot be directly used for prediction on the test data that may contain missing entries.

Therefore, in Ayme et al. (2022), we focus on linear predictors that handle missing values. Under some mild assumptions on the data distribution, the Bayes rule associated with an underlying linear model can be decomposed as a sum of linear predictors corresponding to each missing pattern. We thus propose a rigorous setting to analyze a least-squares type estimator adapted to the variety of missing patterns, and we establish a bound on the excess risk which increases exponentially in the dimension. Then, we leverage the missing data distribution to propose a new algorithm, and derive associated adaptive risk bounds that turn out to be almost minimax optimal.

Outline

The organization of the following chapters mirrors the organization of the axes previously mentioned. Chapter 1 is therefore dedicated to inverse problems and convex regularization. Chapter 2 summarizes my contributions in general supervised learning. Chapter 3 covers different ways of handling missing values in the practice of data science.

I have chosen to indicate my research directions by the symbol ☕ throughout the document. It materializes the quantity of holy beverage[4] necessary (and probably not sufficient) to get there.

[4]Note that cognitive bias makes us believe this is a coffee cup, but I would like to pay tribute to tea cups, herbal tea cups, or even quaint chicory cups that are the great forgotten of mathematical and engineering achievements.

Chapter 1

Contributions in convex regularization

Contents

1.1 Oracle-type bounds for structured CS

During a research visit to Vancouver in 2017 funded by a PIMS-CNRS "Distinguished Visitor" grant, I initiated a collaboration with Ben Adcock (Simon Fraser University) and Simone Brugiapaglia (Concordia University). The purpose of this section is to present the resulting work (Adcock et al., 2021) of this collaboration.

Context In the seminal paper of Candès and Romberg (2007), under a random sign assumption on the s-sparse signal to reconstruct $x \in \mathbb{C}^d$, the authors proposed to draw uniformly at random rows from an isometry $\underline{A} = (\underline{a}_k)_{1 \leq k \leq d}$, leading to stable reconstruction with probability at least $1 - \delta$ with the following required number of measurements:

$$m \gtrsim s \cdot d \max_k \|\underline{a}_k\|_\infty^2 \cdot \ln(d/\delta)\,. \tag{1.1}$$

This result can be of interest when considering totally incoherent transforms such as the Fourier matrix for which $d \max_k \|\underline{a}_k\|_\infty^2 = O(1)$. However, this is not relevant anymore in the case of coherent transforms, such as the Fourier-Haar transform used to model MRI acquisition (meaning that sensing is performed in the Fourier domain, and the signal is represented in the Haar wavelet domain), where $d \max_k \|\underline{a}_k\|_\infty^2 = O(d)$.

In Adcock et al. (2021), our results include the previous ones, but are also extended to: (i) the case of variable density sampling; (ii) stability robustness results when measurements are corrupted

with bounded noise; (iii) structured measurements using blocks of measurements; (iv) optimization of the sampling density with respect to prior information on the signal support, such as structured sparsity.

A general sampling strategy Given some distributions $(F_\ell)_{1\leq\ell\leq m}$ respectively on sets of $p_\ell \times d$ matrices, with $p_\ell \geq 1$ for $\ell = 1, \dots, m$, the sampling strategy consists in drawing m independent matrices $B_1, \dots, B_m$ where $B_\ell \sim F_\ell$ for $\ell = 1, \dots, m$ and forming the sensing matrix as follows:

$$A = \frac{1}{\sqrt{m}} \begin{pmatrix} B_1 \\ \vdots \\ B_m \end{pmatrix}, \qquad \text{with} \qquad B_\ell \sim F_\ell, \quad \text{for} \quad \ell = 1, \dots, m. \tag{1.2}$$

We assume the sampling to be isotropic, in the sense that

$$\mathbb{E}(A^*A) = \mathbb{E}\left(\frac{1}{m}\sum_{\ell=1}^{m} B_\ell^* B_\ell\right) = \mathrm{Id}.$$

In particular, this framework encompasses the two following cases, the former having been considered in my PhD. works (Bigot et al., 2016; Boyer et al., 2019a), the latter in Candès and Plan (2011).

(i) **Block-structured sampling from a finite-dimensional isometry.** Let $(\mathcal{I}_k)_{1\leq k\leq M}$ denote a partition of the set $\{1, \dots, d\}$, i.e. a family of disjoint subsets

$$\mathcal{I}_k \subset \{1, \dots, d\} \quad \text{s.t.} \quad \bigsqcup_{k=1}^{M} \mathcal{I}_k = \{1, \dots, d\}.$$

The rows $(\underline{a}_i)_{1\leq i\leq d} \in \mathbb{C}^n$ of an orthogonal matrix $\underline{A} \in \mathbb{C}^{d\times d}$ can be partitioned accordingly into a block dictionary $(\underline{B}_k)_{1\leq k\leq M}$, such that

$$\underline{B}_k = (\underline{a}_i)_{i\in\mathcal{I}_k} \in \mathbb{C}^{|\mathcal{I}_k|\times d}.$$

Define the random blocks $B_1, \dots, B_m$ to be i.i.d. copies of a random block B such that

$$\mathbb{P}\left(B = \underline{B}_k/\sqrt{\pi_k}\right) = \pi_k, \qquad \text{for} \quad k = 1, \dots, M,$$

where $(\pi_k)_{1\leq k\leq M}$ is a discrete probability distribution on $\{1, \dots, M\}$. Note that in this case, all the distributions (F_ℓ)'s are the same one, characterizing the law of the random block B described right above. The sensing matrix A is then constructed by randomly drawing blocks as follows:

$$A = \frac{1}{\sqrt{m}} (B_\ell)_{1\leq\ell\leq m}. \tag{1.3}$$

Moreover, thanks to the renormalization, the random sensing matrix A satisfies $\mathbb{E}(A^*A) = \mathrm{Id}$.

(ii) **Isolated measurements from a finite-dimensional isometry** (standard CS). This is a particular case of the setting described in (i): each block corresponds to a row of the matrix $\underline{A} = (\underline{a}_1|\underline{a}_2|\dots|\underline{a}_d)^*$. Therefore, the sensing matrix is constructed by stacking random vectors drawn from the set of row vectors $\{\underline{a}_1^*, \dots, \underline{a}_d^*\}$ and can be written as follows:

$$A = \frac{1}{\sqrt{m}} (a_\ell)_{1 \le \ell \le m}, \tag{1.4}$$

where the random vectors $(a_\ell)_{1 \le \ell \le m}$ are i.i.d. copies of a random vector a such that

$$\mathbb{P}(a = \underline{a}_j / \sqrt{\pi_j}) = \pi_j,$$

for all $1 \le j \le d$. Here again all the (F_ℓ)'s consists in the same distribution, designating the law of the random vector a. The isotropy condition, i.e. $\mathbb{E}(A^*A) = \mathbb{E}\left(\frac{a_\ell a_\ell^*}{\pi_\ell}\right) = \mathrm{Id}$, is also satisfied.

We should mention in passing that our sampling strategy can be adapted to the case where $\underline{m}$ blocks of measurements could be sensed deterministically, and $m - \underline{m}$ random blocks would be drawn as previously described. But for the sake of conciseness we will only present the case of randomly drawn blocks.

Main result Let S be the set of s largest absolute entries of a target signal and let F be the probability model used to draw random (possibly block-structured) measurements from a finite-dimensional isometry $\underline{A}$. Our recovery guarantees are based on a notion of local coherence, denoted as $\Lambda(S, F)$, and on a global coherence measure $\Gamma(F)$.

Definition 2. *Consider a block sampling strategy as previously described in (1.2) where $(B_k)_{1 \le k \le m}$ are random blocks such that $B_k \sim F_k$, with $F = (F_k)_{1 \le k \le m}$ the associated collection of probability distributions. Let $S \subset \{1, \dots, d\}$ denote the support of the target vector. Define the quantities $\Lambda(S, F)$, and $\Gamma(F)$ to be positive real numbers such that for all $\ell = 1, \dots, m$, when $B_\ell \sim F_\ell$,*

$$\Lambda(S, F) \ge \left\| B_{\ell,S}^* B_{\ell,S} \right\|_{2 \to 2} \qquad a.s. \tag{1.5}$$

$$\Gamma(F) \ge \|B_\ell\|_{1 \to 2}^2 = \max_{1 \le i \le d} \|B_\ell e_i\|_2^2 \qquad a.s. \tag{1.6}$$

where $(e_i)_{1 \le i \le d}$ denote the vectors of the canonical basis, and for a matrix M the norm $\|\cdot\|_{p \to q}$ is defined by $\|M\|_{p \to q} = \sup_{\|x\|_p \le 1} \|Mx\|_q$.

Theorem 3. *Let $x \in \mathbb{R}^d$ or $\mathbb{C}^d$ be a vector supported on S with $|S| = s \le n/2$, such that $\mathrm{sign}(x_S)$ forms a Rademacher or Steinhaus sequence. Let A be the random sensing matrix defined in (1.2) associated with parameters $\Lambda(S, F)$. Then, for every $0 < \delta < 1$, if*

$$\Lambda(S, F) \ge 50 \cdot \Gamma(F) \cdot \ln(3d/\delta), \tag{1.7}$$

and if

$$m \ge 100 \cdot \Lambda(S, F) \cdot \ln\left(\frac{3d}{\delta}\right),$$

then, with probability at least $1 - \delta$, the signal x is exactly recovered via (2) with probability at least $1 - \delta$.

The proof is based on the construction of a dual certificate, a standard path to non-uniform (signal-based) approaches. Note that similar conditions ensure also stable and robust recovery by solving a quadratically-constrained Basis Pursuit, given a level of noise η. Condition (1.7) actually holds (see Section 5 in Adcock et al. (2021)) in realistic settings such as the case of Fourier-Haar measurements (either sensed in a isolated way or by entire lines of the spatial acquisition domain). Existing CS results for such a constrained acquisition generally involve considering the maximum between different types of complex coherence (e.g. $\Theta(S,F) \geq \|B_\ell^* B_{\ell,S}\|_{\infty\to\infty} = \max_{1\leq i\leq d} \|e_i^* B_\ell^* B_{\ell,S}\|_1$ a.s.). The strength of Theorem 3 is that the simple quantity $\Lambda(S,F)$ is governing the required number of measurements for a signal-based reconstruction. This comes at the price of the random signs assumption. The quantity $\Lambda(S,F)$ seems to be sharp in the sense that it is actually driving a sufficient condition on the number of measurements needed by an oracle-type estimator that would know the support S of the target vector, and that would simply make the inversion of a linear system with s unknowns.

Proposition 4. *For every $0 < \delta < 1$, provided*

$$m \gtrsim \cdot \Lambda(S,F) \cdot \ln\left(\frac{2s}{\delta}\right), \tag{1.8}$$

then, with probability at least $1-\delta$, the matrix A_S has full column rank and the oracle-least squares estimator $x^\star \in \mathbb{C}^n$ of the system $y = Az$, defined by

$$x_S^\star = (A_S)^\dagger y, \quad x_{S^c}^\star = 0. \tag{1.9}$$

matches the target signal x.

The proof relies on the control of the least singular value of A_S, ensuring that the oracle estimator is well-defined. One can note that when $\|A_S^* A_S - \mathrm{Id}_S\|_{2\to 2} \leq \tau$, then $\sigma_{\min}(A_S) \geq \sqrt{1-\tau}$ and $\|(A_S^* A_S)^{-1}\|_{2\to 2} \leq \frac{1}{1-\tau}$. Using Bernstein concentration results, the condition $\|A_S^* A_S - \mathrm{Id}_S\|_{2\to 2} \leq \tau$ is ensured provided that

$$m \geq \frac{1+2\tau/3}{\tau^2/2} \cdot \Lambda(S,F) \cdot \ln\left(\frac{2s}{\delta}\right).$$

☕ We must certainly temper the significance of the results, as we compare the bound obtained in Theorem 3 to the one in Proposition 4, which is only a sufficient condition for the oracle-type estimator to be well-defined. However, this is the only result that I know that is coming closer to understanding the phase transition of the basis pursuit program when the sensing matrix is very structured (as drawn from an orthogonal transform). More precisely, the phase transition in the case of BP can be described as the characterization of a function Ψ giving $m = \Psi(s)$ such that the probability of successful reconstruction is $1/2$. This is a powerful information on the convex program in question since it entails that when $m \gtrsim \Psi(s)$, the exact reconstruction is ensured with high probability, and when $m \lesssim \Psi(s)$, the reconstruction fails w.h.p.. In Amelunxen et al. (2014), the authors describe the phase transitions of several convex programs (including (2)), but their proof strategy heavily relies on the Gaussian assumption of the sensing matrix A (using, in some extent, Gordon's escape through a mesh theorem). This question remains totally open in the case of more structured acquisition such as in (1.4).

Optimizing the sampling strategy With Theorem 3 at hand, considering for instance the case of isolated measurements ii, one can optimize the sampling distribution π in order to minimize the required bound on m, as follows,

$$\forall k \in \{1, \dots, d\}, \quad \pi_k = \pi_k^\Lambda = \frac{\|\underline{a}_{k,S}\|_2^2}{\sum_{\ell=1}^d \|\underline{a}_{\ell,S}\|_2^2}. \tag{1.10}$$

The required number m of measurements in Theorem 3 can be then rewritten as follows:

$$m \gtrsim \sum_{\ell=1}^d \|\underline{a}_{\ell,S}\|_2^2 \cdot \ln(6d/\delta). \tag{1.11}$$

In practice, the sampling distributions can be significantly modified when structure is considered in the sensing vectors and in the sparsity of the target vector. Fix the sensing basis to be the Fourier one. Often in image processing (such as MRI or involving any "natural" image), the representation basis is chosen as a wavelet one: hence the support of the target signal is likely to be structured w.r.t. to the decomposition levels of the wavelet transform, hence the name of sparsity-in-level. In such as case, the optimal sampling distribution is depicted in Figure 1.1 (b)(d) and formally reads:

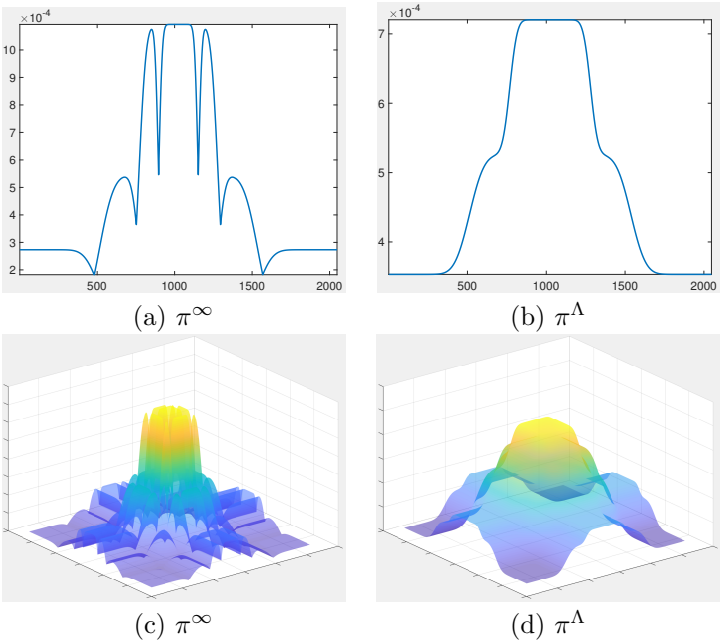
$$\forall k \in \{1, \dots, d\}, \quad \pi_k^\Lambda = \frac{2^{-j(k)} \sum_{p=0}^J 2^{-|j(k)-p|} s_p}{\sum_{\ell=1}^n 2^{-j(\ell)} \sum_{p=0}^J 2^{-|j(\ell)-p|} s_p}, \tag{1.12}$$

where $j(k)$ is the corresponding subband of index k, J is the number of decomposition levels used in the wavelet transform, and s_ℓ is the sparsity of the signal relative to the subband ℓ.

By optimizing the bound obtained in standard results of the literature (Candès and Plan, 2011), in the case of isolated measurements, one would choose

$$\forall k \in \{1, \dots, d\}, \quad \pi_k^\infty = \frac{\|\underline{a}_k\|_\infty^2}{\sum_{\ell=1}^d \|\underline{a}_\ell\|_\infty^2}.$$

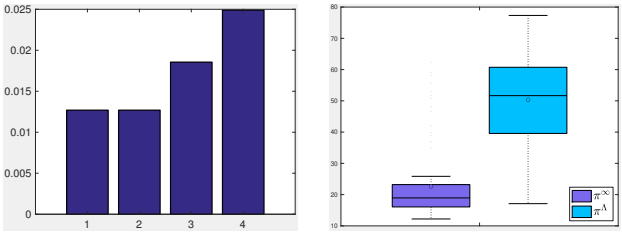
This sampling distribution, as shown in Figure 1.1 (a)(c), does not take into account how the support of the signal of interest interacts with the sensing vectors $(\underline{a}_k)_k$. Given a same number of measurements, drawing samples according to either $(\pi_k^\infty)_k$ or $(\pi_k^\Lambda)_k$ can lead to significant difference in terms of the quality of reconstruction, see Figure 1.2.



(a) π^{∞} (b) π^{Λ}

(c) π^{∞} (d) π^{Λ}

Figure 1.1: Comparison between sampling probability distributions chosen according to different strategies: in (a,b) for 1D signals with a sparse-in-level structure, in (d,e) for 2D signals with the corresponding tensorized structured sparsity. In (a) and (c), the sampling probability distribution π^{∞} is optimized to minimize the global coherence, i.e. $\pi_k \propto \|\underline{a}_k\|_{\infty}^2$; in (b) and (d), the sampling probability distribution π^{Λ} is optimized to minimize an upper bound to $\Lambda(S, \pi)$.



(a) (b)

Figure 1.2: Boxplot of reconstruction PSNR (the higher, the better) in (b) of 100 random signals of length $d = 2048$ having a structured sparsity in the wavelet decomposition as in (a) ($s/n = 6\%$) from 25% measurements drawn according to π^{∞} and π^{Λ} displayed in Figure 1.1 (a,b). More precisely in (a), we represent the sparsity in levels structure of the randomly generated signals: each bar corresponds to the percentage of nonzero coefficients in the subband.

1.2 Sampling rates for ℓ^1-synthesis

Context Most of contributions in the CS literature focus on the reconstruction of vectors, observed through a random sensing matrix A, but sparsely represented in orthogonal bases. In practice, there is a real interest in using redundant transforms: think about a signal which would be the superposition of some vector of the canonical basis (hence, sparse in the canonical basis) and of a sine function (sparse in the Fourier domain); this signal can be therefore more efficiently represented in a (redundant) dictionary resulting from the concatenation of the canonical and Fourier bases.

In such a case, there exist two ways to formulate the reconstruction problem for a level of noise $\eta > 0$:

(i) the ℓ^1-synthesis formulation: consider a signal $x \in \mathbb{R}^d$ sparsely represented in the synthesis operator $D \in \mathbb{R}^{d \times d'}$ ($d' \gg d$) with coefficients $z \in \mathbb{R}^{d'}$, i.e., $x = Dz$, so that z is of low complexity (not x). In order to reconstruct the signal x, one can reconstruct its coefficient z, which leads to the ℓ^1-synthesis program,

$$\min_{z \in \mathbb{R}^{d'}} \|z\|_1 \qquad \text{such that} \qquad \|y - ADz\|_2 \leq \eta. \tag{1.13}$$

(ii) the ℓ^1-analysis formulation: for a signal $x \in \mathbb{R}^d$, consider the analysis operator $\Psi \in \mathbb{R}^{d' \times d}$, so that Ψx is assumed of low complexity. The ℓ^1-analysis program reads as

$$\min_{x \in \mathbb{R}^d} \|\Psi x\|_1 \qquad \text{such that} \qquad \|y - Ax\|_2 \leq \eta. \tag{1.14}$$

Of course, Problems (1.13) and (1.14) are equivalent when Ψ (or D) form an orthogonal basis. However, introducing redundancy in these representation operators changes the geometry of the problem depending on whether we consider its analysis or synthesis version. One could however establish the following link between both. Let $g_K(x) := \inf_{\lambda > 0} \{x \in \lambda K\}$ denote the gauge associated to some convex set K. Consider the gauge $g_{DB_1^{d'}}$, then

$$D\hat{Z} = \inf_{x \in \mathbb{R}^d} g_{DB_1^{d'}}(x) \qquad \text{such that} \qquad \|y - Ax\|_2 \leq \eta, \tag{1.15}$$

where $\hat{Z}$ is the solution set of the ℓ^1-synthesis program (1.13). Moreover, if the atoms $\{d_1, \dots, d_{d'}\}$ of the dictionary D are the extreme points of the level set $\{x, \|\Psi x\|_1 \leq 1\}$, then the solution set of the analysis problem (1.14) matches the one of the synthesis program (1.13). This sheds light on the fact that an analysis program can always be reformulated into a synthesis one.

The ℓ^1-analysis formulation has gained considerable attention in the past years, see e.g., Nam et al. (2013); Candès et al. (2011); Krahmer et al. (2015); Kabanava and Rauhut (2015); Kabanava et al. (2015) compared to its synthesis counterpart Rauhut et al. (2008).

In 2016, Pierre Weiss, Jonas Kahn and I started a collaboration with Maximilian März who was at that time a brilliant PhD. student (Deutsche Qualität) under the supervision of Gitta Kutyniok at TU Berlin. A few years later, we finalized the paper März et al. (2020), recently accepted at Foundations of Computational Mathematics, that I personally find elegant and that yet has difficulty in finding its audience. The rest of the section is dedicated to presenting the insights it contains.

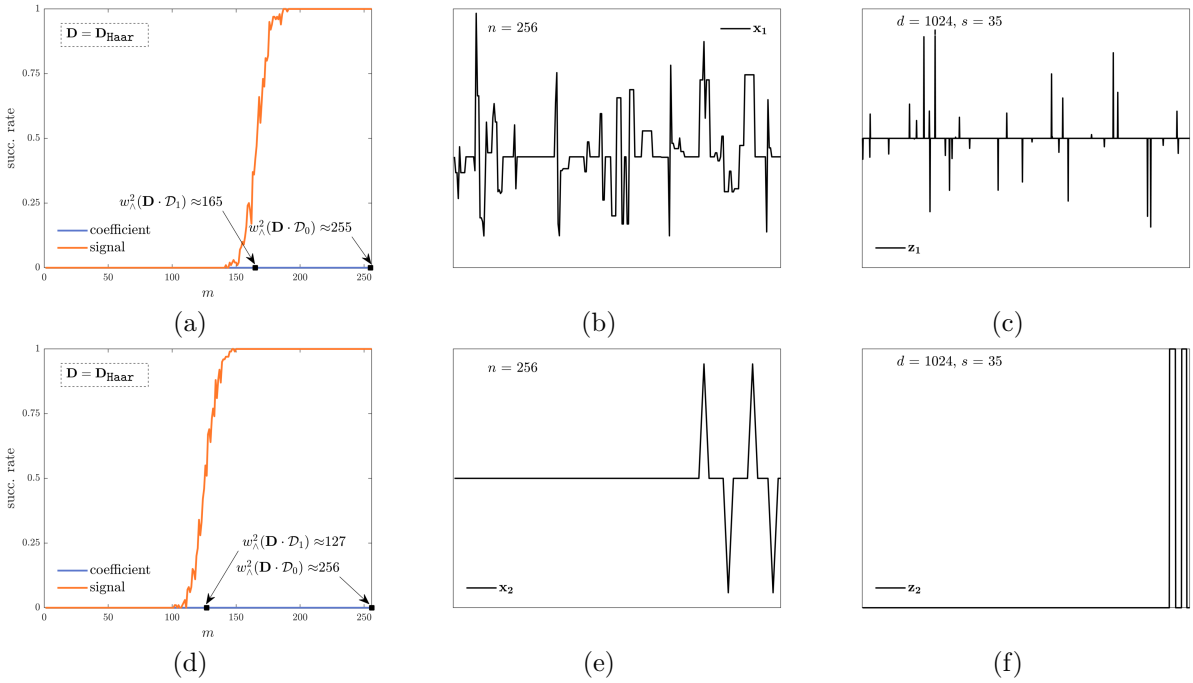


Figure 1.3: Phase transitions of signal recovery: the empirical success rates for the signal recovery are shown in the first column. The different signals are shown in the second column, they admit a sparse representation in a redundant Haar wavelet frame displayed in the third row. Note that in both examples, the coefficient recovery is not possible.

At the risk of sounding repetitive, I want to emphasize the cruel need for non-uniform (i.e. signal-dependent) results in the case of ℓ^1-synthesis, and all the more so as we use redundant representation bases. To convince you, I invite you to look at Figure 1.3, which shows that it is no longer the degree of sparsity of the coefficient representations, which was sufficient to summarize the structure of the signal so far with orthogonal bases, that seems to govern the sampling rate with redundant dictionaries.

Toolkit on phase transition Consider the following generalized regularization framework, involving a convex regularizer f,

$$\min_{x \in \mathbb{R}^d} f(x) \qquad \text{such that} \qquad \|y - Ax\|_2 \leq \eta. \tag{1.16}$$

where the observations $y \in \mathbb{R}^m$ satisfy $y = A\underline{x} + \varepsilon$ with $A \in \mathbb{R}^{m \times d}$ a sensing operator, $\underline{x} \in \mathbb{R}^d$ the target vector, and ε a bounded noise vector such that $\|\varepsilon\|_2 \leq \eta$. A quite straightforward recovery guarantee can be established as follows, involving the descent cone $\mathcal{D}(f, \underline{x})$ of f at $\underline{x}$ defined by $\mathcal{D}(f, \underline{x}) := \text{cone}\left(\{h, f(\underline{x} + h) \leq f(\underline{x})\}\right)$, and the minimal singular value $\sigma_{\min}\left(A, \mathcal{D}(f, \underline{x})\right)$

of A restricted to the set $\mathcal{D}(f, \underline{x})$, i.e., $\sigma_{\min}\left(A, \mathcal{D}(f, \underline{x})\right) := \min_{x \in \mathcal{D}(f,\underline{x})} \|Ax\|_2$

Proposition 5 (Tropp, 2015)**.**

1. *(Noiseless measurements) If $\eta = 0$, then the following conditions are equivalent*
   - *$\hat{x} = \underline{x}$, with $\hat{x}$ the unique solution of (1.16),*
   - $\mathrm{null}(A) \cap \mathcal{D}(f, \underline{x}) = \{0\}$,
   - $\sigma_{\min}\left(A, \mathcal{D}(f, \underline{x})\right) > 0$.
2. *(Noisy measurements) For $\eta > 0$, any solution $\hat{x}$ of (1.16) satisfies the error bound*

$$\|\hat{x} - \underline{x}\|_2 \leq \frac{2\eta}{\sigma_{\min}\left(A, \mathcal{D}(f, \underline{x})\right)}.$$

The quantity $\sigma_{\min}\left(A, \mathcal{D}(f, \underline{x})\right)$ contains all the keys to the success of the reconstruction in both study settings, with or without noise. The bad news is that it is generally NP-hard to evaluate it (as an instance of testing the co-positivity of matrices). The good news is that we can theoretically do it when the sensing operator is assumed to be a Gaussian matrix.

Theorem 6 (Tropp, 2015)**.** *Assume that A is a Gaussian matrix. Then, with a probability larger than $1 - e^{-u^2/2}$,*

$$\sigma_{\min}\left(A, \mathcal{D}(f, \underline{x})\right) \geq \sqrt{m-1} - \omega(\mathcal{D}(f, \underline{x})) - u,$$

where the Gaussian width $\omega(K)$ of a cone K is given by

$$\omega(K) := \mathbb{E}_{g \sim \mathcal{N}(0, \mathrm{Id}_d)} \left[\sup_{h \in K \cap \mathbb{S}^{d-1}} \langle g, h \rangle \right].$$

The Gaussian width $\omega(\mathcal{D}(f, \underline{x}))$ determines a sufficient condition for the (robust) recovery of $\underline{x}$ as soon as $m \gtrsim \omega^2(\mathcal{D}(f, \underline{x})) + 1$. It actually provides more information about the reconstruction in the noiseless case: it governs the phase transition of the recovery success.

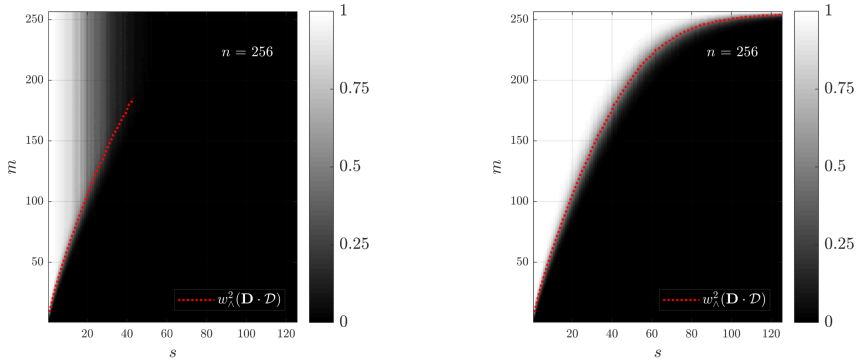
Theorem 7 (Amelunxen et al., 2014)**.** *When A is a Gaussian matrix, and $\eta = 0$,*

- *if $m \geq \omega^2(\mathcal{D}(f, \underline{x})) + 1 - \ln(\delta)\sqrt{d}$, the resolution of Problem (1.16) recovers $\underline{x}$ with probability larger than $1 - \delta$;*
- *if $m \leq \omega^2(\mathcal{D}(f, \underline{x})) + 1 + \ln(\delta)\sqrt{d}$, the resolution of Problem (1.16) fails to recover $\underline{x}$ with probability larger than $1 - \delta$.*

Some selected results Based on the previous tools, in März et al. (2020), we derive sampling rates for the coefficient and signal reconstruction through the following ℓ^1-synthesis approach,

1. first we recover a candidate coefficient $\hat{z} \in \hat{Z}$ by solving (1.13), with $\hat{Z}$ the solution set of (1.13);
2. then we apply the dictionary D to get the signal solution $\hat{x} := D\hat{z} \in \hat{X} := D\hat{Z}$, with $\hat{X}$ the solution set of the signal ℓ^1-synthesis.

We will not comment in details on how our results highlight that the signal reconstruction can depart from the coefficient one, but one can refer to Figure 1.4 to get more intuition. Figure 1.4 shows the phase transitions for both signal and coefficient recoveries. What is striking is that when the sparsity increases, the coefficient reconstruction fails no matter how many measurement we perform, whereas the signal reconstruction is still possible. This emphasizes that the signal can be uniquely determined, without disambiguating its coefficient representation got by solving the ℓ^1-synthesis problem. We also notice that on the left hand side of both plots, when the reconstructions of the coefficient and the signal are concomitant, they seem to share the same phase transition. Sampling rates for coefficient and signal reconstruction are indeed shown to match, so we only



(a) (b)

Figure 1.4: Phase transitions of coefficient and signal recovery by ℓ^1-synthesis. In (a), we show the empirical probability that atomic coefficient representations are successfully recovered via solving (1.13), whereas in (b) we show the empirical probability for the associated signal reconstruction. The underlying dictionary is a redundant Haar wavelet frame with three decomposition levels and the defining s-sparse coefficients are chosen at random. The brightness of each pixel reflects the observed probability of success, reaching from certain failure (black) to certain success (white). The dotted line shows our predictions for the location of the phase transitions.

present in the sequel the developments for the signal recovery.

Theorem 8. *Let $\underline{x} \in \mathbb{R}^d$ be the target signal, $y = A\underline{x} + \varepsilon$ with $A \in \mathbb{R}^{m\times d}$ a Gaussian matrix and $\|\varepsilon\|_2 \leq \eta$. If*

$$m \geq \underline{m} := \omega^2\left(D\mathcal{D}(\|\cdot\|_1, \underline{z})\right)$$

for any ℓ^1-representer $\underline{z} \in \operatorname{argmin}_{\underline{x}=Dz}\|z\|_1$, then for any $\hat{x} \in \hat{X}$,

$$\|\underline{x} - \hat{x}\|_2 \leq \frac{2\eta}{\sqrt{m-1} - \sqrt{\underline{m}-1}}$$

The proof mainly relies on recasting the procedure of the signal reconstruction in the form of (1.16). If, as previously said, the sampling rates of the signal and coefficient recoveries match, the stability to noise can actually drastically differ, as the error bound for the coefficient involves $\sigma_{\min}\left(D, \mathcal{D}(\|\cdot\|_1, \underline{z})\right)$. Now all the efforts are to be deferred on a fair evaluation of the conic mean width $\omega^2\left(D\mathcal{D}(\|\cdot\|_1, \underline{z})\right)$ of a linearly transformed cone. This is the purpose of the following result,

using a brilliant idea of Jonas Kahn: (i) decompose the polyhedral cone $\mathcal{D}(\|\cdot\|_1, \underline{z})$ into its lineality (containing all the lines of the cone) and range parts, (ii) embed its range, which is a pointed polyhedral cone into a circular cone of circumangle α, (iii) bound the initial squared conic mean width by the sum of the squared mean widths of the lineal (this is the easy part as, it is a vector space, and its squared mean width boils down to its dimension) and of the range (using a clever upper bound through (ii) involving the circumangle α and the number of extreme rays of the range - that we know to be finite).

Theorem 9 (Unformal)**.** *Let $C := D\mathcal{D}(\|\cdot\|_1, \underline{z})$, such that $C = C_L \oplus C_R$ with C_L (resp. C_R) the lineality (resp. range) part of C. Then, C_R is a polyhedral cone generated by at most k extreme rays, with k of the order of d, and embedded into a circular cone of circumangle α. An upper-bound on the critical number of measurement $\underline{m}$ is given by*

$$\underline{m} \lesssim \bar{s} + \tan^2(\alpha)\ln(d').$$

where $\bar{s}$ is the maximal degree of sparsity of the ℓ^1-representers $\operatorname{argmin}_{\underline{x}=Dz}\|z\|_1$ of $\underline{x}$.

The bound in Theorem 9 encloses the bound $\bar{s}$ on the squared conic mean width of the lineality part and the bound $\tan^2(\alpha)\ln(d)$ of the squared conic mean width of any $k(\simeq d)$-polyhedral cone embedded into an α-circular cone. Of course, it requires to compute the circumangle α: this is a convex problem! Recall that

$$\cos(\alpha) = \sup_{\|\theta\|_1=1} \inf_{\substack{x\in D\mathcal{D}(\|\cdot\|_1,\underline{z}) \\ \|x\|_2=1}} \langle x, \theta\rangle.$$

We actually manage to theoretically evaluate all these quantities (k and α) by dissecting the convex cones in presence of dictionaries chosen as orthogonal bases, but also as the concatenation of convolution matrices. The former matches (up to log factor) the existing bound on the number of measurements required by the state-of-the-art results, which underlines the sharpness of the derived upper-bound in Theorem 9. The latter provides the first meaningful bound showing that compressive sensing of a signal sparsely represented in a redundant transform is indeed possible, without the need of exact reconstruction of its coefficients!

☕ There is clearly room for improvement, as the quantities considered here remain partly cryptic, and necessitate a case-by-case study depending on the dictionary in play. A first step would be to try to use this approach to evaluate sampling rates for some dictionary class such as tight frames.

1.3 Off-the-grid compressed sensing

Context Until now, the models considered in CS are discrete models, which do not reflect the continuous intrinsic nature of the objects in play. For example, when we think of an astronomy image, there is no reason for galaxies and stars to be positioned naturally on a discretization grid. Off-the-grid CS compensates for this finite-dimensional simplification by considering that the object to be reconstructed is no longer a vector, but an atomic Radon measure. The degree of sparsity then becomes the number of Dirac masses in the target measure. This refers to an old problem considered in the 1930's by Beurling (1938). After a long scientific winter, Yohann De Castro during his PhD. thesis has brought it up to date in 2012 (De Castro and Gamboa, 2012; Azais et al., 2015), and then Emmanuel Candès and Carlos Fernandez-Granda made it even more popular, see Candès and Fernandez-Granda (2013, 2014) but also Bredies and Pikkarainen (2013).

I met Yohann De Castro and Joseph Salmon during the summer school of Saint-Flour, edition 2015. We were enjoying the lecture given by Sara Van de Geer, on the finite-dimensional square-root lasso (Owen, 2007; Belloni et al., 2011; Sun and Zhang, 2012, 2013; Chrétien and Darses, 2014), which is a version of the sparse regression problem, whose goal is to estimate simultaneously the sparse vector of interest and the unknown noise level contaminating the observations. To make the memory of Saint-Flour last, and as Yohann is a specialist of the continuous counterpart of the lasso, we started a collaboration aiming at adapting these robust strategies dissected in the lecture notes (van de Geer, 2016) to the case of the infinite dimension. While some of the works may not have a major impact on the community, they do nonetheless remain significant to a young academic who is plagued by doubts and uncertainty. I sincerely thank Yohann and Joseph for their support, their enthusiasm in this project and their friendship since then.

Setting We aim at estimating an atomic measure $\underline{\mu} = \sum_{i=1}^{s} \underline{\alpha}_i \delta_{t_i} \in \mathcal{M}(\mathbb{T})$, where $\mathcal{M}(\mathbb{T})$ is the space of Radon measures supported on the one-dimensional torus $\mathbb{T}$, so that $(\underline{\alpha})_{1\leq i\leq s} \in \mathbb{R}^s$ and $(\underline{t}_i)_{1\leq i\leq s} \in \mathbb{T}^s$ respectively denote the amplitudes and the spike locations of the target measure. We observe

$$y = \mathcal{F}_m(\underline{\mu}) + \varepsilon \in \mathbb{C}^{2m+1} \tag{1.17}$$

where $\mathcal{F}_m$ is the sensing operator mapping a Radon measure to its m first Fourier coefficients, i.e.

$$\begin{aligned} \mathcal{F}_m : \mathcal{M}(\mathbb{T}) &\longrightarrow \mathbb{C}^{2m+1} \\ \mu &\longmapsto \left(\int_{\mathbb{T}} \exp(-2\imath\pi k t)\mu(dt) \right)_{|k|\leq m}, \end{aligned}$$

and $\varepsilon \in \mathbb{C}^{2m+1}$ is a complex Gaussian vector, such that $\varepsilon = \mathfrak{Real}(\varepsilon) + \imath\mathfrak{Im}(\varepsilon)$, $\mathfrak{Real}(\varepsilon), \mathfrak{Im}(\varepsilon) \sim \mathcal{N}(0, \underline{\sigma}^2\mathrm{Id})$ with unknown $\underline{\sigma} > 0$. Since the noise level is unknown, we define and study the Concomitant Beurling Lasso (CBLasso)[1], which is a convex penalization of a joint log-likelihood estimator:

$$(\hat{\mu}, \hat{\sigma}) \in \operatorname*{argmin}_{\substack{\mu\in\mathcal{M}(\mathbb{T}) \\ \sigma>0}} \frac{1}{2n\sigma} \|y - \mathcal{F}_m(\mu)\|_2^2 + \frac{\sigma}{2} + \lambda\|\mu\|_{TV} \tag{1.18}$$

where $\|\cdot\|_{TV}$ is the total variation norm, defined by duality on the space of bounded continuous functions equipped with the ℓ^∞-norm,

$$\|\mu\| := \sup_{\|f\|_\infty\leq 1} \mathfrak{Real}\left(\int_{\mathbb{T}} \bar{f} d\mu \right).$$

When the measure is atomic, the TV norm boils down to the ℓ^1-norm of the spike amplitudes, hence the continuous counterpart of the ℓ^1-norm.

When the solution is reached for $\hat{\mu}, \hat{\sigma} > 0$, optimality conditions provide:

$$\begin{cases} \hat{\sigma} &= \|y - \mathcal{F}_m(\hat{\mu})\|_2 / \sqrt{n} \\ \hat{\mu} &\in \operatorname{argmin}_{\mu\in\mathcal{M}(\mathbb{T})} \frac{1}{2n\sigma} \|y - \mathcal{F}_m(\mu)\|_2 + \lambda\|\mu\|_{TV} \end{cases}$$

[1] Persifiers could see a "Claire Boyer Lasso" there.

where the last expression looks like a (continuous) lasso problem without any square on the data-fidelity term, hence the name of square-root lasso.

Regarding the numerical resolution, we resort to solve an SDP formulation of the dual problem of (1.18). Note that since then, other algorithmic approaches have been explored, in particular working directly with the primal formulation, see Denoyelle et al. (2019).

Proposition 10. *Denote* $\Delta = \left\{c \in \mathbb{C}^{2m+1} : \|\mathcal{F}_m^*(c)\|_\infty \leq 1, \quad m\lambda^2\|c\|_2^2 \leq 1\right\}$, *the dual problem of the CBLasso* (1.18) *reads*

$$\hat{c} \in \operatorname*{argmax}_{c\in\Delta} \lambda\langle y, c\rangle. \tag{1.19}$$

The following equation holds between the primal and dual solutions:

$$y = m\lambda\hat{\sigma}\hat{c} + \mathcal{F}_m(\hat{\mu}). \tag{1.20}$$

The dual vector $\hat{c}$ actually corresponds to coefficients of the so-called dual polynomial $\hat{p} = \mathcal{F}_m^*(\hat{c})$. By strong duality, $\hat{p}$ interpolates the spike signs of $\hat{\mu}$ (which is informative as soon as the dual polynomial is non-constant). By the constraint $\|\mathcal{F}_m^*(\hat{c})\|_\infty \leq 1$ in Proposition 10, the support of $\hat{\mu}$ is then included in the roots of the derivative of the dual polynomial. The dual problem has the advantage to work with finite-dimensional objects but involves functional constraint. Luckily, they admit an SDP representation.

Lemma 11 (Candès and Fernandez-Granda, 2013)**.** *For any* $c \in \mathbb{C}^{2m+1}$, *the following conditions are equivalent,*

(i) $\|\mathcal{F}_m^*(c)\|_\infty^2 \leq 1$,

(ii) $\exists \Lambda \in \mathbb{C}^{(2m+1)\times(2m+1)}$ *s.t.* $\begin{cases} \Lambda^* = \Lambda, \\ \begin{pmatrix} \Lambda & c \\ c^* & 1 \end{pmatrix} \succeq 0, \\ \sum_{i=1}^{n-j+1} \Lambda_{i,i-j-1} = \delta_{j,1} \quad \forall j, \end{cases}$

where $\delta_{k,\ell}$ *is the Kronecker symbol.*

With this at hand, the dual problem (1.19) can be reformulated as

$$\max_{\substack{c\in\mathbb{C}^{m'} \\ \Lambda\in\mathbb{C}^{m'\times m'}}} \lambda\langle y, c\rangle \text{ such that } \begin{cases} \Lambda^* = \Lambda, \\ \begin{pmatrix} \Lambda & c \\ c^* & 1 \end{pmatrix} \succeq 0, \\ \sum_{i=1}^{n-j+1} \Lambda_{i,i-j-1} = \delta_{j,1} \quad \forall j, \\ \begin{pmatrix} \mathrm{Id}_{m'} & \lambda\sqrt{m'}c \\ \lambda\sqrt{m'}c^* & 1 \end{pmatrix} \succeq 0. \end{cases} \tag{1.21}$$

setting $m' = 2m + 1$. Finally, the CBLasso is solved by taking the following path:

1. For a fixed $\lambda > 0$, compute $\hat{c}$ the solution of (1.21). To do so, use for instance the `cvx` toolbox (Grant and Boyd, 2014);
2. Identify the potential support $\left\{\hat{t}_j\right\}_{j=1,\dots,\hat{s}}$ of $\hat{\mu}$ by computing the roots of $1-|\hat{p}|^2 = 1-|\mathcal{F}_m^*(\hat{c})|^2$. Then form the design/sampling matrix $X \in \mathbb{C}^{(2m+1)\times\hat{s}}$, defined by $X := \left(\exp(-2\imath\pi k\hat{t}_j)\right)_{|k|\leq m, j=1,\dots,\hat{s}}$;

3. Solve a finite-dimensional square-root lasso with inputs (X, λ, y) to get the amplitudes $\hat{\alpha}$ and the noise level $\hat{\sigma}$;
4. Output $\hat{\mu} = \sum_{j=1}^{\hat{s}} \hat{\alpha}\delta_{\hat{t}_j}$.

Contributions First, we show that the conditions (such as compatibility, restricted eigenvalue or restricted isometry properties) guaranteeing the uniform reconstruction of any sparse measure are proscribed in such a problem, so that we restrict the study to measure-dependent (non-uniform) reconstruction results. Then we obtain prediction results of the following form.

Theorem 12. *Assume that*

- *(sampling rate condition)* $\lambda \cdot \|\underline{\mu}\|_{TV}/\sqrt{2}\underline{\sigma} \lesssim 1$,
- *(separation condition) the distance between the spikes in* $\underline{\mu}$ *is at least* $1.26/m$.

Then, the estimator $\hat{\mu}$ *solution to CBLasso with* $\lambda \gtrsim \sqrt{\log(n)/n}$ *satisfies w.h.p.,*

$$\frac{1}{m}\left\|\mathcal{F}_m(\hat{\mu} - \underline{\mu})\right\|_2^2 \lesssim s\lambda^2\underline{\sigma}^2 = O\left(\frac{s\underline{\sigma}^2 \log n}{n}\right).$$

Proofs are adapted from Tang et al. (2014), amended by the Rice method for a non-Gaussian process to deal with the noise level dependency in the bounds. This bound actually matches the fast convergence rate exhibited in Tang et al. (2014) when the noise level is assumed to be known. The proof relies on exploiting the properties of a dual certificate, which is in this specific instance, a trigonometric polynomial (lying in the range of $\mathcal{F}_m^*$) satisfying first-order optimality (KKT) conditions of Problem (1.18). We also establish localization results ensuring that the mass in the reconstructing measure mainly lies in "near" regions (determined by the inverse of the frequency cut m) of the original spikes. We also show that provided that a given Dirac mass localized at $\underline{t}$ with amplitude $|\underline{\alpha}| > C\underline{\sigma}s\lambda$ large enough, for a numerical constant $C > 0$, then there exists a reconstructed spike at $\hat{t}$ in $\hat{\mu}$, such that

$$\operatorname{dist}(\hat{t}, \underline{t}) \lesssim \frac{1}{n}\sqrt{\frac{C\underline{\sigma}s\lambda}{|\underline{\alpha}| - C\underline{\sigma}s\lambda}}.$$

Again, the Rice method is used in conjunction of the approaches developed in Azais et al. (2015); Tang et al. (2014); Fernandez-Granda (2013) for the spike detection. Overall, the spike deconvolution with unknown noise level through CBLasso enjoys the same guarantees as when using Beurling lasso with known noise level, provided that the signal-to-noise ratio is bounded. We also provide guarantees that the no-overfitting regime occurs, i.e., that $\|\mathcal{F}_m(\hat{\mu}) - y\|_2^2 > 0$, as required. A numerical resolution is possible by a root-finding search, and we ensure the necessary condition that the dual polynomial is never constant.

☕ As an ongoing work with Yohann De Castro and Vincent Duval, we are studying how to leverage off-the-grid approaches to tackle low-rank tensor denoising. This would give a new light to this problem that has so far been studied through trend filtering (van de Geer and Ortelli, 2021; Gong et al., 2020); this could have a wider spin-off, particularly for the problem of tensor PCA (Richard and Montanari, 2014; Arous et al., 2020). The problem of tensor denoising can be actually lifted to the space of measures: assume we observed a d-dimensional k-way tensor Y such that

$$Y = \underline{X} + W$$

where $\underline{X}$ is an unknown d-dimensional symmetrical k-tensor supposed of rank r, and W is a noise tensor. Therefore, $\underline{X}$ admits the following expansion

$$\underline{X} = \sum_{\ell=1}^{r} \underline{\alpha}_\ell \underline{x}_\ell^{\otimes k}$$

for some amplitudes $(\underline{\alpha}_\ell)_{1\leq\ell\leq r}$ and unit d-dimensional vectors $(\underline{x}_\ell)_{1\leq\ell\leq r}$. This parameterization of rank-r tensors can be embedded into the space of Radon measures, so that denoising Y to estimate $\underline{X}$ amounts to reconstructing a target measure

$$\underline{\mu} = \sum_{\ell=1}^{r} \underline{\alpha}_\ell \delta_{\underline{x}_\ell} \in \mathcal{M}\left(\mathbb{S}^{d-1}\right).$$

On can then recast the tensor reconstruction into a Beurling lasso program using, instead of Fourier sampling, the following sensing operator:

$$\begin{aligned} \Phi : \mathcal{M}\left(\mathbb{S}^{d-1}\right) &\longrightarrow \mathbb{R}^d \\ \mu &\longmapsto \int_{\mathbb{S}^{d-1}} x^{\otimes k} d\mu(x). \end{aligned}$$

1.4 Representer theorem for convex regularization

The last part of this chapter is dedicated to the work Boyer et al. (2019b), which somewhat bridges the gap between the community of inverse problems, which has been around until now, and that of machine learning. This work is the result of two competing teams, respectively based in Toulouse and Paris, that decided to join forces to deliver a single, higher impact paper. I thank all the collaborators, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric de Gournay and Pierre Weiss for this nice output.

Context Representer theorems (Schölkopf et al., 2001) are well-known in machine learning as they are the reason of the practical success of kernel machines. They characterize structural properties of minimizers of a regularized empirical risk, without actually needing to explicitly solve the problem. This prior can be injected in the empirical risk minimization so that learning even in infinite-dimensional spaces can be recast into a finite-dimensional, and then scalable, optimization problem.

The purpose of this section is to extend representer theorems to a general convex optimization program. In some specific instances, some were already known. Consider for example the ℓ^1-minimization problem under m-dimensional equality affine constraints, i.e., the basis pursuit program (2). Chen and Donoho (1994) show that (2) admits m-sparse solutions, that is to say, some solution $\hat{x} \in \mathbb{R}^d$ to the basis pursuit can be decomposed as

$$\hat{x} = \sum_{\ell=1}^{m} \alpha_\ell e_{\sigma(\ell)}$$

for some amplitudes $(\alpha_\ell)_\ell$, $(e_\ell)_\ell$ being the vectors of the d-dimensional canonical basis, and σ some permutation on $\{1, \ldots, d\}$. This result is of interest if $m \ll d$, meaning that solutions of low-complexity always exist. The attentive reader will have already noticed that the "atoms" $(\pm e_\ell)_\ell$

used in the decomposition of $\hat{x}$ are actually extreme points of the ℓ^1-ball, i.e., extreme points of level sets of the regularizer in play. This observation can be extrapolated to a higher level of abstraction, encompassing standard variational problems.

Contributions Let E be a vector space. Consider the following generalized basis problem

$$\min_{u \in E} R(u) \quad \text{such that} \quad \Phi u = y \tag{$\mathcal{P}$}$$

where $R : E \to \mathbb{R} \cup \{+\infty\}$ is a convex regularizer, and $\Phi : E \to \mathbb{R}^m$ is a linear map. We characterize the extreme points structure of the solution set (provided that the latter is non-empty), which can be rewritten as the intersection of a peculiar level set of R and the preimage of y by Φ, i.e., $\{u \in E : R(u) \leq \min(\mathcal{P})\} \cap \Phi^{-1}(\{y\})$.

Theorem 13. *Let E be a vector space, $R : E \to \mathbb{R} \cup \{+\infty\}$ a convex function, and $\Phi : E \to \mathbb{R}^m$ linear such that argmin($\mathcal{P}$) is nonempty. Assume that $\{R \leq \min(\mathcal{P})\}$ is linearly closed and contains no line.*

If $\min(\mathcal{P}) > \inf_E R$, then each extreme point of argmin($\mathcal{P}$) is

- *a convex combination of (at most) M extreme points of $\{R \leq \min(\mathcal{P})\}$,*
- *or a convex combination of (at most) $M - 1$ points, each an extreme point or in an extreme ray of $\{R \leq \min(\mathcal{P})\}$.*

If $\min(\mathcal{P}) = \inf_E R$, the previous description holds involving an extra point in both cases. The proof relies on two steps: (i) showing that each extreme point of argmin($\mathcal{P}$) belongs to a face of $\{R \leq \min \mathcal{P}\}$ of dimension at most $M - 1$, (ii) use Klee's extension (Dubins, 1962; Klee, 1963) of Caratheodory's theorem for unbounded sets.

Note that the results presented in this section are obtained using geometrical considerations, and without relying on optimality conditions, as standard ML representer theorems do.

This theorem allows to characterize the solutions of various problems involving the most popular penalties:

- in the case of linear programming: Let $\psi \in \mathbb{R}^n$ be a vector and $\Phi \in \mathbb{R}^{m \times n}$ be a matrix and consider the following linear program in standard (or equational) form:

$$\inf_{\substack{u \in \mathbb{R}^n_+ \\ \Phi u = y}} \langle \psi, u \rangle \tag{1.22}$$

  The extreme points of its solution set can be shown to be m-sparse, i.e. of the form

$$u^\star = \sum_{i=1}^{m} \alpha_i e_i, \quad \alpha_i \geq 0; \tag{1.23}$$

- in the case of ℓ^1-analysis prior, we consider an analysis operator $L \in \mathbb{R}^{p \times n}$ and $\Phi \in \mathbb{R}^{m \times n}$ be a matrix, and we want to solve

$$\inf_{\substack{u \in \mathbb{R}^n \\ \Phi u = y}} \|Lu\|_1 . \tag{1.24}$$

Assuming that L is surjective (so $p \leq n$), we can show that the solutions can be written as

$$u^\star = \sum_{i \in I} \alpha_i L^+ e_i + u_K, \quad \alpha_i \in \mathbb{R}, \tag{1.25}$$

where $u_K \in \ker(L)$ and $I \subset \{1, \dots, p\}$ is a set of cardinality $|I| \leq m - \dim(\Phi \ker(L))$.

- in the case of semi-definite programming, we aim at solving

$$\inf_{\substack{M \succeq 0 \\ \Phi(M) = y}} \langle A, M \rangle, \tag{1.26}$$

and extreme points of its solution set are shown to be matrices of rank m;

- in the case of nuclear norm minimization, as extreme points of the regularizer level set are matrices of rank-one, we directly obtain that we recover rank-m solutions;

- towards an infinite-dimensional setting, in the case of moment problem: Let Ω be a compact metric space, $\mathcal{M}(\Omega)$ be the set of Radon measures on Ω and let $\mathcal{M}_+(\Omega) \subseteq \mathcal{M}(\Omega)$ be the cone of nonnegative measures on Ω. Let ψ and $(\phi_i)_{1 \leq i \leq m}$ denote a collection of continuous functions on Ω. Now, let $\Phi : \mathcal{M}(\Omega) \to \mathbb{R}^m$ be defined by $(\Phi\mu)_i = \langle \phi_i, \mu \rangle$, where $\langle \phi_i, \mu \rangle := \int_\Omega \phi_i d\mu$, and consider the following linear program in standard form

$$\inf_{\substack{\mu \in \mathcal{M}_+(\Omega) \\ \Phi\mu = y}} \langle \psi, \mu \rangle. \tag{1.27}$$

Provided that the solution set of the previous problem is non-empty, its extreme points are m-sparse, i.e. of the form:

$$\mu^\star = \sum_{i=1}^m \alpha_i \delta_{x_i}, \ x_i \in \Omega, \ \alpha_i \geq 0; \tag{1.28}$$

- in the case of total variation gradient penalization which is defined as for any locally integrable function u as

$$TV(u) := \sup \left(\int u \mathrm{div}(\phi)\, dx, \phi \in C_c^1(\mathbb{R}^d)^d, \sup_{x \in \mathbb{R}^d} \|\phi(x)\|_2 \leq 1 \right).$$

If the above quantity is finite, we say that u has bounded variation and its gradient Du is a Radon measure, with

$$TV(u) = \int_{\mathbb{R}^d} |Du| = \|Du\|_{(\mathcal{M}(\mathbb{R}^d))^d}.$$

The extreme points of the TV ball have been described by Fleming (1957) and Ambrosio et al. (2001) as indicator of simple sets (i.e. simply connected sets with no holes). This leads to the description of the extreme points of the solution set as a sum of m indicator of simple sets. This description has in particular inspired the PhD. work of Romain Petit at Inria Paris under the supervision of Vincent Duval and Yohann De Castro to propose efficient Frank-Wolfe-based algorithms to solve this type of problems.

Chapter 2

Contributions in machine learning

Contents

After being recruited as a lecturer in 2016 at Sorbonne University (SU), I took the responsibility of machine learning courses at master 2 level, but also the direction of the statistics program of the master of mathematics and applications of SU. Decisive scientific encounters and a consequent investment in teaching have therefore gradually oriented my research towards the highly competitive domain of statistical and machine learning.

In the following, I present some of my contributions in learning, and more particularly in supervised learning. Supervised learning is the science of prediction: how from training examples $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ can one infer a link between the inputs $X_1, \ldots, X_n$ and their corresponding outputs $Y_1, \ldots, Y_n$? And more importantly how can one generalize the learned link on these examples to unseen data to be able to predict Y_{new} from only X_{new}? This is the central question of supervised machine learning, to which tree-based methods and neural networks provide a powerful and efficient answer.

Outline The predictors mentioned just above are usually obtained by minimization of an (empirical) risk. With this in mind, I present in Section 2.1 two new learning algorithms addressing this issue. They respectively belong to the classes of boosting methods and stochastic optimization, and have been respectively developed in collaboration with two of my colleagues in the LPSM, Maxime Sangnier and Antoine Godichon-Baggioni, in the works Fouillen et al. (2022) and Boyer and Godichon-Baggioni (2020). In Section 2.2, I explore the possible connections between tree-based methods and neural networks, 2 classes of powerful machine learning predictors, and how one can benefit from the other. In particular, I present the developments of Arnould et al. (2021)

and Arnould et al. (2022). Finally, in Section 2.3, I investigate a hot topic related to ML predictors achieving a zero empirical risk but still performing well on unseen data. The so-called benign overfitting regime is in particular studied in the case of interpolating random forests (Arnould et al., 2022).

The last tree referred papers have been written during the PhD. thesis of Ludovic Arnould that I co-advise with Erwan Scornet (Ecole Polytechnique). Patrick Lutz, a master 2 intern with incredible scientific maturity, now in PhD. at Boston University, also made it possible to complete a research project numerically involved.

2.1 New learning algorithms

2.1.1 Accelerated proximal boosting

Boosting algorithms, and more specifically gradient boosting ones, consist in a class of methods, currently regarded as one of the best off-the-shelf learning techniques for tabular data in several real-world situations. They act by sequentially aggregating weak learners, to build a more complex and accurate model. This assembly is iteratively performed, taking into account the performance of the model built so far. Gradient boosting can be seen as a greedy optimization procedure (similar to gradient descent), aimed at minimizing an empirical risk over the set of linear combinations of weak learners.

Setting More formally, given a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, supervised learning aims at explaining Y given X, thanks to a measurable function $f^\star \colon \mathcal{X} \to \mathbb{R}$. In this context, $f^\star(X)$ may represent several quantities, depending on the task at hand, for which the most notable examples are the conditional expectation $x \in \mathcal{X} \mapsto \mathbb{E}[Y|X = x]$ or the conditional quantiles of Y given X for regression, as well as the regression function $x \in \mathcal{X} \mapsto \mathbb{P}(Y = 1|X = x)$ for ± 1-classification. Often, this target function $f^\star$ is a minimizer of the risk $\mathbb{E}(\ell(Y, f(X)))$ over all measurable functions f, where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a suitable convex loss function (respectively the square function and the pinball loss in the regression examples previously mentioned).

Since the distribution of (X, Y) is generally unknown, the minimization of the risk is out of reach. One would rather deal with its empirical version instead:

$$\min_{f \in \mathrm{span}(\mathcal{F})} \mathcal{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) \tag{2.1}$$

where $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ is a training sample of i.i.d. copies of (X, Y), and $\mathcal{F}$ is a class of functions from $\mathcal{X}$ to $\mathbb{R}$ (usually chosen as the class of classification and regression trees). In practice the objective function is often regarded as a function from $\mathbb{R}^n$ to $\mathbb{R}$, considering that it depends on f only through the training samples $(f(X_1), \ldots, f(X_n)) \in \mathbb{R}^n$. Therefore, in the following we will denote by $\nabla_n \mathcal{R}_n(f)$ the n-dimensional gradient of $\mathcal{R}_n$ with respect to the evaluations $(f(X_1), \ldots, f(X_n)) \in \mathbb{R}^n$.

Gradient boosting If Adaboost (Freund and Schapire, 1997) is the seminal boosting algorithm by excellence, it can be actually viewed as an instance of gradient boosting algorithms (Friedman, 2001) when using the particular exponential loss. The principles of a general gradient boosting approach are recalled in Algorithm 1. When the function ℓ, used in the definition of $\mathcal{R}_n$, is not

Algorithm 1 Gradient boosting.

Input: $\gamma_1, \dots, \gamma_T > 0$ (gradient steps).

1: Set $f_0 \in \operatorname{argmin}_{g \in \mathcal{F}} \mathcal{R}_n(g)$ (initialization).
2: **for** $t = 0$ **to** $T - 1$ **do**
3: Compute $r \leftarrow -\nabla_n \mathcal{R}_n(f_t)$ (pseudo-residuals).
4: Compute $g_{t+1} \in \arg\min_{g \in \mathcal{F}} \left\| (g(X_1), \dots, g(X_n))^\top - r \right\|_2$.
5: Set $f_{t+1} \leftarrow f_t + \gamma_{t+1} g_{t+1}$. (update).
6: **end for**

Output: f_T.

differentiable with respect to its second argument, gradient boosting just uses a subgradient in the place of the gradient $\nabla_n \mathcal{R}_n(f_t)$. This is, of course, convenient but as subgradients are not necessarily descent directions, they may damage the convergence of (sub-)gradient boosting. When dealing with non-differentiable convex losses, an optimizer normally has the reflex to resort to proximal algorithms. For this reason, we propose a new procedure for non-differentiable loss functions ℓ, which consists in adapting the proximal point algorithm (Nesterov, 2013) to functional optimization.

Proposal: proximal boosting methods The simple idea underlying the proximal boosting algorithm proposed in Fouillen et al. (2022), is to replace the update direction of the optimization variable in Algorithm 1 by a proximal direction (instead of a subgradient), which is[1]

$$\operatorname{Prox}_n^{\lambda_{t+1}} \mathcal{R}_n(f) := \frac{1}{\lambda_{t+1}} \Big((f(X_1), \dots f(X_n)) - \operatorname{prox}_{\lambda_{t+1} \mathcal{R}_n}(f(X_1), \dots f(X_n)) \Big),$$

where $\lambda_{t+1} > 0$ is a proximal step. Thus, proximal boosting computes the pseudo-residuals based on $\operatorname{Prox}_n^{\lambda_{t+1}} \mathcal{R}_n(f_t)$ instead of $\nabla_n \mathcal{R}_n(f_t)$ and leaves the rest unchanged. This new algorithm is very intuitive and proved to converge at the expected rate for differentiable loss functions, while obtaining a rate of convergence for non-differentiable loss functions remains an open question. To remedy this limitation, we have introduced a variant of this algorithm, named *residual proximal boosting* (see Algorithm 2) and inspired by Grubb and Bagnell (2011), which incorporates a mechanism making it possible to control the approximation error made at each iteration and to obtain a convergence rate under weak assumptions.

This algorithm can be shown to converge without requiring strong convexity or differentiability of the objective function, but relying on an edge hypothesis, quite classical in the boosting literature, see Grubb and Bagnell (2011). It characterizes the approximation capacity of the class $\mathcal{F}$, as follows: there exists $\zeta \in (0, 1]$ such that:

$$\forall r \in \mathbb{R}^n, \qquad \exists g \in \mathcal{F} : \|g(X_1^n) - r\|^2 \leq (1 - \zeta^2) \|r\|^2.$$

A set of weak learners $\mathcal{F}$ satisfying such an assumption is said to have edge ζ.

[1] Note that this corresponds to the update direction in a proximal point algorithm.

Algorithm 2 Residual proximal boosting.

Input: $\lambda_1, \ldots, \lambda_T > 0$ (proximal steps).

1: Set $f_0 \in \arg\min_{g \in \mathcal{F}} \mathcal{R}_n(g)$, $\Delta_0 \leftarrow 0$ (initialization).
2: **for** $t = 0$ **to** $T - 1$ **do**
3: Compute $r \leftarrow -\operatorname{Prox}_n^{\lambda_{t+1}} \mathcal{R}_n(f_t)$ (pseudo-residuals).
4: Compute $g_{t+1} \in \arg\min_{g \in \mathcal{F}} \left\| (g(X_1), \ldots, g(X_n))^\top - (r + \Delta_t) \right\|$.
5: Set $f_{t+1} \leftarrow f_t + \lambda_{t+1} g_{t+1}$.
6: Set $\Delta_{t+1} \leftarrow r + \Delta_t - g_{t+1}(X_1^n)$.
7: **end for**

Output: f_T.

Theorem 14 (Unformal)**.** *Under the edge hypothesis, assume that $\mathcal{R}_n$ is convex and L-Lipschitz continuous for some $L > 0$. Let $(f_t)_t$ be any sequence generated by Algorithm 2 and $f_{best} \in \arg\min_{1 \leq t \leq T} \mathcal{R}_n(f_t)$. Assume that there exists $f^\star \in \arg\min_{f \in \operatorname{span} \mathcal{F}} \mathcal{R}_n(f)$ and that $\|(f_t(X_1), \ldots, f_t(X_n))\| \leq R$ and $\|(f^\star(X_1), \ldots, f^\star(X_n))\| \leq R$ for some $R > 0$ and all t. Then, choosing $\lambda_t = \frac{1}{\sqrt{t}}$ leads to:*

$$C(f_{best}) - C(f^\star) \leq \frac{2R^2}{\sqrt{T}} + \frac{40G^2}{\zeta^4 \sqrt{T}} + \frac{2G^2}{\zeta^4 T^{\frac{3}{2}}}.$$

Since, apart from the boosting world, the convergence rate of the proximal point method for non-smooth functions is $O(1/t)$ (respectively $O(1/\sqrt{t})$ for the subgradient method), one may expect that proximal boosting converges in $O(1/t)$ (while subgradient boosting is in $O(1/\sqrt{t})$ (Grubb and Bagnell, 2011)) but the previous result states a worst case convergence rate in $O(1/\sqrt{t})$. The latter is in fact not that surprising: for L-smooth and κ-strongly convex objectives, gradient descent converges in $O\left(\left(1 - \frac{\kappa}{L}\right)^t\right)$ while gradient boosting converges in $O(1/t)$. This highlights that the approximation step used in boosting iterations is prone to damage the convergence rate.

Numerical experiments on synthetic data confirm the theoretical convergence results and show a significant impact of the newly introduced parameter λ. Correctly tuned, this parameter provides a noticeable improvement of proximal-based boosting over gradient-based boosting for non-differentiable loss function, from both the optimization and the statistical points of view.

As standard proximal methods can be accelerated, we also provide accelerated versions of the proximal algorithm and conduct a thorough numerical study on different real-world datasets. Incorporating Nesterov's acceleration to proximal boosting, as done with gradient boosting by Biau et al. (2019a) introduce instabilities in (both) algorithms, leading to divergence on the training and the test sets. Our experience is that accelerated boosting is very sensitive to hyperparameters and thus tricky to tune. Despite the fact that these procedures rarely provide good generalization results, accelerated proximal boosting seems to perform better than its gradient counterpart for non-differentiable losses, and early stopping helps to prevent overfitting.

2.1.2 Theoretical study of stochastic Newton's algorithm

Learning is often performed through the minimization of an (empirical) risk:

$$\min_{f_\theta \in \mathcal{F}} \mathcal{R}(f) := \mathbb{E}\left[\ell(Y, f_\theta(X))\right] \quad \text{or} \quad \min_{f \in \mathcal{F}} \mathcal{R}_n(f) := \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$$

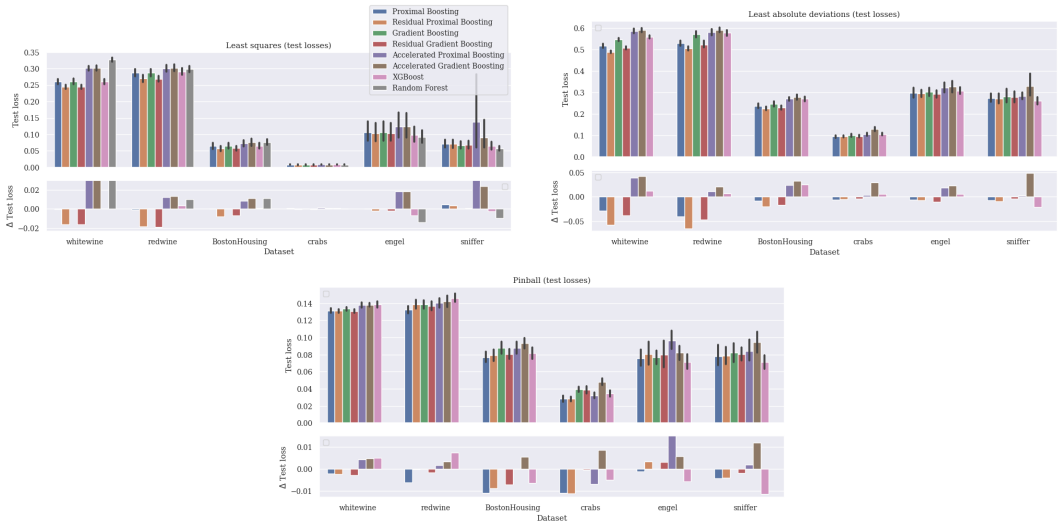


Figure 2.1: Losses on test datasets for different losses. Δ *Test loss* refers to the increment of the loss from that of gradient boosting. Proposed methods are depicted in blue, orange and green.

for some loss function ℓ, some generic pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, a training samples $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ i.i.d. copies of (X, Y), and a class $\mathcal{F}$ of parameterized functions $\{f_\theta; \theta \in \mathbb{R}^d\}$. First-order online algorithms have become hegemonic to tackle both problems: by a low computational cost per iteration (in the case of empirical risk minimization) or the ability to process online data (in the case of risk minimization), they allow performing machine learning tasks on large and real-world datasets; see, for instance, the review paper Bottou et al. (2018), including a unified view on both problems.

SGD, the ruler Stochastic gradient methods (SGD) and their averaged versions are shown to be theoretically asymptotically efficient (Polyak and Juditsky, 1992a; Pelletier, 2000), and more recent works focus on the non-asymptotic behavior of these estimates (Bach and Moulines, 2013; Godichon-Baggioni, 2019; Gadat and Panloup, 2022): more precisely, it was proven that, under mild assumptions, averaged estimates can converge at a rate of order $O(1/n)$ where we let n denote the size of the dataset (and the number of iterations as well, in a streaming setting). However, these first-order online algorithms can be shown in practice to be very sensitive to the Hessian structure of the risk they are supposed to minimize. To address this issue, first-order optimization strategies involving adaptive learning rates have been also considered in the literature, see Adagrad (Duchi et al., 2011) or Adadelta (Zeiler, 2012). In view of avoiding highly costly iterations, they mimic second-order algorithms by approximating the Hessian matrix by exclusively using gradient information or by assuming a diagonal structure of it (making its inversion much easier). If this type of algorithms have been incredibly popularized, they may still under-perform when the Hessian structure is complex. To address this issue, the attention may be deported towards stochastic second-order algorithms.

Stochastic Newton algorithms Stochastic second-order algorithms have been studied in the literature, essentially under two different prisms: on the one hand, Byrd et al. (2016) relies on limited-memory BFGS updates. Specifically, local curvature is captured through (subsampled) Hessian-vector products, instead of differences of gradients, so that the cost is close to the one of standard SGDs. This quasi-Newton method requires a good conditioning of the estimated Hessian inverses, and can be seen as a refinement of mini-batches gradient algorithms, which is not explicitly derived for online purposes. On the other hand, Leluc and Portier (2020) introduce a conditioned SGD based on a preconditioning of the gradient direction. The preconditioning matrix is typically an estimate of the inverse Hessian at the optimal point. The proposed conditioned SGD entails a full inversion of the estimated Hessian, requiring $O(d^3)$ operations per iteration in general, which is less compatible with large-scale data.

Contributions In Boyer and Godichon-Baggioni (2020), we consider a unified and general framework that includes various applications of machine learning tasks, for which we propose a stochastic weighted Newton algorithm in which an estimate of the Hessian is constructed and *easily* updated over iterations using genuine second-order information. Given a particular structure of the Hessian estimates that will be encountered in the most classical applications, this algorithm leverages from the possibility to directly update the inverse of the Hessian matrix at each iteration in $O(d^2)$ operations, generalizing a trick introduced in the context of logistic regression in Bercu et al. (2020).

Algorithms The stochastic Newton algorithm is defined by

$$\theta_{n+1} = \theta_n - \frac{1}{n+1}\overline{H}_n^{-1}\nabla_\theta \ell\left(Y_{n+1}, f_{\theta_n}(X_{n+1})\right)$$

where $\overline{H}_n$ is $\mathcal{F}_n = \sigma\left(X_1, Y_1, \ldots, X_n, Y_n\right)$ measurable. Its weighted version reads as

$$\begin{aligned}\tilde{\theta}_{n+1} &= \tilde{\theta}_n - \gamma_{n+1}\overline{S}_n^{-1}\nabla_\theta \ell\left(Y_{n+1}, f_{\theta_n}(X_{n+1})\right)\\ \theta_{n+1,\tau} &= \theta_{\tau,n} + \tau_{n+1}\left(\tilde{\theta}_{n+1} - \theta_{n,\tau}\right)\end{aligned}$$

where $\gamma_n = \frac{c_\gamma}{n^\alpha}$, with $c_\gamma > 0$, $\alpha \in (1/2, 1)$ and $\overline{S}_n$ is $\mathcal{F}_n = \sigma\left(X_1, Y_1, \ldots, X_n, Y_n\right)$ measurable. By choosing different sequences $(\tau_n)_n$, one can play more or less on the strength of the last iterates in the optimization. This strategy can be indeed motivated by limiting the effect of bad initialization of the algorithms. For instance, choosing $\tau_n = \frac{1}{n+1}$ leads to the "usual" averaging in stochastic algorithms. This can be generalized to other weighting, e.g. $\tau_n = \frac{\log(n+1)^\omega}{\sum_{k=0}^n \log(k+1)^\omega}$ with $\omega > 0$.

The choice of the weights $(\tau_n)_n$ are assumed to satisfy the conditions of Mokkadem and Pelletier (2006): there is $\tau > \max\{1/2, -1/2\}$ such that

$$n\left(1 - \frac{\tau_{n-1}}{\tau_n}\right) \xrightarrow[n\to+\infty]{} -1 \qquad \text{and} \qquad n\tau_n \xrightarrow[n\to\infty]{} \tau$$

For both averaging previously introduced, since $\tau = 1$, one has the following unified iterates

$$\theta_{n,1} = \frac{1}{\sum_{k=0}^n \log(k+1)^\omega}\sum_{k=0}^n \log(k+1)^\omega \tilde{\theta}_k, \tag{2.2}$$

where choosing $\omega = 0$ leads to the "usual" averaging technique.

Theoretical study The iterates of the weighted stochastic Newton algorithm are shown to asymptotically converge.

Theorem 15 (Unformal)**.** *Assuming that the convergence of Hessian estimates and their inverse is controlled in terms of the operator norm,*

$$\|\theta_{\tau,n} - \theta^\star\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s.$$

and

$$\sqrt{n}\left(\theta_{\tau,n} - \theta^\star\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2}{2\tau - 1} H^{-1}\Sigma H^{-1}\right)$$

with

$$H := \nabla^2_\theta \mathbb{E}\left[\ell(Y, f_{\theta^\star}(X))\right] \qquad \textit{and} \qquad \Sigma := \mathbb{E}\left[\nabla_\theta \ell\left(Y, f_{\theta^\star}(X)\right) \nabla_\theta \ell\left((Y, f_{\theta^\star}(X))\right)^\top\right].$$

Note that the averaging techniques mentioned above give

$$\sqrt{n}\left(\theta_{n,1} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

so that these estimates are asymptotically efficient for both considered averaging methods (the usual and the logarithmic ones). But the choice of weights remains crucial as it can damage the asymptotic variance. Indeed, using another non-uniform weighting of the form $\theta_{n,1+\omega} = \frac{1}{\sum_{k=0}^n (k+1)^\omega} \sum_{k=0}^n (k+1)^\omega \tilde{\theta}_k$ leads to

$$\sqrt{n}\left(\theta_{n,1+\omega} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{(1+\omega)^2}{2\omega + 1} H^{-1}\Sigma H^{-1}\right).$$

Implementation trick Stochastic Newton algorithms involves the inverse of the Hessian estimates at each iteration, requiring generally $O(d^3)$ operations. One can take advantage of a special form of the Hessian estimate $\overline{H}_n = (n+1)^{-1} H_n$:

$$H_n = H_0 + \sum_{k=1}^n u_k \Phi_k \Phi_k^\top, \tag{2.3}$$

with H_0 symmetric and positive, $u_k = u_k\left(X_k, \theta_{k-1}\right) \in \mathbb{R}$ and $\Phi_k = \Phi_k\left(X_k, \theta_{k-1}\right) \in \mathbb{R}^d$. Indeed, the inverse H_{n+1}^{-1} can be easily updated thanks to Riccati's formula (Duflo, 1997), i.e.

$$H_{n+1}^{-1} = H_n^{-1} - u_{n+1}\left(1 + u_{n+1}\Phi_{n+1}^T H_n^{-1}\Phi_{n+1}\right)^{-1} H_n^{-1}\Phi_{n+1}\Phi_{n+1}^T H_n^{-1}, \tag{2.4}$$

making the update costs $O(d^2)$. In such a case, one can consider the filtration generated by the sample again, i.e. $\mathcal{F}_n = \sigma\left(X_1, \ldots, X_n\right)$. The construction of these types of recursive estimates of the inverse of the Hessian can be made explicit in the cases of linear, logistic and softmax regressions.

Applications We perform numerical simulations in the case of linear, logistic and softmax models, showing the benefit of stochastic Newton iterates over standard (and averaged) SGD ones, in particular when the covariance structure is complex. We include in Figure 2.2 the results for the logistic regression.

☕ One should remark that the main limitation of this work is that it does not capture the benefit of stochastic Newton iteration, over stochastic gradient ones (which were known to be already asymptotically optimal), whereas the improvement is visible on practical experiments (provided the practitioner is willing to pay a computational cost in $O(d^2)$ instead of $O(d)$). This highlights that there is still avenue for deriving theoretical guarantees for stochastic Newton algorithms able to show improvement in non-asymptotic regimes.

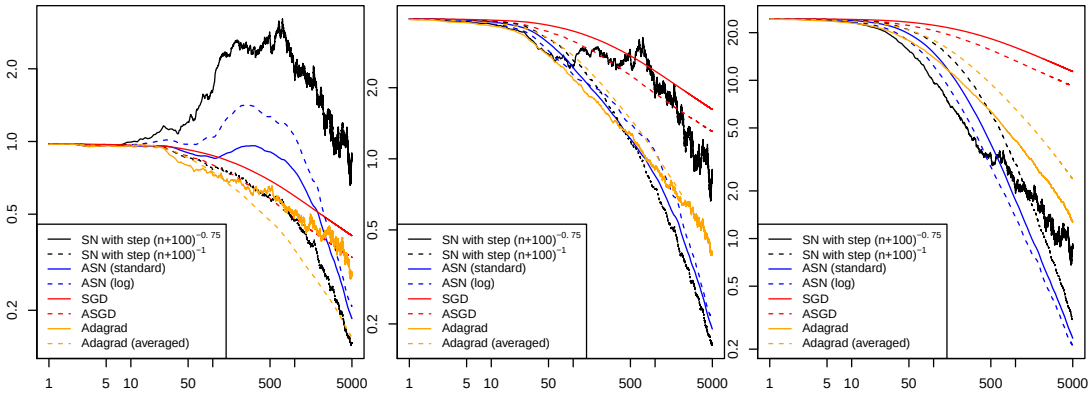


Figure 2.2: (Logistic regression with correlated Gaussian variables) Mean-squared error of the distance to the optimum θ with respect to the sample size for different initializations: $\theta_0 = \theta + r_0 U$, with $r_0 = 1$ (left), $r_0 = 2$ (middle) or $r_0 = 5$ (right).

2.2 Connections between tree-based methods and neural networks

Tree ensemble methods on the one hand, and neural networks on the other hand, have met great success in the machine learning community for their predictive performances. In this section, we explore in two different ways how one technique can benefit from the other to form more powerful models.

2.2.1 Analyzing the tree-layer structure of deep forests

What is Deep Forest? The Deep Forest (DF) algorithm (Zhou and Feng, 2019) has received a lot of attention in recent years in various applications ranging from hyperspectral image processing, medical imaging, drug interactions to fraud detection. It consists in a hybrid learning procedure in which random forests are used as elementary components of a neural network. More precisely, each layer of DF is composed of an assortment of Breiman's forests (Breiman, 2001) and Completely-Random Forests (CRF) (Fan et al., 2003), fed by the outputs of the previous layer and the initial input (entailing therefore residual connections), see Figure 2.3. Each module of this NN architecture being non-differentiable, each layer are trained in cascade, one after the other (without any backpropagation training). This results in an ever more complex forest architecture, and if the empirical evidence of its relevancy can be outlined on real-world datasets, the real asset of such procedure over standard forests still

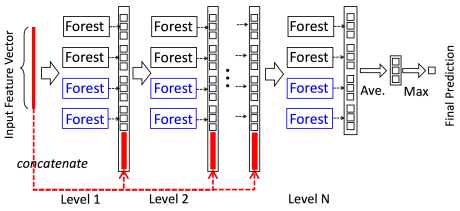


Figure 2.3: (from Zhou and Feng (2019)) Deep Forest architecture in which Breiman's forests (resp. CRF) are depicted in black (resp. blue).

remains theoretically unclear.

Numerical and theoretical analyses of DF As the performances of DF have already been validated by the literature (see Zhou and Feng, 2019), the main goals of our study in Arnould et al. (2021) are (i) to quantify the potential benefits of DF over RF, and (ii) to understand the mechanisms at work in such complex architectures.

Point (i) Deep Forests contain an important number of tuning parameters. Apart from the traditional parameters of random forests, DF architecture depends on the number of layers, the number of forests per layer, the type and proportion of random forests to use (Breiman or CRF). In Zhou and Feng (2019), the default configuration is set to 8 forests per layer, 4 CRF and 4 RF, 500 trees per forest (other forest parameters are set to `scikit-learn` (Pedregosa et al., 2011) default values), and layers are added until 3 consecutive layers do not show score improvement. We show in particular that much lighter configuration can be on par with DF default configuration, leading to a drastic reduction of the number of parameters in few cases. For most datasets, considering DF with two layers is already an improvement over the basic RF algorithm. However, the performance of the overall method is highly dependent on the structure of the first random forests, which leads to stability issues.

Point (ii) To theoretically understand the benefit of using cascades of tree-based methods instead of a single instance, we study a simplified version of the DF architecture. By establishing tight lower and upper bounds on the risk, we prove that a shallow tree-network may outperform an individual tree in the specific case of a well-structured dataset if the first encoding tree is rich enough. This is developed in the sequel as it is a first step to understand the interest of extracting features from trees, and more generally the benefit of tree networks.

We assume to have access to a dataset $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of i.i.d. copies of the generic pair (X, Y) with X living in $[0,1]^d$ and $Y \in \{0,1\}$ being the label associated to X. We consider a context of binary classification (in regression): let $r(x) = \mathbb{E}[Y|X = x]$ be the regression function and for any function f, its quadratic risk is defined as $R(f) = \mathbb{E}[(f(X) - r(X))^2]$, where the expectation is taken over $(X, Y, \mathcal{D}_n)$.

Given a decision tree, we denote by $L_n(X)$ the leaf of the tree containing X and $N_n(L_n(X))$ the number of data points falling into $L_n(X)$. The prediction of such a tree at point X is given by

$$\hat{r}_n(X) = \frac{1}{N_n(L_n(X))} \sum_{X_i \in L_n(X)} Y_i$$

with the convention $0/0 = 0$, i.e. the prediction for X in a leaf with no observations is arbitrarily set to zero.

Definition 16 (Shallow centered tree network)**.** *The shallow tree network consists in two trees in cascade:*

- ***(Encoding layer)*** *The first-layer tree is a* cycling centered tree of depth k. *It is built independently of the data by splitting recursively on each variable, at the center of the cells. The first cut is made along the first coordinate, the second along the second coordinate, etc. The tree construction is stopped when exactly k cuts have been made. For each point X, we extract the empirical mean $\bar{Y}_{L_n(X)}$ of the outputs Y_i falling into the leaf $L_n(X)$ and we pass the new feature $\bar{Y}_{L_n(X)}$ to the next layer, together with the original features X.*

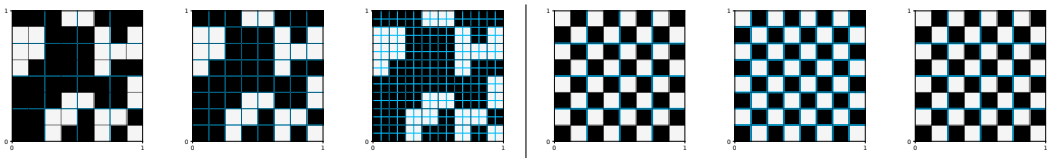


Figure 2.4: Arbitrary chessboard (left) or regular chessboard (right) data distribution for $k^\star = 6$ and 40 black cells (p is not displayed here). Partition of the (first) encoding tree of depth 4, 6, 8 (from left to right) is displayed in blue. The optimal depth of a single centered tree for this chessboard distribution is 6.

- ***(Output layer)*** *The second-layer tree is a* centered tree of depth k' *for which a cut can be performed at the center of a cell along a raw feature (as done by the encoding tree) or along the new feature $\bar{Y}_{L_n(X)}$. In this latter case, two cells corresponding to $\{\bar{Y}_{L_n(X)} < 1/2\}$ and $\{\bar{Y}_{L_n(X)} \geq 1/2\}$ are created.*

The resulting predictor composed of the two trees in cascade, of respective depth k and k', trained on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ is denoted by $\hat{r}_{k,k',n}$.

Note in particular that $\hat{r}_{k,0,n}(X)$ is the prediction given by the first encoding tree only and outputs, as a classical tree, the mean of the Y_i's falling into a leaf containing X.

We manage to exhibit a *very structured data* setting in which a tree network outperforms a single tree. The input data X is assumed to be uniformly distributed over $[0,1]^d$ and $Y \in \{0,1\}$. Let $k^\star$ be a multiple of d and let $p \in (1/2, 1]$. We build a regular partition of the space with *cells* $C_1, \ldots, C_{2^{k^\star}}$ of generic form

$$\prod_{k=1}^{d} \left[\frac{i_k}{2^{k^\star/d}}, \frac{i_k + 1}{2^{k^\star/d}} \right),$$

for $i_1, ..., i_d \in \{0, ..., 2^{k^\star/d} - 1\}$. We arbitrary assign a color (black or white) to each cell, which has a direct influence on the distribution of Y in the cell. More precisely, for x in a given cell C,

$$\mathbb{P}[Y = 1 | X = x] = \begin{cases} p & \text{if } C \text{ is a black cell,} \\ 1 - p & \text{if } C \text{ is a white one.} \end{cases} \tag{2.5}$$

This distribution corresponds to a *generalized chessboard* structure, see Figure 2.4.

The theoretical upper and lower bounds on the excess risk of one or two layer tree networks are summarized in Table 2.1.

This shows that there exists a benefit from using a tree network when the first-layer tree is deep enough. In this case, the risk of the shallow tree network is $O(1/n)$ whereas that of a single tree is $O(2^k/n)$. In presence of complex and highly structured data in the sense that one can identify similar distributions in different areas of the input space, the shallow tree network benefits from a variance reduction phenomenon by a factor 2^k. Overall, first-layer trees, and more generally the first layers in DF architecture, can indeed act as efficient data-driven encoders, helping to reduce the variance of the subsequent predictive layers.

	Consider the worst data setting for the cascade tree network (i.e. regular chessboard). Then, if the encoding is biased ($k < k^\star$) (shallow first tree)	Consider the general chessboard when the encoding is unbiased (over-optimal 1st tree depth, $k \geq k^\star$).
Risk of the first tree $\mathcal{R}(\hat{r}_{k,0,n})$	$\underbrace{\left(p - \frac{1}{2}\right)^2}_{\text{bias}} + \underbrace{O\left(\frac{2^k}{n}\right)}_{\text{variance}}$	$\Theta\left(\underbrace{\left(\frac{2^k p(1-p)}{n}\right)}_{\text{variance}} + C_1 \alpha_k^n\right)$
Risk of the second tree $\mathcal{R}(\hat{r}_{k,1,n})$ (after one cut)	$\geq \underbrace{\left(p - \frac{1}{2}\right)^2}_{\text{bias}}$	$\Theta\left(\underbrace{\frac{2p(1-p)}{n}}_{\text{variance}} + C_2 \beta_{k,p}^n\right)$

Table 2.1: Summary of the lower and upper bounds on a one or two-layer tree network depending on the depth of the first "encoding" tree.

2.2.2 New initialization technique for MLP learning

Neural networks (NN) are now widely used in many domains of machine learning, in particular when dealing with very structured data. They indeed provide state-of-the-art performances for applications with images or text. However, neural networks still perform poorly on tabular inputs, for which tree ensemble methods remain the gold standards (Grinsztajn et al., 2022). This section is dedicated on presenting a practical approach, from the original paper Lutz et al. (2022), which improves the performances of the former by using the strengths of the latter.

NN & tabular data In the absence of a suitable architecture for handling tabular data, the Multi-Layer Perceptron (MLP) architecture (Rumelhart et al., 1986) remains the obvious choice due to its generalist nature. Apart from the large number of parameters, one difficulty of MLP training arises from the non-convexity of the loss function (see, e.g., Sun, 2020). In such situations, the initialization of the network parameters (weights and biases) are of the utmost importance, since it can influence both the optimization stability and the quality of the minimum found. Typically, such initializations are drawn according to independent uniform distributions with a variance decreasing w.r.t. the size of the layer (He et al., 2015). In the end, how MLP can be used to handle tabular data remains unclear, especially since a corresponding prior in the MLP architecture adapted to the correlations of the input is not obvious, to say the least.

Very recently, specific NN architectures have been proposed to deal specifically with tabular data. For example, TabNet (Arik and Pfister, 2021) uses a sequential self-attention structure to detect relevant features and then applies several networks for prediction. SAINT (Somepalli et al., 2021), on the other hand, uses a two-dimensional attention structure (on both features and samples) organized in several layers to extract relevant information which is then fed to a classical

MLP. These methods typically require a large amount of data, since the self-attention layers and the output network involve numerous MLP.

Tree-based methods & NN Several solutions have been proposed to leverage the correspondence between tree-based methods and NN, in order to develop more efficient models for processing tabular data. For example, TabNN (Ke et al., 2018) first trains a GBDT on the available data, then extracts a group of features per individual tree, compresses the resulting groups, and uses a tailored Recursive Encoder based on the structure of these groups (with an initialization based on the tree leaves). Therefore, TabNN employs pre-trained tree-based methods to design more efficient NN.

Conversely, Sethi (1990), Brent (1991), and later Welbl (2014), Richmond et al. (2015) and Biau et al. (2019b) propose to translate decision trees into very specific MLP (made of 3 layers) and use GD training to improve upon the original tree-based method. Such procedures can be seen as a way to relax and generalize the partition geometry produced by trees and their aggregation. To our knowledge, such translations have not been used to boost the training of *general* NN architectures.

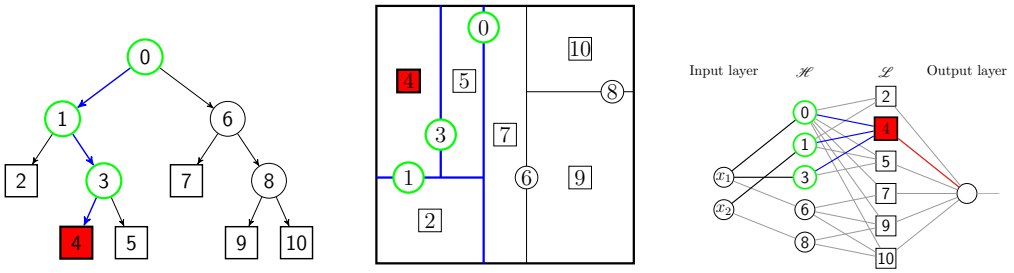


Figure 2.5: (from Biau et al. (2019b)) Illustration of a decision tree, its induced feature space partition and its corresponding MLP translation on a problem with 2 input variables. The activation functions involved in the network on the right are indicator ones.

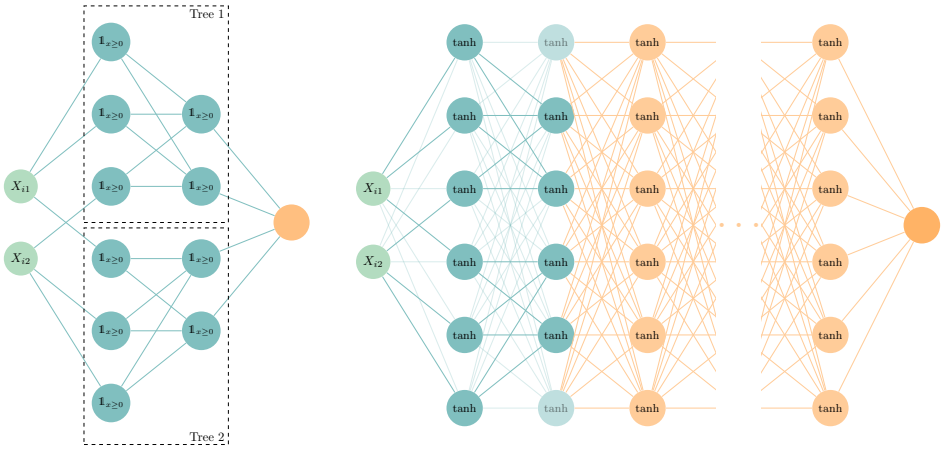
Proposal In the work Lutz et al. (2022), we propose a new method to initialize a potentially deep MLP for learning tasks with tabular data:

(i) First a tree-based predictor (boosting or random forests) is trained;

(ii) Secondly, the tree-based predictor is translated into a 3-layer NN architectures (following the procedure of Biau et al. (2019b)), see Figure 2.5, in which typical indicator functions involved in the tree prediction are relaxed by the use of tanh functions instead. Note that this relaxation may induce a performance loss for the predictors in play;

(iii) Finally, use this translation, to initialize the 2 first layers of an MLP (forget the last layer of the translation), the deeper ones being classically randomly initialized.

The whole procedure is represented in Figure 2.6. When the tree-based predictor used for initialization is from boosting (resp. a random forest), we talk about "GBDT init" (resp. "RF init") in the numerical results. Note that the boosting algorithm used in the sequel is XGBoost (Chen and Guestrin, 2016).



(a) (b)

Figure 2.6: Illustration of the initialization technique on an MLP with 2 inputs and 1 output. In (a), a pre-trained tree-based method composed of 2 trees is represented in a NN fashion involving indicator functions as activation functions. In (b), an MLP of arbitrary depth and involving tanh activation functions is represented at initialization: the weights of the first two layers are initialized using the information captured in (a) (note that all connections marked in transparent blue are initialized to 0). The weights of the subsequent layers are randomly initialized (orange edges).

Numerical results The proposed procedure is shown to outperform the widely used uniform initialization of MLP (default initialization in Pytorch Paszke et al., 2019), denoted "random init", as follows.

1. **Improved performances.** For tabular data, the predictive performances of the MLP after training are improved compared to MLP that use a random initialization. Our procedure also outperforms more complex deep learning procedures based on self-attention and is on par with classical tree-based methods (such as XGBoost Chen and Guestrin (2016)).

2. **Faster optimization.** The optimization following a tree-based initialization is boosted in the sense that it enjoys a faster convergence towards a (better) empirical minimum: a tree-based initialization results in faster training of the MLP.

3. **Even better in large width regimes.** The standard search spaces for the MLP width used in the literature usually involve a few hundred neurons per layer (e.g. up to 100 neurons in Borisov et al., 2021); yet, in this work, we consider MLP with a width up to 2048 neurons. Large MLP are actually very beneficial for tree-based initialization methods as they allow the use of more expressive tree-based models in the initialization step. We have conducted numerical experiments in order to compare the performance of an MLP with random/GDBT initializations and various widths. There is no gain in prediction by using wider (thus more complex) NN, when randomly initialized. However, an MLP initialized with GBDT significantly benefits from enlarging the NN width (justifying a fixed width of 2048 for tree-based initialized MLP). This confirms the idea that tree-based initialization helps revealing relevant

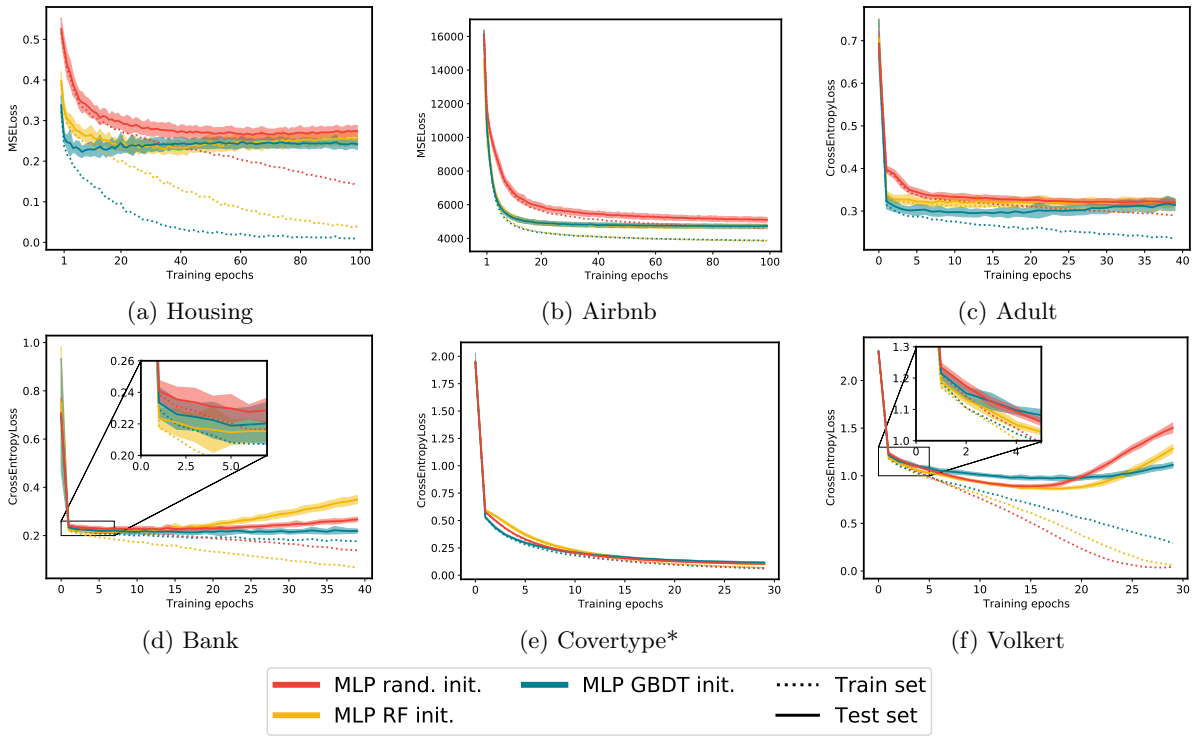


Figure 2.7: Optimization behaviour of randomly, RF and GBDT initialized MLP evaluated over a 5 times repeated (stratified) 5-fold of each data set, according to Protocol P1. The lines and shaded areas report the mean and standard deviation. *evaluation on a single 5-fold cross validation.

features to the MLP, all the more as the width increases, and by doing so, boosts the MLP performance after training.

4. **Preserved weight distribution after training.** Finally, we investigate the structure that tree-based initialization induces on the MLP *after* training, by looking at the distribution of the weights layer by layer before and after training. It indicates that the weight distribution on the first two layers change significantly during training when the MLP is randomly initialized: the weights are uniformly distributed at epoch 0 but appear to be Gaussian after training. When RF or GBDT initializers are used instead, the weights of the first two layers are sparsely distributed at epoch 0 by construction, and their distribution is preserved during training (notice the logarithmic y-axis for these plots in Figure 2.8).

☕While our procedure is quite generic, some restrictions are noticeable. First, our analysis only allows to initialize neural networks with tanh activation functions; removing this limitation by considering ReLU is a good avenue for future work. Therefore, our work highlights that SGD optimizer fails to find the best final weights starting from a random and unstructured initialization. Hence, the implicit regularization that operates with SGD in some specific theoretical framework such as diagonal network, see Woodworth et al. (2020), does not seem strong enough in practical

Model \ UCI data set	Housing	Airbnb	Diamonds	Adult	Bank	Blastchar	Heloc	Higgs	Covertype	Volkert
	MSE ↓	MSE ↓ (x10^3)	MSE ↓ (x10^{-3})	AUC ↑ (in %)	AUC ↑ (in %)	AUC ↑ (in %)	AUC ↑ (in %)	AUC ↑ (in %)	Acc. ↑ (in %)	Acc. ↑ (in %)
Random Forest	0.263±0.009	5.39±0.13	9.80±0.35	91.6±0.3	92.8±0.3	**84.5±1.2**	91.3±0.6	80.4±0.1	83.6±0.1	64.2±0.3
GBDT	**0.208±0.010**	**4.71±0.15**	**7.38±0.28**	**92.7±0.3**	**93.3±0.3**	**84.7±1.0**	**92.1±0.4**	82.8±0.1	**97.0±0.0**	71.3±0.4
Deep Forest	0.225±0.008	**4.68±0.16**	8.23±0.29	91.8±0.3	92.9±0.2	**83.7±1.2**	90.3±0.5	81.2±0.0*	92.4±0.1*	66.3±0.4
MLP rand. init.	0.258±0.011	5.07±0.16	15.5±12.5	90.5±0.4	91.0±0.3	81.4±1.2	80.1±0.1	83.2±0.3	96.7±0.0	72.2±0.4
MLP Xavier init.	0.263±0.012	5.05±0.17	12.4±6.19	90.5±0.5	90.8±0.5	81.7±1.3	79.9±1.1			72.1±0.4
MLP LUSV init.	0.295±0.018	4.99±0.14	14.1±5.00	90.5±0.5	90.2±0.5	**84.3±1.2**	79.9±0.9			70.8±0.5
MLP WT prun.	0.248±0.011	5.26±2.11	9.83±5.07	90.6±0.4	90.9±0.5	**84.4±1.2**	89.6±0.7	82.9±0.1	**97.0±0.0**	71.5±0.4
MLP RF init.	0.222±0.009	**4.66±0.16**	7.93±0.22	92.1±0.3	92.4±0.4	**84.4±1.2**	**91.7±0.4**	**83.6±0.1**	96.7±0.0	**74.1±0.4**
MLP GBDT init.	**0.206±0.007**	**4.70±0.09**	8.15±0.35	92.2±0.3	92.5±0.3	**84.6±1.2**	91.5±0.6	83.0±0.0	96.2±0.0	73.5±0.5
MLP DF init.	0.234±0.016	**4.81±0.13**	8.28±0.24	91.9±0.4	92.2±0.3	**84.2±1.0**	91.4±0.6	83.3±0.1*	94.5±0.3*	71.3±0.5
SAINT	0.258±0.011	**4.81±0.15**	17.7±3.83	91.6±0.3	92.2±0.4	**84.0±0.8**	90.2±0.7	**83.7±0.1***	96.6±0.1*	70.1±0.4

Table 2.2: Best mean scores and standard deviations based on a 5 times repeated (stratified) 5-fold cross validation. For each data set, predictors performing at least as well as the best over all (resp. best DL) score up to its standard deviation are highlighted in **bold** (resp. underlined). For each model, HP have been chosen via the "optuna" library with 100 iterations.

NN architectures to lead to a good optimum, without a specific (sparse) weight initialization. The fact that the weight sparsity is preserved through training could potentially result from a more complex SGD regularization, which will be definitely interesting to study in future works.

2.3 Interpolating regimes in random forests

Random Forests (RF, Breiman, 2001) have proven to be very efficient algorithms, especially on tabular data sets as previously highlighted. As any machine learning (ML) algorithm, Random Forests and Decision Trees have been analyzed and used according to the bias-variance trade-off. Regularization parameters have been introduced in order to control the variance while still reducing the bias. For instance, one can increase the variety of the constructed trees (by playing either with bootstrap samples or feature subsampling) or control the tree structure (by limiting either the number of points falling within each leaf or the maximum depth of all trees).

A statistical crisis In this line, statistical wisdom suggests that very complex models, interpolating training data, will be poor at predicting unseen examples. However, this paradigm has been completely questioned in recent years, by the identification of benign overfitting regimes: in particular, deeper and larger neural networks, achieving a zero training error, still empirically exhibit high predictive performances (Goodfellow et al., 2016).

Regarding parametric methods, benign overfitting has been exhibited and well understood in linear regression (Bartlett et al., 2020; Tsigler and Bartlett, 2020; Liang et al., 2020). Many researchers currently study the *implicit bias* or *implicit regularization* of stochastic gradient (SGD) strategies used during neural network training: the optimization of an over-parametrized one-hidden-layer neural network via SGD will converge to a minimum of minimal norm with good generalization properties in a regression setting (Bach and Chizat, 2021), or with maximal margin in a classification setting (Chizat and Bach, 2020).

Regarding non-parametric methods, practitioners have noticed the good performances of high-depth RFs for a long time (by default, several ML libraries such as the popular scikit-learn grow trees until pure leaves are reached). More recently, the use of interpolating (or very deep) trees for

boosting and bagging methods has been advocated in Wyner et al. (2017). Indeed, Wyner et al. (2017) believe that the *self-averaging* process at hand in RF (or in boosting methods) also produces an implicit regularization that prevents the interpolating algorithm from overfitting. Note that the regularization properties of RF have also been studied in the light of their complexity (Buschjäger and Morik, 2021) and tree depth (Zhou and Mentch, 2021). This phenomenon can be put in parallel with the results proved in Devroye et al. (1998); Belkin et al. (2019) where they show that an interpolating kernel method using a singular kernel (similar to $K(x) = ||x||^{-\alpha}\mathbb{1}_{||x||\leq 1}$) is consistent, reaching minimax convergence rate for β-Hölder regular functions.

In the specific case of random forests, while it is widely known that fully-grown decision trees interpolate and, in turn, have bad predictive performances, the statistical properties of interpolating RF have yet to be analyzed.

Study of the consistency of interpolating RF In Arnould et al. (2022), we study the trade-off between interpolation and consistency in the context of RF regression. Contrary to Nadaraya-Watson methods involving singular kernels that interpolate regardless of the bandwidth parameter, note that RF interpolate only for a specific choice of the depth, thus restricting the regime in which interpolation and consistency may occur in concordance.

(Analysis for centered RF) Centered RF have been historically studied for theoretical purposes as simplified versions of Breiman RF (Breiman, 2001) that are widely used in practice. These RF involve centered trees, in which cuts are always performed in the middle of current cells, therefore independently of the data in play. Theoretically, we prove that interpolation regimes and consistency cannot be achieved simultaneously for such non-adaptive centered RF. Note that we can only study their non-consistency in a "mean" interpolating regime, i.e. when each tree of the RF counts only one training data point per cell in expectation. In such a case, the major problems arise from empty cells in tree partitions (likely to appear as the cuts are performed independently of the data). Therefore, we study two versions of centered-type RF, called void-free CRF and kernel RF (KeRF), for which the aggregation rule is changed in order to avoid taking into account empty cells. More precisely, void-free CRF do not take into account empty cells and KeRF are built by averaging over all connected data points. These methods are proved to be consistent for larger tree depths, respectively in a noiseless and in a general noisy setting, but still for a "mean" interpolation regime only.

(Analysis for median RF) The Median RF, studied e.g. in Duroux and Scornet (2018), is composed of median trees which first randomly choose the direction to cut over and then cut at the median of the data points contained in the current cells. In order to avoid points falling on a cell boundary, whenever the number of points n_c in the cell is odd, the cut is made at the quantile $(n_c + 1)/2n_c$. This type of forests therefore inherently avoid the presence of empty cells. For the first time, we manage to prove the consistency of an infinite interpolating median RF, i.e. involving an infinite number of fully-grown median trees.

Theorem 17 (Upper bound on the risk of the median forest)**.** *Consider a generic pair* $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ *of random variables such that* $Y = f^\star(X) + \varepsilon$*, where* $||\partial_\ell f^\star||_\infty^2$ *exists for all* $\ell \in \{1, \dots, d\}$*,* X *is uniformly distributed on* $[0, 1]^d$ *and the noise* ε *satisfies* $\mathbb{E}[\varepsilon|X] = 0$ *almost surely, and* $\mathbb{V}[\varepsilon] = \sigma^2$*.*

Consider $n \geq 16$ *i.i.d. observations, where* n *is a power of two, distributed as the generic pair*

(X, Y). *Then, the risk of the infinite median forest* $f_{\infty,n}^{\mathrm{MedRF}}$ *built on this data set satisfies*

$$\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{MedRF}}(X) - f^\star(X)\right)^2\right] \leq C_1 d \left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty^2\right) \left(1 - \frac{3}{4d}\right)^{\log_2 n} + \sigma^2 C_{2,d} (\log_2 n)^{-(d-1)/2}. \tag{2.6}$$

with $C_1 \leq 256 \exp\left(\frac{55+\sqrt{5}}{2-\sqrt{2}}\right)$ *and* $C_{2,d} = \left(32 \ \exp\left(\frac{10}{\sqrt{2}-1}\right)\right)^d d^{d/2}$.

In particular, the infinite median forest is consistent, that is

$$\lim_{n\to\infty} \mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{MedRF}}(X) - f^\star(X)\right)^2\right] = 0. \tag{2.7}$$

The proof essentially relies on controlling the volume of the resulting cells. This is the first result showing the consistency of complex predictors relying on fully-grown trees. In previous works, the tree depth was indeed always carefully chosen to comply with the traditional "bias-variance" compromise, and thus preventing the choice of high depths. If bootstrap is often considered as a way to prevent overfitting when dealing with interpolators, here we show that the key randomization mechanism at work in RF is sufficient on its own to reach consistency in spite of interpolation (and without the help of bootstrapping). Finally, it is worth noting that if an overfitting regime is benign for the consistency of Median RF, it seems to be nonetheless malignant for the convergence rate.

Remark 18 (Bootstrap or no bootstrap?)**.** *In our theoretical analysis, by switching the RF bootstrap off, we manipulate trees and forests interpolating the entire training dataset, i.e. achieving a zero training error (complying with interpolation considered in a NN context). We therefore focus only on the diversification of the trees to retrieve consistency despite interpolation. By turning the bootstrap on, the interpolation mode could be relaxed: each fully-grown tree would interpolate p% of the training data (but the resulting infinite forest will not interpolate any datapoint with probability 1). Bootstrap can actually help turning interpolators into consistent estimates, see for instance Biau and Devroye (2015) for the nearest neighbour predictor. Zhou and Mentch (2021) consider bootstrap for interpolating RF in classification (which radically differs from regression). They observe that with bootstrap on, the probability of interpolation tends to zero with n. This is actually corroborated in the case of regression with median RF: in Scornet (2015), the consistency of median RF with fully grown trees is established provided that the size* $a_n \in \{1, \ldots, n\}$ *of the bootstrap samples is such that* $a_n \to \infty$ *and* $a_n/n \to 0$*. Therefore, bootstrap can preserve the RF consistency by counteracting tree interpolation: as the sample size n grows, the fully-grown trees are less and less interpolating the training data (since the fraction* a_n/n *of samples seen by each tree tends to 0).*

(Analysis for Breiman RF) Numerical experiments show that Breiman RF are consistent when exactly interpolating, i.e. when the whole data set is used to build each fully-grown tree. Breiman forests rely on CART trees (classification and regression trees) which construction is driven by sophisticated mechanisms. On the theoretical side, it remains very challenging to bound the risk of interpolating Breiman RF. We control instead the volume of the interpolation zone for an infinite Breiman RF (without bootstrap).

Proposition 19. *Consider an infinite Breiman forest constructed without bootstrap. Suppose that for a given configuration of the training data, all cuts have a probability strictly greater than 0 to appear. Define the interpolation zone of a tree-based predictor as the area of the input space on*

which the prediction relies only on one training data point. Let $\mathcal{A}_{min}$ be the area of the input space corresponding to the intersection of the interpolation zones of all possible CART trees. Then, the volume of the minimal interpolation zone verifies

$$\mathbb{E}\left[\mu(\mathcal{A}_{min})\right] \leq \frac{1}{n^{d-1}}\left(1-2^{-n}\right)^{d}.$$

This volume tends to 0 at a polynomial rate in the number of samples n and an exponential rate in the dimension d. This supports the idea that the decay of the interpolation volume could be fast enough to retrieve consistency despite interpolation. More specifically, this is a *necessary* condition for consistency: the volume of the area where the prediction involves only a finite number of points (*a fortiori* the interpolation zone) should tend to 0. Indeed, it is not possible to average the noise of the training dataset with only a finite number of points. Therefore the noise in this area is of order σ^2. Proposition 19 portends the predominant *self-averaging* property of adaptive RF, and hence underpins the idea of good capabilities of Breiman RF in interpolation regimes.

☕ For semi-adaptive forests such as median RF, increasing the tree depth towards the interpolation regime results in a reduced bias and the variance reduction phenomenon only results from the split randomization effect. The higher the dimension, the more diversified the trees, the stronger the averaging effect and the variance reduction should be. This remains to be understood for the complex Breiman RF. A prelude could be to study the consistency of Breiman RF in an asymptotic dimensional regime. The idea would be to help the RF benefit from an increase of the dimension by improving its averaging effect and helping reduce the variance. Of course, in such a setting, the variance is only one part of the story, and a control on the bias becomes a real hindrance (as the approximation error may explode), unless extra model assumptions are formulated to bound the bias (if not to make it tend to 0). This would echo in particular the behavior of ridgeless least squares estimator in modern interpolation regimes, see Hastie et al. (2022).

☕ I want also to step back to discuss about the practical success of interpolating and powerful predictors such as RF but above all, neural networks. The recent enthusiasm for theoretically analyzing the behaviour of interpolators (specially in a NN context) brought the statistical community face to face with its own limitations: the statistical viewpoint related to the "bias-variance dilemma" had become the dominant paradigm of supervised machine learning, stimulating a mathematical research in this direction and increasingly sophisticated research on uniform theoretical bounds (involving VC dimension, Rademacher complexities, etc.). However, the latter remain insufficient to explain the good predictive behavior of ever more complex methods. This echoes in some way the rupture encountered in the compressed sensing community when realizing the need for non-uniform results. It is important to recognize that the good generalization of neural networks is a proven fact in practice and is maintained across a variety of architectures, optimization methods and data sets. The ability of neural networks to generalize to new data reveals a fundamental interplay between the underlying mathematical structures of the hypothesis class involved, the learning (and thus the optimization algorithms used), and the nature of the data. In this respect, bridging the gap between optimization and the statistical power of the constructed predictors is a major challenge, which is in my opinion today the central and extremely delicate issue in this highly competitive field. To this end, I would like to dedicate research efforts towards a unified theory of learning that combines optimization and statistics, which would be able to reflect the observed practical success of learning methods in the established theoretical bounds.

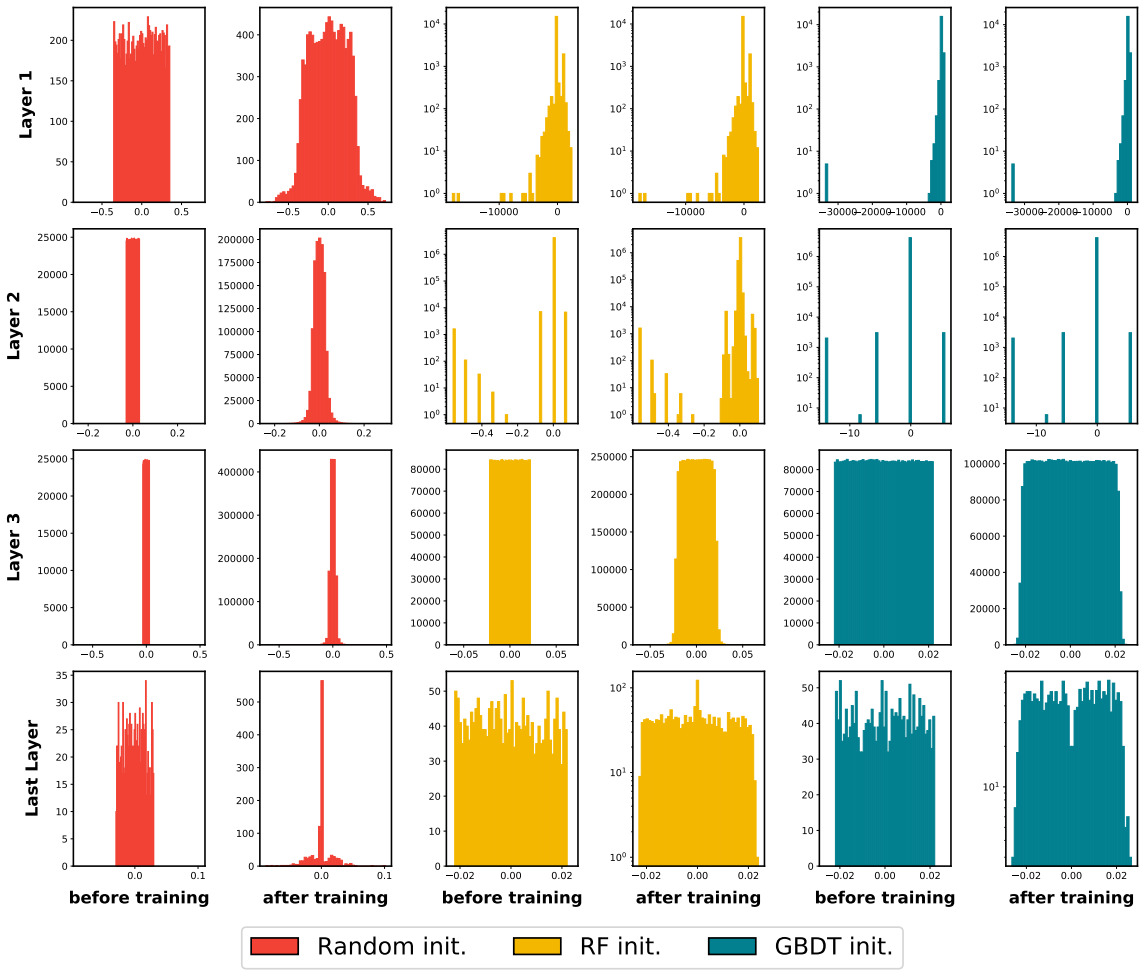


Figure 2.8: Histograms of the first three and last layers' weights before and after the MLP training on Housing. Comparison between random, RF and GBDT initializations.

Chapter 3

Handling missing values in statistical learning

Contents

The increasing availability of datasets and the multiplication of sources offer hopes for understanding, interpreting and predicting many phenomena. However one of the ironies of the so-called "big data" era is that missing data are unavoidable: the more data there are, the more missing data there are. Indeed, missing data can occur for many reasons: unanswered questions in a survey, lost data, sensing machines that fail, aggregation of multiple sources, etc. Classical statistical or machine learning methods cannot be directly applied on the datasets which contain missing values.

A naive solution is then to simply discard the missing values: by removing either the incomplete variables or the missing individuals, i.e. either some columns or some rows of the dataset. However, deleting data is not a good answer in most cases for two main reasons:

(i) this may result in erasing a large proportion of the initial data (think about a dataset in which each variable can be missing with probability 0.05, the proportion of complete individuals in expectation is roughly 60% in dimension 10, but 0,5% in dimension 100 !);

(ii) the complete observations can be not necessarily representative of the overall population leading to bias in subsequent analyses (think about a poll in which rich people would be less incline to reveal their income, the complete individuals would be therefore circumscribed to low salaries only).

Therefore two strategies can be envisaged when ML encounters missing data. The first one consists in imputing the missing values to form a completed dataset on which standard ML algorithms can be subsequently applied. This is an appealing solution, to which much effort has been devoted.

However, it must be conceded that this may introduce a bias in the imputed data and inevitably reduce the variability of the data. The second approach entails new statistical methods, specifically designed for a particular learning task or estimation problem to handle missing values.

I have contributed to both aspects in the recent years. Julie Josse introduced me to this problem in 2018, which was actually closely related to the matrix completion problems that I previously encountered in the compressed sensing community, but with radically different tools from optimization instead of statistics. We have co-advised the PhD. work of Aude Sportisse on these topics, with dedicated efforts in the case where the missing mechanism is said to be informative. We were also keen to illustrate and apply our methods on the Traumabase® dataset, which is a medical register fed from 19 French trauma centers and which counts clinical measurements (250 variables) on 20,000 traumatized patients from the scene of the accident to the hospital admission. This is an opportunity to pay tribute to the extraordinary work of its founder Sophie Hamada (†). Since 2021, I also co-advise the PhD. thesis of Alexis Ayme with Erwan Scornet and Aymeric Dieuleveut on related themes.

In order to present most of my works in this field, I have made some editorial choices: I decided to go through one of the most canonical models of statistics, the linear model. This seemingly simple setting yet remains unresolved in the case of missing input entries. Partial answers can be formulated but they actually depend on the question at stake.

Outline In Section 3.1, as a prelude to linear regression, I present very briefly three different manners to impute missing entries of the design matrix. The first two rely on low-rank assumptions and are specifically designed to handle "informative" missing values, they correspond to the works Sportisse et al. (2020c), Sportisse et al. (2020b). The last one, drawn from Muzellec et al. (2020), is a heuristics based on optimal transport tools proving to be a versatile imputation method. In Section 3.2, I study how the estimation of the model parameter can be performed with data arriving in a streaming fashion, and with the help of an algorithm that has become essential in the practice of data science, the stochastic gradient algorithm. This section gathers results from Sportisse et al. (2020a). Finally in Section 3.3, I discuss theoretical considerations henceforth focusing on the task of predicting the output using a linear model when the inputs may be missing.

The following works will then not be addressed: Descloux et al. (2022), Sportisse et al. (2021).

3.1 Imputation as a preprocessing step

3.1.1 Low-rank models for informative missing data

The low-rank model has become very popular in recent years (Kishore Kumar and Schneider, 2017) and it plays a key role in many scientific and engineering tasks, including denoising (Gavish and Donoho, 2017), collaborative filtering (Yang et al., 2018), genome-wide studies (Leek and Storey, 2007; Price et al., 2006), and functional magnetic resonance imaging (Candès et al., 2013). It is also a very powerful solution for dealing with missing values (Josse et al., 2016; Kallus et al., 2018). Indeed, the low-rank assumption can be considered as an accurate approximation for many matrices as detailed by Udell and Townsend (2017). For instance, the low-rank approximation makes sense when either, one can consider that a limited number of individual profiles exist or, dependencies between variables can be established.

Zoology of missing mechanisms The missing-data mechanism is said (i) missing completely at random (MCAR) when the occurrence of the missing data is totally independent of the data, (ii) missing at random (MAR) when the unavailability of the data depends on the values of observed variables and (iii) missing not at random (MNAR) when the process that causes the missing data depends on the values of missing variable, and possibly observed ones too. This last case is often qualified as "informative" as the fact that the variable is missing may give information about its value. MCAR and MAR settings are usually handled more easily than the challenging MNAR case. In the literature, a mechanism derived from the general MNAR scenario is often considered (Mohan, 2018), when the unavailability of a missing variable $X_{.j}$ only depends on the values of $X_{.j}$ themselves. It is the so-called *self-masked* MNAR mechanism.

Low rank model and fixed effects Our work Sportisse et al. (2020c) focuses on imputing a data set containing MNAR values, under a prior of a low-rank model with fixed effects. More precisely, the data matrix $X \in \mathbb{R}^{n\times d}$ is considered as the sum of an underlying low-rank matrix $\Theta \in \mathbb{R}^{n\times d}$ corrupted by an additive Gaussian noise:

$$X = \Theta + \epsilon, \text{where} \begin{cases} \Theta \text{ with rank } r < \min\{n, p\}, \\ \epsilon_{i.} \overset{\perp}{\sim} \mathcal{N}(0_n, \sigma^2 \mathrm{I}_{n\times n}), \forall i \in \{1, \ldots, n\}, \end{cases}$$

and to contain MNAR values. The aim is to estimate Θ and to use it in turn, to impute missing values in X. Most of the existing methods previously reviewed do not consider the case of MNAR data.

Our contribution is two-fold. We first propose to maximize the joint distribution of the data and the missing-data pattern using an EM algorithm. The missing-data pattern is modeled with the selection models specification and a self-masked mechanism is assumed. As the E-step has no closed form, a Monte Carlo approximation is performed and coupled with the Sampling Importance Resampling (SIR) algorithm (Gordon et al., 1993). The M-step is penalized by the nuclear norm (as a convex relaxation of the rank penalty), i.e.

$$\Theta^{r+1}, \phi^{r+1} \in \underset{\Theta,\phi}{\operatorname{argmax}}\, Q(\Theta, \phi; \theta^r, \phi^r) + \lambda \|\Theta\|_\star,$$

and solved by using an accelerated proximal gradient algorithm, called Fast Iterative Soft-Thresholding Algorithm (FISTA) (Beck and Teboulle, 2009), which converges faster than the traditional non-accelerated version ISTA. However, the whole method can be computationally costly and relies on the specification of the missing-data mechanism. The second contribution is to suggest an efficient surrogate estimation, without specifying the missing-data mechanism, by concatenating the data matrix and the missing-data mask $\Omega \in \{0, 1\}^{n\times d}$ (coding for the occurrence of missing values in X) as $X^{\text{aug}} = [X, (1 - \Omega)]$. A low-rank structure on this new matrix is assumed in order to take into account the relationship between the variables and the mechanism. The optimization can thus be performed as if the data were M(C)AR, because we assume that the information of the missing-data mechanism is already encoded in $(\mathbf{1}_{n\times d} - M)$. For this, we use the algorithm in Robin et al. (2020) which deals with mixed data, as X is assumed to contain continuous variables, and as $(\mathbf{1}_{n\times d} - \Omega)$ is a binary matrix.

Through a study on synthetic data, the model-based method proves to be extremely relevant when few variables are missing and the implicit method, which models the mask using a binomial distribution, is much less costly in terms of computation time and allows a better imputation. The

performances of our methods are assessed on the Traumabase dataset, when the aim is to complete the data before using it to predict if the doctors should administrate tranexomic acid to patients with traumatic brain injury, that would limit excessive bleeding.

☕ With a lot more perspective on the issue, I would not attack the problem in the same way today. First, a low-rank model with random effects would allow to study identifiability issues depending on the missing mechanism in play. Then, an algorithm solving the initial problem efficiently when many input variables may be missing with different MNAR mechanisms even in the simple setting of Gaussian data has yet to be developed. This is subject to active discussions with Kimia Nadjahi, Marine Le Morvan & Julie Josse.

A low-rank model with random effects (PPCA) for MNAR data In Sportisse et al. (2020b), the data matrix is considered to be generated under a *fully-connected* probabilistic principal component analysis (PPCA) model, in the following sense

$$X = \mathbf{1}\alpha + WB + \epsilon, \text{with} \begin{cases} W = (W_{1.}|\dots|W_{n.})^\top, \text{ with } W_{i.} \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0_r, \mathrm{Id}_{r\times r}) \in \mathbb{R}^r, \\ B \text{ of rank } r < \min\{n, d\}, \\ \alpha \in \mathbb{R}^d \text{ and } \mathbf{1} = (1\dots 1)^\top \in \mathbb{R}^n, \\ \epsilon = (\epsilon_{1.}|\dots|\epsilon_{n.})^\top, \text{ with } \epsilon_{i.} \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0_d, \sigma^2 \mathrm{I}_{d\times d}) \in \mathbb{R}^d, \end{cases}$$

where σ^2 and r are known, and so that the loading matrix B are of full rank. This model implies that the rows of X are independent and Gaussian with mean α and covariance matrix $\Sigma = B^\top B + \sigma^2 \mathrm{I}_{d\times d}$. From a theoretical point of view, we first discuss and prove the identifiability of the parameters of the PPCA and of the missing-data mechanism, by assuming a self-masked MNAR mechanism.

Then, in presence of (general) MNAR values, we propose a strategy to estimate the coefficients matrix B based on estimations of the mean and the covariance matrix. We show that they can be consistently estimated in the available-case analysis when only the observed cases are used and when the noise is assumed to tend to 0. In order to derive such estimators, we leverage linear connections that can be established between variables under the fully-connected PPCA assumption. Two strategies to derive mean and covariance estimators are suggested: by using algebraic arguments or graphical models. The latter is inspired by Mohan et al. (2018), which considers linear models with a self-masked mechanism.

This method has the great advantage of being specification-free for the missing-data mechanism and of dealing with MNAR data, possibly coupled with M(C)AR data, resulting in a realistic missing scenario. This comes at a high computational cost for estimation, and by requiring enough complete data available. To assess the proposed methodology, experiments are conducted on synthetic data and on two real datasets including the Traumabase dataset and a recommendation system dataset.

3.1.2 Imputation using optimal transport tools

Imputation methods, which consist in filling missing entries with plausible values, are very appealing as they allow to both get a guess for the missing entries as well as to perform (with care) downstream machine learning methods on the completed data. Efficient methods include, among others, methods based on low-rank assumptions (Hastie et al., 2015), as previously discussed, iterative random forests (Stekhoven and Bühlmann, 2012) and imputation using variational autoencoders (Mattei and Frellsen, 2019; Ivanov et al., 2018). A desirable property for imputation methods is that

they should preserve the joint and marginal distributions of the data. Non-parametric Bayesian strategies (Murray and Reiter, 2016) or recent approaches based on generative adversarial networks (Yoon et al., 2018) are attempts in this direction. However, they can be quite cumbersome to implement in practice.

Proposal We argue in Muzellec et al. (2020) that the optimal transport (OT) toolbox constitutes a natural, sound and straightforward alternative. Indeed, optimal transport provides geometrically meaningful distances to compare discrete (empirical) distributions, and therefore data. This work focuses on an application of entropic optimal transport to missing data imputation based on the intuitive fact that two random subsets of data (batches) drawn from the same initial dataset should have similar distributions.

More particularly, we propose an imputation criterion whose loss function uses the (differentiable) Sinkhorn divergence. Indeed, entropic regularization, in addition to being an efficient method to make optimal transport more computationally tractable in a discrete setting, has also proven to be a way to improve the unfavorable sampling complexity of OT.

A fast introduction to Sinkhorn divergences. Let $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$, $\beta = \sum_{i=1}^{n'} b_i \delta_{y_i}$ be two discrete distributions, described by their supports $(x_i)_{i=1}^{n} \in \mathbb{R}^{n \times p}$ and $(y_i)_{i=1}^{n'} \in \mathbb{R}^{n' \times p}$ and weight vectors $a \in \Delta_n$ and $b \in \Delta_{n'}$. Optimal transport compares α and β by considering the most efficient way of transporting the masses a and b onto each-other, according to a ground cost between the supports. The (2-)Wasserstein distance corresponds to the case where this ground cost is quadratic:

$$W_2^2(\alpha, \beta) := \min_{P \in U(a,b)} \langle P, M \rangle \,, \tag{3.1}$$

where $U(a,b) := \{P \in \mathbb{R}^{n \times n'} : P\mathbb{1}_{n'} = a, P^\top \mathbb{1}_n = b\}$ is the set of transportation plans, and $M = \left(\|x_i - y_j\|^2\right)_{ij} \in \mathbb{R}^{n \times n'}$ is the matrix of pairwise squared distances between the support points. W_2 is not differentiable and requires solving a costly linear program via network simplex methods (Peyré et al., 2019, §3). Entropic regularization alleviates both issues: consider

$$\mathrm{OT}_\varepsilon(\alpha, \beta) := \min_{P \in U(a,b)} \langle P, M \rangle + \varepsilon h(P), \tag{3.2}$$

where $\varepsilon > 0$ and $h(P) := \sum_{ij} p_{ij} \log p_{ij}$ is the negative entropy. Then, $\mathrm{OT}_\varepsilon(\alpha, \beta)$ is differentiable and can be solved using Sinkhorn iterations (Cuturi, 2013). However, due to the entropy term, OT_ε is no longer positive. This issue is solved through debiasing, by subtracting auto-correlation terms. Let

$$S_\varepsilon(\alpha, \beta) := \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon(\alpha, \alpha) + \mathrm{OT}_\varepsilon(\beta, \beta)). \tag{3.3}$$

Equation (3.3) defines the Sinkhorn divergences (Genevay et al., 2018), which are positive, convex, and can be computed with little additional cost compared to entropic OT (Feydy et al., 2019). Sinkhorn divergences hence provide a differentiable and tractable proxy for Wasserstein distances, and will be used in the following.

OT gradient-based methods Not only are the OT metrics described above good measures of distributional closeness, they are also well-adapted to gradient-based imputation methods. Indeed, let $X_{K:}, X_{L:}$ be two batches drawn from X, for K and L two subsets of $\{1, \ldots, n\}$. Then, gradient updates for $\mathrm{OT}_\varepsilon(\mu_m(X_{K:}), \mu_m(X_{L:})), \varepsilon \geq 0$ w.r.t. a point $X_{k:}$ of $X_{K:}$ $(k \in K)$ correspond to taking steps along the so-called barycentric transport map. Indeed, with (half) quadratic costs, it holds

(Cuturi and Doucet, 2014, §4.3) that

$$\nabla_{X_{k:}} \mathrm{OT}_\varepsilon(\mu_m(X_{K:}), \mu_m(X_{L:})) = \sum_\ell P^\star_{k\ell}(X_{k:} - X_{\ell:}),$$

where $P^\star$ is the optimal (regularized) transport plan. Therefore, a gradient based-update is of the form

$$X_{k:} \leftarrow (1-t)X_{k:} + t\sum_\ell P^\star_{k\ell} X_{\ell:}. \tag{3.4}$$

In a missing value imputation context, Equation (3.4) thus corresponds to updating values to make them closer to the target points given by transportation plans.

The imputation algorithm We propose a method using the previous criterion aiming at imputing missing values for quantitative variables by minimizing OT distances between batches. First, missing values of any variable are initialized with the mean of observed entries plus a small amount of noise (to preserve the marginals and to facilitate the optimization). Then, batches are sequentially sampled and the Sinkhorn divergence between batches is minimized with respect to the imputed values, using gradient updates (here using RMSprop (Tieleman and Hinton, 2012)). Taking a step back, one can see that Algorithm 3 essentially uses Sinkhorn divergences between

Algorithm 3 Batch Sinkhorn Imputation

Input: $X \in (\mathbb{R} \cup \{\texttt{NA}\})^{n\times d}$, $\Omega \in \{0,1\}^{n\times d}$, $\alpha, \eta, \varepsilon > 0$, $n \geq m > 0$,
Initialization: for $j = 1, \ldots, d$,

- for i s.t. $\Omega_{ij} = 0$ (missing entries),
  $\hat{X}_{ij} \leftarrow \overline{X^{obs}_{:j}} + \varepsilon_{ij}$, with $\varepsilon_{ij} \sim \mathcal{N}(0, \eta)$ and $\overline{X^{obs}_{:j}}$ corresponds to the mean of the observed entries in the j-th variable (missing entries)
- for i s.t. $\Omega_{ij} = 1$ (observed entries),
  $\hat{X}_{ij} \leftarrow X_{ij}$

for t = 1, 2, ...,t_{max} **do**
  Sample two sets K and L of m indices
  $\mathcal{L}(\hat{X}_{K:}, \hat{X}_{L:}) \leftarrow S_\varepsilon(\mu_m(\hat{X}_{K:}), \mu_m(\hat{X}_{L:}))$
  $\hat{X}^{(imp)}_{K\cup L:} \leftarrow \hat{X}^{(imp)}_{K\cup L:} - \alpha\mathrm{RMSprop}(\nabla_{X^{(imp)}_{K\cup L:}}\mathcal{L})$
end for
Output: $\hat{X}$

batches as a loss function to impute values for a model in which "one parameter equals one imputed value".

Although Algorithm 3 is straightforward, a downside is that it cannot directly generate imputations for out-of-sample data points with missing values. Hence, we also propose a natural extension - that we will not describe here in more details- to fit parametric imputation models, provided they are differentiable with respect to their parameters. The parametric algorithm is trained in a round-robin fashion similar to iterative conditional imputation techniques, as implemented for instance in the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011).

Numerical experiments These methods are showcased in extensive experiments on a variety of datasets and for different missing values proportions and mechanisms, including the difficult case of informative missing entries. The relevancy of the proposed method is shown in comparison to the most classical baselines (such as mean imputation, imputation by chained equations (ice) (van Buuren and Groothuis-Oudshoorn, 2011), softimpute (Hastie et al., 2015)), but also to deep learning techniques (MIWAE (Mattei and Frellsen, 2019), GAIN (Yoon et al., 2018), imputation via VAEs (Ivanov et al., 2018)). As noticeable result, we highlight that the linear round-robin model matches or out-performs `scikit`'s iterative imputer (ice) on MAE and RMSE scores for most datasets. Since both methods are based on the same cyclical linear imputation model but with different loss functions, this shows that the batched Sinkhorn loss is well-adapted to imputation with parametric models. Moreover, the proposed OT-based methods consistently outperform DL-based methods, and have the additional benefit of having a lower variance in their results overall.

We also provide toy examples including two-dimensional datasets with strong structures, such as an S-shape, half-moon(s), or concentric circles, see Figure 3.1. A 20% missing rate is introduced (void rows are discarded), and imputations performed using Algorithm 3 or the **ice** method are compared to the ground truth dataset. While the ice method is not able to catch the non-linear structure of the distributions at all, the Sinkhorn approach performs efficiently by imputing faithfully to the underlying complex data structure (despite the two half-moons and the S-shape being quite challenging). This is remarkable, since Algorithm 3 does not rely on any parametric assumption for the data. This underlines in a low-dimensional setting the flexibility of the proposed method. Finally, note that the trailing points which can be observed for the S shape or the two moons shape come from the fact that Algorithm 3 was used as it is, i.e. with pairs of batches *both* containing missing values, even though these toy examples would have allowed to use batches without missing values. In that case, we obtain imputations that are visually indistinguishable from the ground truth.

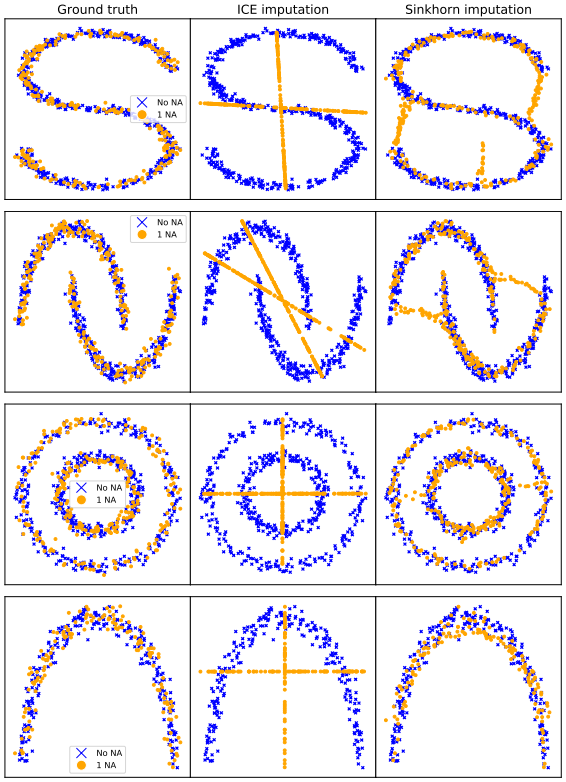


Figure 3.1: Examples w/ 20 % MCAR values on toy datasets. Blue points have no missing values, orange points have one missing value on either coordinate. **ice** outputs conditional expectation imputations (irrelevant due to the high non-linearity of these examples). Since Algorithm 3 does not assume a parametric form for the imputations, it is able to satisfyingly impute missing values.

☕ The imputation method is extremely flexible in the sense that it can be used to train a imputation model (with a cyclic scheme), or to perform an imputation without making any parametric assumption on the underlying distribution of the data. This represents a real applied contribution, but the theoretical ana-

lysis of such methods is still missing, even in simpler cases. At least the consistency of such a procedure should be guaranteed in a MCAR (and perhaps even MAR) mechanism. This would be a necessary step to understand in depth the limitations of the proposed approach.

3.2 Model estimation with online missing data

Context Stochastic gradient algorithms (SGD) (Robbins and Monro, 1951) play a central role in machine learning problems, due to their cheap computational cost and memory per iteration. There is a vast literature on its variants, for example using averaging of the iterates (Polyak and Juditsky, 1992b), some robust versions of SGD (Nemirovski et al., 2009; Juditsky et al., 2011) or adaptive gradient algorithms like Adagrad (Duchi et al., 2011); and on theoretical guarantees of those methods (Moulines and Bach, 2011; Bach and Moulines, 2013; Dieuleveut et al., 2017; Shamir and Zhang, 2013; Hazan and Kale, 2011; Needell et al., 2014).

Standard linear regression & SGD The data $(X_{i\cdot}, Y_i)_{1\leq i\leq n}$ are considered to be i.i.d. observations obeying the following linear model

$$Y_i = X_{i:}^\top \beta^\star + \epsilon_i,$$

where $Y_i \in \mathbb{R}$, $X_{i:} \in \mathbb{R}^d$, and $\epsilon_i \in \mathbb{R}$ are respectively the outcome variable, the input covariates and the noise term for the individual i; $\beta^\star \in \mathbb{R}^d$ denotes the model parameter, thus the object of attention in a model estimation context. The latter is also characterized as one minimizer of the following quadratic risk

$$\min_\beta \mathbb{E}_{X,Y}[(Y - X^\top \beta)^2] := R(\beta).$$

In practice as the data distribution is unknown, stochastic optimization strategies can be deployed, relying on the observations $(X_{i:}, Y_i)_i$, and using that a "local" gradient $(X_{i:}^\top \beta - Y_i)X_{i:}$, associated only to the i-th training point, can be seen as an unbiased estimation of the gradient of R at β.

SGD techniques are therefore natural and efficient answers to perform regression in large-scale settings and/or with data coming in a streaming fashion. However, if the training data happens to contain missing entries, the risk minimization with incomplete data becomes intractable and the usual results cannot be directly applied.

Contributions In Sportisse et al. (2020a), we study a debiased averaged stochastic gradient (SGD) algorithm to perform linear regression when the input data may be missing completely at random (MCAR).

The input variables $X_{i:}$ are assumed to come along, and to contain heterogeneous MCAR data, meaning that the probabilities $(p_j)_{1\leq j\leq d}$ to be observed may be specific for each input variable (but independent of the data values). To still perform linear regression despite missing values, we propose

(i) to naively impute the missing values by zero in order to get completed covariates $\tilde{X}_{i:} = X_{i:} \odot (\mathbf{1}_{n\times d} - \Omega_{i:})$ where $\Omega_{i:} \in \{0,1\}^d$ accounts for the missing pattern of the i-th input data;

(ii) to account for the imputation error by debiasing the gradients of the averaged SGD algorithm.

The averaging method has been indeed shown to stabilise the behaviour of the algorithm and reduce the impact of noise, resulting in better convergence rates (Bach and Moulines, 2013). This concretely means that instead of considering the traditional iterates β_k of SGD, the averaged SGD uses the Polyak-Ruppert averaged iterates $\bar{\beta}_k$ (Polyak and Juditsky, 1992b), which allows to account for all the iterates, robustifying the optimization dynamics

$$\beta_k = \beta_{k-1} - \alpha \tilde{g}_k(\beta_{k-1})$$
$$\bar{\beta}_k = \frac{1}{k+1} \sum_{i=0}^{k} \beta_i,$$

with $\tilde{g}_k$ unbiased stochastic gradients, such that $\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \mid \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1})$, with the filtration $\mathcal{F}_{k-1}$ with respect to $(X_{1:}, Y_1, \Omega_{1:}, \ldots, X_{k-1:}, Y_{k-1}, \Omega_{k-1:})$. Please refer to Algorithm 4 for details about the averaged SGD handling missing values in the case of linear regression.

Algorithm 4 Averaged SGD for linear regression with heterogeneous missing data

Input: data $(\tilde{X}_{i:})_{1\leq i\leq n}$ (imputed-by-0 inputs), $(Y_i)_{1\leq i\leq n}, \alpha$ (step size)
Initialize $\beta_0 = 0_d$
Set $P = \operatorname{diag}\left((p_j)_{j\in\{1,\ldots,d\}}\right) \in \mathbb{R}^{d\times d}$ (probabilities to be observed)
for $k = 1$ **to** n **do**
$\tilde{g}_k(\beta_{k-1}) = P^{-1}\tilde{X}_{k:}\left(\tilde{X}_{k:}^\top P^{-1}\beta_{k-1} - y_k\right)$
$\quad -(\mathrm{I}-P)P^{-2}\operatorname{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^\top\right)\beta_{k-1}$ (debiased gradient)
$\beta_k = \beta_{k-1} - \alpha\tilde{g}_k(\beta_{k-1})$
$\bar{\beta}_k = \frac{1}{k+1}\sum_{i=0}^{k}\beta_i = \frac{k}{k+1}\bar{\beta}_{k-1} + \frac{1}{k+1}\beta_k$ (averaging)
end for

Note that this type of SGD algorithm is also studied by Ma and Needell (2018) using the same strategy but this work assumes MCAR data, is restricted to the finite-sample setting in which the convergence rate is only conjectured, and is not suitable for the high-dimensional setting (as requiring a strong convexity parameter of the loss function). The main contribution of our work consists of adapting the powerful supervised learning SGD algorithm to deal with missing values, adapted both to the streaming setting, when the data arrive progressively, and to the high dimension setting, without adding strong parametric assumptions. These are the main advantages of our work compared to classical methods such as multiple imputation or the EM algorithm.

From a theoretical point of view, under weak assumptions on the observations, we derive a convergence rate of $\mathcal{O}(k^{-1})$ for our algorithm in the streaming setting.

Theorem 20. *Assume that for any i, $\|X_{i:}\| \leq \gamma$ almost surely for some $\gamma > 0$, and consider any constant step-size $\alpha \leq \frac{1}{2L}$, where L is the Lipschitz constant of the debiased gradients.*

Then Algorithm 4 ensures that, for any $k \geq 0$:

$$\mathbb{E}\left[R\left(\bar{\beta}_k\right) - R(\beta^\star)\right] \leq \frac{1}{2k}\left(\underbrace{\frac{\sqrt{c(\beta^\star)d}}{1-\sqrt{\alpha L}}}_{\text{variance term}} + \underbrace{\frac{\|\beta_0 - \beta^\star\|}{\sqrt{\alpha}}}_{\text{bias term}}\right)^2,$$

where

$$c(\beta^\star) = \underbrace{\underbrace{\frac{\mathrm{Var}(\epsilon_k)}{p_{\min}^2}}_{\textit{classical term}} + \underbrace{\left(\frac{(2+5p_{\min})(1-p_{\min})}{p_{\min}^3}\right)\gamma^2\|\beta^\star\|^2}_{\textit{multiplicative noise (induced by naive imputation)}}}_{\textit{increasing with the missing values rate}},$$

and $p_{\min} = \min_{j=1,\dots d}\ p_j$ *being the minimal probability to be observed.*

This convergence rate is remarkable, because it is optimal for the least squares regression and similar to the rate without missing values (Bach and Moulines, 2013). This result also sheds some light on the management of missing values in ML as discussed in the following remarks.

Remark 21 (What about discarding incomplete observations instead?)**.** *Another approach to solve the missing data problem is to discard all observations that have at least one missing feature. The probability that one input is complete, under our missing data model is* $\prod_{j=1}^d p_j$*. In the homogeneous case, the number of complete observations* k_{co} *out of a* $k-$*sample thus follows a binomial law* $k_{co} \sim \mathcal{B}(k, p^d)$*. With only those few observations, the statistical lower bound is* $\mathrm{Var}(\epsilon_k)d/k_{co}$*. In expectation, by Jensen inequality, we get that the lower bound on the risk is larger than* $\mathrm{Var}(\epsilon_k)d/kp^d$*.*

Our strategy thus leads to an upper-bound *which is typically* p^{d-3} *times smaller than the* lower bound *on the error of any algorithm relying only on complete observations. For a large dimension or a high percentage of missing values, our strategy is thus provably several orders of magnitude smaller than the best possible algorithm that would only rely on complete observations - e.g., if* $p = 0.9$ *and* $d = 40$*, the error of our method is at least 50 times smaller.*

Remark 22 (What do you prefer between having access to 200 half-incomplete training points or 100 complete training points?)**.** *An important question in practice is to understand how much information has been lost because of the incompleteness of the observations. In other words, it is better to access 200 input/output pairs with a probability 50% of observing each feature on the inputs, or to observe 100 input/output pairs with complete observations?*

Without missing observations, the variance bound in the expected excess risk is given by the previous theorem with $p_m = 1$*: it scales as* $\mathcal{O}\left(\frac{\mathrm{Var}(\epsilon_k)d}{k}\right)$*, while with missing observations it increases to* $\mathcal{O}\left(\frac{\mathrm{Var}(\epsilon_k)d}{kp_m^2} + \frac{C(X,\beta^\star)}{kp_m^3}\right)$*. As a consequence, the variance upper bound is larger by a factor* p_m^{-1} *for the estimator derived from* k *incomplete observations than for* $k \times p_m$ *complete observations. This suggests that there is a higher gain to collecting fewer complete observations (e.g., 100) than more incomplete ones (e.g., 200 with* $p = 0.5$*). However, one should keep in mind that this observation is made by comparing upper bounds thus does not necessarily reflect what would happen in practice.*

In order to assess the convergence behavior and the relevance of our algorithm, we conduct experiments on synthetic data and on real datasets including the Traumabase dataset.

☕ In this work, we thoroughly study the impact of missing values for the SGD algorithm when performing least squares regression. It raises several questions, including how this can be extended to other types of missing data, but also to generalized loss functions beyond the linear regression framework (for which it is challenging to build a debiased gradient estimator from observations with missing values).

3.3 Consistency of linear models with missing input data

In this section, we remain dedicated to a linear model

$$Y = X^\top \beta^\star + \varepsilon$$

with $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ the input/output variables, the model parameter $\beta^\star \in \mathbb{R}^d$ and a noise term ε. In such a linear regression framework, when missing data occur in the input variables X, most of the literature focuses on the *model estimation* -as in the previous section-, i.e. on the estimation of $\beta^\star$, (Little, 1992; Jones, 1996; Robins et al., 1994), using sometimes a sparse prior leading to the Lasso estimator (Loh and Wainwright, 2012) or the Dantzig selector (Rosenbaum and Tsybakov, 2010). The robust estimation literature (Dalalyan and Thompson, 2019; Chen and Caramanis, 2013) can be also invoked to handle missing values, as they can be reinterpreted as a multiplicative noise in linear models.

On the other hand, *prediction* with missing values in a parametric framework (even under the canonical linear model) is in fact not an easy task. Indeed, the prediction task is distinct from model estimation: estimated model parameters cannot be directly used to predict on a test sample containing missing values as well. As a matter of fact, the occurrence of missing data turns the linear regression problem into a semi-discrete one of very high complexity. Furthermore, establishing risk bounds -even without missing values- for random designs is already a challenge as studied by Györfi et al. (2006); Audibert and Catoni (2011); Dieuleveut et al. (2017) and more recently by Mourtada (2019).

Bayes predictor and NA Without missing values, the Bayes predictor (associated to the quadratic loss) is given as the linear function defined with the model parameter $\beta^\star$. However, when missing data occur, the Bayes predictor $f^\star$ can be decomposed according to the possible missing data patterns $m \in \{0,1\}^d$, as

$$f^\star(X_{\text{obs}(\Omega)}, \Omega) = \mathbb{E}\left[Y | X_{\text{obs}(\Omega)}, \Omega\right] = \sum_{m \in \{0,1\}^d} f_m^\star(X_{\text{obs(m)}}) \mathbb{1}_{\Omega = m},$$

where $X_{\text{obs}(m)}$ (resp. $X_{\text{mis}(m)}$) is the vector of observed components (resp. unobserved components) of X, and $f_m^\star(X_{\text{obs}(m)}) := \mathbb{E}\left[Y | X_{\text{obs}(m)}, \Omega = m\right]$ can be seen as the Bayes predictor conditionally on the event "$\Omega = m$". The function $f_m^\star$ can be written as

$$f_m^\star(X_{\text{obs}(m)}) = \beta_0 + \beta_{\text{obs}(m)}^\top X_{\text{obs}(m)} + \beta_{\text{mis}(m)}^\top \mathbb{E}\left[X_{\text{mis}(m)} | X_{\text{obs}(m)}, \Omega = m\right].$$

Thus, $f_m^\star$ remains linear in the observed variables X_{obs}, provided that $x \mapsto \mathbb{E}\left[X_{\text{mis(m)}} | X_{\text{obs(m)}} = x, \Omega = m\right]$ is a linear function. This is guaranteed by Le Morvan et al. (2020) in various settings: Gaussian input variables with M(C)AR data or with Gaussian self-masking mechanisms[1], Gaussian pattern-mixture models, or even independent covariates.

[1] this is a particular case of MNAR data for which $\mathbb{P}(\Omega|X) := \prod_{j=1}^d \mathbb{P}(\Omega_j|X_j) = \prod_{j=1}^d \varphi(X_j; a_j, b_j)$ where φ is the Gaussian density with parameters $(a_j)_j$ and $(b_j)_j$.

Contributions In Ayme et al. (2022), we study pattern-by-pattern predictors for regression with missing input variables. First, we provide a distribution-free excess risk bound for the pattern-by-pattern least squares estimator:

$$\hat{f}(X_{\mathrm{obs}(\Omega)}, \Omega) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{\mathrm{obs}(\Omega)}) \mathbb{1}_{\Omega = m}$$

with $\hat{f}_m$ being the traditional least squares estimator trained only on the samples sharing the same fixed missing pattern m.

Theorem 23 (Unformal)**.** *Assume that the input underlying covariates are sub-Gaussian and that $f^\star_m$ is L-Lipschitz for any missing pattern $m \in \{0,1\}^d$. Consider $\hat{f}$ to be the pattern-by-pattern least squares estimator trained with n samples. Then,*

$$\mathbb{E}\left[\left(\hat{f}(X_{obs(\Omega)}, \Omega) - f^\star(X_{obs(\Omega)}, \Omega)\right)^2\right] \lesssim (\log(n)+1)\big(\sigma^2_{\mathrm{na}} \vee L^2\big) 2^d \frac{d}{n} + \mathrm{Approx}(f^\star, \mathcal{F}),$$

where $\sigma^2_{na} := \sup_{(x,m)} \mathbb{V}\left[Y | (X_{obs(\Omega)}, \Omega) = (x, m)\right]$, and $\mathrm{Approx}(f^\star, \mathcal{F})$ is an approximation error made by approximating $f^\star$ with the help of the class $\mathcal{F}$ of linear pattern-by-pattern regressors.

Theorem 23 is the first theoretical result that provides a control on the excess risk of a least-square-type predictor under very general assumptions on the input variables distribution and without any assumption on the missing pattern distribution. In particular, the bound ensures that when $n > d2^d$, the considered predictor is better than the zero one. This curse of dimensionality is unfortunately unavoidable as it naturally arises for some specific worst-case distributions, for instance when all the missing patterns are equiprobable.

In what follows, we leverage the distribution of the missing patterns in order to derive less pessimistic theoretical bounds. To this end, we propose a refined version of the predictor previously introduced. The new predictor $\hat{f}^{(\tau)}$, for $\tau \in [0,1]$ is obtained by combining the previous pattern-by-pattern least squares predictors, but only for all patterns $m \in \{0,1\}^d$ that appear with an empirical proportion larger than τ. For the least frequent missing pattern, $\hat{f}^{(\tau)}$ arbitrarily returns 0.

Theorem 24 (Unformal)**.** *Under the same assumptions as in Theorem 23, set $p = (p_m)_{m \in \{0,1\}^d}$ to be the probability distribution of the missing patterns. Choose $\tau = d/n$ with n the sample size, then the generalization bound for the predictor $\hat{f}^{(d/n)}$ reads as*

$$\mathbb{E}\left[\left(\hat{f}^{(\tau)}(X_{obs(\Omega)}, \Omega) - f^\star(X_{obs(\Omega)}, \Omega)\right)^2\right] \lesssim (\log(n)+1)\big(\sigma^2_{\mathrm{na}} \vee L^2\big) \mathfrak{C}_p\left(\frac{d}{n}\right) + \mathrm{Approx}(f^\star, \mathcal{F}),$$

where the missing patterns distribution complexity $\mathfrak{C}_p(\tau)$ is defined by

$$\mathfrak{C}_p(\tau) := \sum_{m \in \{0,1\}^d} p_m \wedge \tau. \tag{3.5}$$

Similar bounds are obtained for a general threshold $\tau \geq 1/n$, but the obtained upper-bound is actually minimized for the choice $\tau = d/n$. Theorem 24 is the first result controlling the excess risk of a pattern-by-pattern least-square-type predictor with a bound depending on the missing pattern distribution through the complexity $\mathfrak{C}_p$, and that holds for any type of missing patterns (including MCAR, MAR, MNAR scenarios). The adaptivity of $\mathfrak{C}_p$ to the missing pattern distribution can be illustrated as follows. Imagine that with a high probability $1-\delta$, the missing pattern belongs to a class of K distinct missing patterns. Then,

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \lesssim K\frac{d}{n} + \delta.$$

This reflects the good learning ability of the regressor $\hat{f}^{(d/n)}$ when there are few frequent missing patterns.

Optimality of the contributions Finally the optimality of this bound is also studied. To do so, consider the class of problems $\mathcal{P}_p(\sigma, L)$ assumed to satisfy the following conditions: for all $\mathbb{P} \in \mathcal{P}_p(\sigma, L)$

(i) $\forall m \in \{0,1\}^d, \mathbb{P}(\Omega = m) = p_m$,

(ii) $Y = X^\top \beta^\star + \epsilon$ where $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$,

(iii) X is a sub-Gaussian vector, and $f_m^\star$ is L-Lipschitz for any missing pattern $m \in \{0,1\}^d$,

(iv) $\text{Approx}(f^\star, \mathcal{F}) = 0$.

For any missing pattern distribution p, one can show that

$$\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right) \lesssim \min_{\hat{f}} \max_{\mathbb{P} \in \mathcal{P}_p(\sigma,L)} \mathbb{E}_{\mathbb{P}}\left[\left(\hat{f} - f^\star\right)^2\right].$$

where the minimum is over all predictor $\hat{f}$. Since $\mathfrak{C}_p\left(\frac{1}{n}\right) \geq d^{-1}\mathfrak{C}_p\left(\frac{d}{n}\right)$, this lower bound emphasizes that previous results are sharp up to a factor d. This lower bound can be shown to be of the same order as that of the lower bound when we restrict the class $\mathcal{P}_p(\sigma, L)$ to MAR settings only, meaning that the worst-case scenario of $\mathcal{P}_p(\sigma, L)$ is as hard as the one of $\mathcal{P}_p(\sigma, L)$ restricted to MAR settings. While the MAR hypothesis is known to facilitate the inference framework (the former actually originates from the latter, see Rubin (1976)), we emphasize here that MAR scenarios do not help prediction purposes.

☕ In a high-dimensional setting $d \gg n$, the maps are shuffled, such that this is not feasible to compute a prediction function per pattern anymore, since there are not enough observations to learn all the patterns. A naive but relatively common strategy is to (i) first impute by 0 (or an arbitrary constant), in order to obtain a completed dataset and (ii) train the predictor in question (and thus a single predictor) on it. It is easy to be convinced that this method is biased. Indeed, the imputed dataset no longer follows the same law as the underlying initial data. We are currently working to show that under certain assumptions on the missing data mechanism, this bias decreases and even tends towards 0 when the dimension increases. The bias introduced by this naive imputation

can be also analyzed through the prism of a bias that would be introduced by a standard ridge regression based on complete data (relevant in a high dimensional context). This would mean that imputing by an arbitrary constant is actually harmless for prediction purposes in a high-dimensional setting. We also hope to make a parallel with dropout strategies (Srivastava et al., 2014; Gal and Ghahramani, 2016) used in neural network training.

Bibliography

Ben Adcock, Claire Boyer, and Simone Brugiapaglia. On oracle-type local recovery guarantees in compressed sensing. *Information and Inference: A Journal of the IMA*, 10(1):1–49, 2021.

Luigi Ambrosio, Vicent Caselles, Simon Masnou, and Jean-Michel Morel. Connected components of sets of finite perimeter and applications to image processing. *Journal of the European Mathematical Society*, 3(1):39–92, 2001.

Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

Ludovic Arnould, Claire Boyer, and Erwan Scornet. Analyzing the tree-layer structure of deep forests. In *International Conference on Machine Learning*, pages 342–350. PMLR, 2021.

Ludovic Arnould, Claire Boyer, and Erwan Scornet. Is interpolation benign for random forests? *arXiv preprint arXiv:2202.03688*, 2022.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor pca. *The Annals of Probability*, 48(4):2052–2087, 2020.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Near-optimal rate of consistency for linear models with missing values. In *International Conference on Machine Learning*, pages 1211–1243. PMLR, 2022.

Jean-Marc Azais, Yohann De Castro, and Fabrice Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.

Francis Bach and Lenaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization, 2021.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

Francis Bach et al. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Bernard Bercu, Antoine Godichon, and Bruno Portier. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1): 348–367, 2020.

Arne Beurling. Sur les intégrales de fourier absolument convergentes et leur applicationa une transformation fonctionelle. In *Ninth Scandinavian Mathematical Congress*, pages 345–366, 1938.

G. Biau, B. Cadre, and L. Rouvière. Accelerated Gradient Boosting. *Machine Learning*, 108(6): 971–992, 2019a. ISSN 0885-6125.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

Gérard Biau, Erwan Scornet, and Johannes Welbl. Neural random forests. *Sankhya A*, 81(2): 347–386, 2019b.

Jérémie Bigot, Claire Boyer, and Pierre Weiss. An analysis of block sampling strategies in compressed sensing. *IEEE transactions on information theory*, 62(4):2125–2139, 2016.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey, 2021. URL https://arxiv.org/abs/2110.01889.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Claire Boyer and Antoine Godichon-Baggioni. On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions, 2020. URL https://arxiv.org/abs/2011.09706.

Claire Boyer, Nicolas Chauffert, Philippe Ciuciu, Jonas Kahn, and Pierre Weiss. On the generation of sampling schemes for magnetic resonance imaging. *SIAM Journal on Imaging Sciences*, 9(4): 2039–2072, 2016.

Claire Boyer, Yohann De Castro, and Joseph Salmon. Adapting to unknown noise level in sparse deconvolution. *Information and Inference: A Journal of the IMA*, 6(3):310–348, 2017.

Claire Boyer, Jérémie Bigot, and Pierre Weiss. Compressed sensing with structured sparsity and structured acquisition. *Applied and Computational Harmonic Analysis*, 46(2):312–350, 2019a.

Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019b.

Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Richard P Brent. Fast training algorithms for multilayer neural nets. *IEEE Transactions on Neural Networks*, 2(3):346–354, 1991.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Sebastian Buschjäger and Katharina Morik. There is no double-descent in random forests. *arXiv preprint arXiv:2111.04409*, 2021.

Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

Emmanuel Candès and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

Emmanuel J Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

Emmanuel J Candès and Yaniv Plan. A probabilistic and ripless theory of compressed sensing. *IEEE transactions on information theory*, 57(11):7235–7254, 2011.

Emmanuel J Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

Emmanuel J Candès, Yonina C Eldar, Deanna Needell, and Paige Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31 (1):59–73, 2011.

Emmanuel J Candès, Carlos A Sing-Long, and Joshua D Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61 (19):4643–4657, 2013.

Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37. Société Mathématique de France France, 2012.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Shaobing Chen and David Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Yudong Chen and Constantine Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 383–391, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/chen13d.html.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Stéphane Chrétien and Sébastien Darses. Sparse recovery with unknown variance: a lasso-type approach. *IEEE Transactions on Information Theory*, 60(7):3970–3988, 2014.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, page 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, number 2, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

Arnak S. Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ell-1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems 32*, pages 13188–13198, 2019. URL http://arxiv.org/pdf/1904.06288.

Yohann De Castro and Fabrice Gamboa. Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding frank–wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 2019.

Pascaline Descloux, Claire Boyer, Julie Josse, Aude Sportisse, and Sylvain Sardy. Robust lasso-zero for sparse corruption and model selection with missing covariates. *Scandinavian Journal of Statistics*, 2022.

Luc Devroye, Laszlo Györfi, and Adam Krzyżak. The hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Lester E Dubins. On extreme points of convex sets. *Journal of Mathematical Analysis and Applications*, 5(2):237–244, 1962.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Marie Duflo. *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. ISBN 3-540-57100-0. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

Roxane Duroux and Erwan Scornet. Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128, 2018.

Wei Fan, Haixun Wang, Philip S Yu, and Sheng Ma. Is random model better? on its accuracy and efficiency. In *Third IEEE International Conference on Data Mining*, pages 51–58. IEEE, 2003.

Carlos Fernandez-Granda. Support detection in super-resolution. *arXiv preprint arXiv:1302.3921*, 2013.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 2681–2690, 2019.

Wendell H. Fleming. Functions with generalized gradient and generalized surfaces. *Annali di Matematica Pura ed Applicata*, 44(1):93–103, 1957.

Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

Erwan Fouillen, Claire Boyer, and Maxime Sangnier. Proximal boosting: aggregating weak learners to minimize non-differentiable losses. *Neurocomputing*, 2022. doi: 10.48550/ARXIV.1808.09670. URL https://arxiv.org/abs/1808.09670.

Yoav Freund and Robert .E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Jerome Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Jean-Jacques Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. *Stochastic Processes and their Applications*, 2022.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29, 2016.

Matan Gavish and David L Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

Antoine Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. *ESAIM: Probability and Statistics*, 23:841–873, 2019.

Xiao Gong, Wei Chen, Jie Chen, and Bo Ai. Tensor denoising using low-rank tensor train decomposition. *IEEE Signal Processing Letters*, 27:1685–1689, 2020. doi: 10.1109/LSP.2020.3025038.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 107–113. IET, 1993.

Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*, 2022.

Alex Grubb and J. Andrew Bagnell. Generalized Boosting Algorithms for Convex Optimization. In *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, Washington, USA, 2011.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

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

Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. *arXiv preprint arXiv:1806.02382*, 2018.

Michael P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91:222–230, 1996.

Julie Josse, Sylvain Sardy, and Stefan Wager. denoiser: A package for low rank matrix estimation. *Journal of Statistical Software*, 2016.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Maryia Kabanava and Holger Rauhut. Analysis ℓ^1-recovery with frames and gaussian measurements. *Acta Applicandae Mathematicae*, 140(1):173–195, 2015.

Maryia Kabanava, Holger Rauhut, and Hui Zhang. Robust analysis ℓ^1-recovery from gaussian measurements and total variation minimization. *European Journal of Applied Mathematics*, 26 (6):917–929, 2015.

Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal inference with noisy and missing covariates via matrix factorization. *arXiv preprint arXiv:1806.00811*, 2018.

Guolin Ke, Jia Zhang, Zhenhui Xu, Jiang Bian, and Tie-Yan Liu. Tabnn: A universal neural network solution for tabular data. 2018.

N Kishore Kumar and Jan Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.

Victor Klee. On a theorem of dubins. *Journal of Mathematical Analysis and Applications*, 7(3): 425–427, 1963. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(63)90063-5. URL https://www.sciencedirect.com/science/article/pii/0022247X63900635.

Felix Krahmer, Deanna Needell, and Rachel Ward. Compressive sensing with redundant dictionaries and structured measurements. *SIAM Journal on Mathematical Analysis*, 47(6):4606–4629, 2015.

Carole Lazarus, Pierre Weiss, Nicolas Chauffert, Franck Mauconduit, Loubna El Gueddari, Christophe Destrieux, Ilyess Zemmoura, Alexandre Vignaud, and Philippe Ciuciu. Sparkling: variable-density k-space filling curves for accelerated t2*-weighted mri. *Magnetic resonance in medicine*, 81(6):3643–3661, 2019a.

Carole Lazarus, Pierre Weiss, Franck Mauconduit, Alexandre Vignaud, and Philippe Ciuciu. 3d sparkling for accelerated ex vivo t2*-weighted mri with compressed sensing. In *ISMRM 2019-27th Annual Meeting & Exhibition*, 2019b.

Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR, 2020.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

Rémi Leluc and François Portier. Towards asymptotic optimality with conditioned stochastic gradient descent. *arXiv preprint arXiv:2006.02745*, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

Roderick JA Little. Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237, 1992.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637 – 1664, 2012. doi: 10.1214/12-AOS1018. URL https://doi.org/10.1214/12-AOS1018.

Patrick Lutz, Ludovic Arnould, Claire Boyer, and Erwan Scornet. Sparse tree-based initialization for neural networks. *arXiv preprint arXiv:2209.15283*, 2022.

Anna Ma and Deanna Needell. Stochastic gradient descent for linear systems with missing data. *Numerical Mathematics: Theory, Methods and Applications*, 12(1):1–20, 2018. ISSN 2079-7338. doi: https://doi.org/10.4208/nmtma.OA-2018-0066. URL http://global-sci.org/intro/article_detail/nmtma/12689.html.

Maximilian März, Claire Boyer, Jonas Kahn, and Pierre Weiss. Sampling rates for ℓ^1-synthesis. *arXiv preprint arXiv:2004.07175*, 2020.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

Karthika Mohan. On handling self-masking and other hard missing data problems. 2018.

Karthika Mohan, Felix Thoemmes, and Judea Pearl. Estimation with incomplete data: The linear case. In *IJCAI*, pages 5082–5088, 2018.

Abdelkader Mokkadem and Mariane Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Ann. Appl. Probab.*, 16(3):1671–1702, 2006.

Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv preprint arXiv:1912.10754*, 2019.

Jared S Murray and Jerome P Reiter. Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020.

Sangnam Nam, Mike E Davies, Michael Elad, and Rémi Gribonval. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30–56, 2013.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013.

Art B Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443(7): 59–72, 2007.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Mariane Pelletier. Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.*, 39(1):49–72, 2000. ISSN 0363-0129. doi: 10.1137/S0363012998308169. URL http://dx.doi.org/10.1137/S0363012998308169.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30:838–855, 1992a.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992b.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

Holger Rauhut, Karin Schnass, and Pierre Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory*, 54(5):2210–2219, 2008.

Emile Richard and Andrea Montanari. A statistical model for tensor pca. *Advances in neural information processing systems*, 27, 2014.

David L Richmond, Dagmar Kainmueller, Michael Y Yang, Eugene W Myers, and Carsten Rother. Relating cascaded random forests to deep convolutional neural networks for semantic segmentation. *arXiv preprint arXiv:1507.07583*, 2015.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, and Robert Tibshirani. Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association*, 115(531):1292–1303, 2020.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

Mathieu Rosenbaum and Alexandre B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620 – 2651, 2010. doi: 10.1214/10-AOS793. URL https://doi.org/10.1214/10-AOS793.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

David E Rumelhart, Geoffrey E Hinton, James L McClelland, et al. A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(45-76):26, 1986.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Ishwar Krishnan Sethi. Entropy nets: from decision trees to neural networks. *Proceedings of the IEEE*, 78(10):1605–1613, 1990.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training, 2021. URL https://arxiv.org/abs/2106.01342.

Aude Sportisse, Claire Boyer, Aymeric Dieuleveut, and Julie Josse. Debiasing averaged stochastic gradient descent to handle missing values. *Advances in Neural Information Processing Systems*, 33:12957–12967, 2020a.

Aude Sportisse, Claire Boyer, and Julie Josse. Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Advances in Neural Information Processing Systems*, 33:7067–7077, 2020b.

Aude Sportisse, Claire Boyer, and Julie Josse. Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643, 2020c.

Aude Sportisse, Christophe Biernacki, Claire Boyer, Julie Josse, Matthieu Marbac Lourdelle, Gilles Celeux, and Fabien Laporte. Model-based clustering with missing not at random data. *arXiv preprint arXiv:2112.10425*, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Ruo-Yu Sun. Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, 8(2):249–294, 2020.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research*, 14:3385–3418, 2013. URL http://jmlr.org/papers/v14/sun13a.html.

Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Near minimax line spectral estimation. *IEEE Transactions on Information Theory*, 61(1):499–512, 2014.

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

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Madeleine Udell and Alex Townsend. Nice latent variable models have log-rank. *ArXiv*, abs/1705.07474, 2017. URL http://arxiv.org/abs/1705.07474.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, Articles*, 45(3):1–67, 2011. ISSN 1548-7660.

Sara van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

Sara van de Geer and Francesco Ortelli. Tensor denoising with trend filtering. *Mathematical Statistics and Learning*, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Johannes Welbl. Casting random forests as artificial neural networks (and profiting from it). In *German Conference on Pattern Recognition*, pages 765–771. Springer, 2014.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.

Chengrun Yang, Yuji Akimoto, Dae Won Kim, and Madeleine Udell. Oboe: Collaborative filtering for automl initialization. *arXiv preprint arXiv:1808.03233*, 2018.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, pages 5689–5698, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Siyu Zhou and Lucas Mentch. Trees, forests, chickens, and eggs: when and why to prune trees in a random forest. *arXiv preprint arXiv:2103.16700*, 2021.

Zhi-Hua Zhou and Ji Feng. Deep forest. *National Science Review*, 6(1):74–86, 2019.

List of publications by the author

Journals

12. On the asymptotic rate of convergence of Stochastic Newton algorithms and their Weighted Averaged versions

    C. Boyer, A. Godichon-Baggioni

    Computational optimization and applications, Springer (2022)

11. Proximal boosting: aggregating weak learners to minimize non-differentiable losses

    E. Fouillen, C. Boyer, M. Sangnier

    Neurocomputing, Elsevier (2022)

10. Robust Lasso-Zero for sparse corruption and model selection with missing covariates

    P. Descloux, C. Boyer, J. Josse, A. Sportisse, S. Sardy

    Scandinavian Journal of Statistics (2022)

9. Sampling rates for ℓ^1-synthesis

    M. März, C. Boyer, J. Kahn, P. Weiss

    Foundations of Computational Mathematics (FoCM) (2022)

8. Imputation and low-rank estimation with Missing Non At Random data

    A. Sportisse, C. Boyer, J. Josse

    Statistics & Computing, Springer (2020)

7. On oracle-type local recovery guarantees in compressed sensing

    B. Adcock, C. Boyer, S. Brugiapaglia

    Information & Inference (2020)

6. On representer theorems and convex regularization.

    C. Boyer, A. Chambolle, Y. De Castro, V. Duval, F. de Gournay, P. Weiss

    SIAM Journal on Optimization (2019)

5. Compressed sensing with structured sparsity and structured acquisition
   C. Boyer, J. Bigot, P. Weiss
   Applied and Computational Harmonic Analysis (2017)

4. Adapting to unknown noise level in sparse deconvolution
   C. Boyer, Y. De Castro, J. Salmon
   Information and Inference (2017)

3. On the generation of sampling schemes for Magnetic Resonance Imaging
   C. Boyer, N. Chauffert, P. Ciuciu, J. Kahn, P. Weiss
   SIAM Journal on Imaging Sciences, Volume 9, Issue 4, pp. 1525-2098 (2016)

2. An analysis of block sampling strategies in compressed sensing
   J. Bigot, C. Boyer, P. Weiss
   IEEE Transactions on Information Theory (2016)

1. An algorithm for variable density sampling with block-constrained acquisition
   C. Boyer, P. Weiss and J. Bigot
   SIAM Imaging Science (2014)

International conferences

10. Is interpolation benign for random forest regression?
    L. Arnould, C. Boyer, E. Scornet
    International Conference on Artificial Intelligence and Statistics (AISTATS 2023)

9. Sparse tree-based initialization for neural networks
   P. Lutz, L. Arnould, C. Boyer, E. Scornet
   International Conference on Learning Representations (ICLR 2023)

8. Near-optimal rate of consistency for linear models with missing values
   A. Ayme, C. Boyer, A. Dieuleveut, E. Scornet
   International Conference on Machine Learning (ICML 2022)

7. Analyzing the tree-layer structure of Deep Forests
   L. Arnould, C. Boyer, E. Scornet
   International Conference on Machine Learning (ICML 2021)

6. Debiasing Stochastic Gradient Descent to handle missing values
   A. Sportisse, C. Boyer, A. Dieuleveut, J. Josse
   Conference on Neural Information Processing Systems (NeurIPS 2020)

5. Estimation and imputation in Probabilistic Principal Component Analysis with Missing Not At Random data

   A. Sportisse, C. Boyer, J. Josse

   Conference on Neural Information Processing Systems (NeurIPS 2020)

4. Missing Data Imputation using Optimal Transport

   B. Muzellec, J. Josse, C. Boyer, M. Cuturi

   International Conference on Machine Learning (ICML 2020)

3. Convex Regularization and Representer Theorems

   C. Boyer, A. Chambolle, Y. De Castro, V. Duval, F. de Gournay, P. Weiss

   iTWIST (2018)

2. Sampling by blocks of measurements in compressed sensing

   J. Bigot, C. Boyer, P. Weiss

   Proc. SampTA (2013)

1. HYR2PICS: Hybrid Regularized Reconstruction for Combined Parallel Imaging and Compressive Sensing in MRI

   C. Boyer, P. Ciuciu, P. Weiss and S. Mériaux

   ISBI (2012)

National conferences

2. Sur la génération de schémas d'échantillonnage compressé en IRM

   P. Weiss, N. Chauffert, C. Boyer, P. Ciuciu

   GRETSI (2015)

1. Échantillonnage compressé avec acquisition structurée et parcimonie structurée

   C. Boyer, J. Bigot, P. Weiss

   GRETSI 2015

Submitted works

2. Naive imputation implicitly regularizes high-dimensional linear models

   A. Ayme, C. Boyer, A. Dieuleveut, E. Scornet

1. Model-based Clustering with Missing Not At Random Data

   A. Sportisse, C. Biernacki, C. Boyer, J. Josse, M. Marbac Lourdelle, G. Celeux, F. Laporte