Model Consistency for Learning with Mirror-Stratifiable Regularizers

Jalal Fadili
Normandie Université

Guillaume Garrigos
Unversité Paris-Diderot

Jérôme Malick
CNRS and LJK, Grenoble

Gabriel Peyré
CNRS and ENS Paris

Abstract

Low-complexity non-smooth convex regularizers are routinely used to impose some structure (such as sparsity or low-rank) on the coefficients for linear predictors in supervised learning. Model consistency consists then in selecting the correct structure (for instance support or rank) by regularized empirical risk minimization. It is known that model consistency holds under appropriate non-degeneracy conditions. However such conditions typically fail for highly correlated designs and it is observed that regularization methods tend to select larger models. In this work, we provide the theoretical underpinning of this behavior using the notion of mirror-stratifiable regularizers. This class of regularizers encompasses the most well-known in the literature, including the ℓ_1 or trace norms. It brings into play a pair of primal-dual models, which in turn allows one to locate the structure of the solution using a specific dual certificate. We also show how this analysis is applicable to optimal solutions of the learning problem, and also to the iterates computed by a certain class of stochastic proximal-gradient algorithms.

1 Introduction

Regularized empirical risk minimization. We consider a general set-up for supervised learning where, given an input/output space $\mathcal{X} \times \mathcal{Y}$ endowed with a probability measure ρ, one wants to learn an estimator $f : \mathcal{X} \to \mathcal{Y}$ satisfying $f(x) \approx y$ for ρ-a.e. pair of data $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We restrict ourselves to the case where $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^p \times \mathbb{R}$, with p being the dimension of the feature space, and we search for an estimator that is linear in $\mathcal{X}$, meaning that f can be written $f_{w_0}(x) = \langle x,\, w_0 \rangle$ for some coefficient vector $w_0 \in \mathbb{R}^p$. A standard *modeling assumption* is that, among the minimizers of a quadratic expected risk, w_0 possesses some form of simplicity or low-complexity (e.g. sparsity or low-rank). In other words, w_0 is assumed to be the unique solution of

$$\min_{w \in \mathbb{R}^p} \left\{ R(w) : w \in \operatorname*{Argmin}_{w' \in \mathbb{R}^p} \mathbb{E}_\rho \left[(\langle w',\, \mathbf{x} \rangle - \mathbf{y})^2 \right] \right\} \quad (\mathrm{P}_0)$$

where $R : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semi-continuous (l.s.c.) convex regularizer, and $\mathbb{E}_\rho[\cdot]$ is the expectation of the random variable $(\mathbf{x}, \mathbf{y})$ w.r.t. the probability measure ρ.

In practice (P_0) cannot be solved directly because one does not have access to ρ; only a sequence of n independent and identically distributed (i.i.d.) pairs $(x_i, y_i)_{i=1}^n$ sampled from ρ is available. The conventional approach is then to consider a solution $\widehat{w}_{\lambda,n}$ of a penalized empirical risk minimization (ERM) of the form

$$\min_{w \in \mathbb{R}^p} \lambda R(w) + \frac{1}{2n} \sum_{i=1}^{n} (\langle x_i,\, w \rangle - y_i)^2. \quad (\mathrm{P}_{\lambda,n})$$

The regularization parameter $\lambda > 0$ is tuned as a (decreasing) function of n, balancing appropriately between fitting the data and inducing some desirable property promoted by the regularizer R.

Tracking the structure of the solution. A theoretical question in statistical learning is to understand how close a solution $\widehat{w}_{\lambda,n}$ of $(\mathrm{P}_{\lambda,n})$ comes to w_0. If $\widehat{w}_{\lambda_n,n} \to w_0$ (convergence being usually considered in probability) as $n \to +\infty$ with $\lambda_n \to 0$, then the estimator is said to be *consistent*. One is also generally interested in stating estimation rates, and a linear estimation rate corresponds to $\|\widehat{w}_{\lambda_n,n} - w_0\| \sim n^{-\frac{1}{2}}$ (to be understood in probability). Note that we are here discussing guarantees on the estimation risk and not on the prediction risk (i.e. on w and not on $f_w(x) = \langle w,\, x \rangle$), which is more challenging. In this paper, we investigate *model consistency*, that is, whether $w_{\lambda_n,n}$ and w_0 share the same structure for appropriately chosen λ_n and n large enough. Existing results on the subject heavily

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

rely on a non-degeneracy condition at w_0, which is often referred as an "irrepresentable condition" (see more details and references in Remark 1). In this case, one can show that for n large enough and $\lambda_n \sim n^{-1/2}$, model consistency holds; see for ℓ_1 (Zhao and Yu, 2006), ℓ_1-ℓ_2 (Bach, 2008a), nuclear norm (Bach, 2008b) and more generally for the class of partly-smooth functions (Vaiter et al., 2014). The first goal of this paper is to go one step further by formally analyzing the general and challenging case where the non-degeneracy assumption cannot be guaranteed.

Tracking the structure of proximal algorithms. Similar consistency questions arise for the approximations of solutions computed by stochastic proximal algorithms used to solve ($\mathrm{P}_{\lambda,n}$). Many non-smooth low-complexity structure-promoting regularizers are such that their proximal operator is easy to compute either explicitly (as for the ℓ_1 norm or the trace norm) or approximately to good precision (as for the total variation in one-dimension). Proximal-gradient algorithms are then the methods of choice for solving the structured optimization problem ($\mathrm{P}_{\lambda,n}$). For large-scale machine learning problems, one would typically prefer stochastic versions of these algorithms, which need only one observation to proceed with the iterate; see e.g. (A. Defazio and Lacoste-Julien, 2014; Xiao and Zhang, 2014). The second goal is then to understand if these iterates and w_0 share the same structure induced by R. This complements the existing convergence analysis of these algorithms; pointers to relevant literature are given in Section 3.

Paper organization. As explained above, this paper has two goals about general model consistency for (i) regularized learning models and (ii) stochastic algorithms for solving them. The low-complexity induced by popular regularizers reveals primal-dual partitions which allow us to localize optimal solutions and track iterates. Section 2 recalls the notion of mirror-stratifiable regularizers which provides this structural complexity partition. Then Section 3 states our model recovery results and discusses their originality with respect to the existing literature. The rationale and the milestones of the proofs are sketched in Section 4; details and technical results are established in the supplementary material. Finally Section 5 provides numerical illustrations of our results, giving theoretical justification of typical observed behaviors of stochastic algorithms.

2 Low-complexity models

Low-complexity and stratification. In this paper, we study model consistency for a large class of regularizers, and under few structural assumptions. Our results strongly rely on duality arguments, and on a structure induced by ∂R (where ∂R is the subdifferential of R). To track the structure of solutions, we introduce an appropriate *stratification* $\mathcal{M} = \{M_i\}_{i\in I}$ of $\mathrm{dom}(\partial R) \subset \mathbb{R}^p$ (where $\mathrm{dom}(\partial R) := \{w \in \mathbb{R}^p \; : \; \partial R(w) \neq \emptyset\}$), which is a finite partition such that for any strata M and M'

$$M \cap \mathrm{cl}(M') \neq \emptyset \;\Rightarrow M \subset \mathrm{cl}(M')$$

(where cl stands for the topological closure of the set). Because this is a partition, any element $w \in \mathrm{dom}(\partial R)$ belongs to a unique stratum, which we denote M_w. A stratification also induces a partial ordering $\leqslant$ as follows

$$M \leqslant M' \iff M \subset \mathrm{cl}(M') \iff M \cap \mathrm{cl}(M') \neq \emptyset. \quad (1)$$

With such ordering, it is natural to see some strata as being "smaller" than others, and, by extension, to say that the elements of such small strata have a low-complexity.

Example 1. *Most regularizers R used in machine learning naturally come up with a stratification, in the sense that they promote solutions belonging to small (for the relation $\leqslant$) strata M.*

- *Lasso (Tibshirani, 1996): the simplest example is the ℓ_1 norm where $R(w) = \sum_i |w_i|$, where the strata are the sets of vectors $M_I = \{w \in \mathbb{R}^p \; : \; \mathrm{supp}(w) = I\}$, where $I \subset \{1, \cdots, p\}$.*
- *Nuclear (a.k.a. trace) norm (Fazel, 2002): this is another popular example where $R(w)$ is the ℓ_1 norm of the singular values of w, and where the strata are the manifolds of fixed-rank matrices: $M_r = \{w \in \mathbb{R}^{p_1 \times p_2} \; : \; \mathrm{rank}(w) = r\}$, where $r \in \{0, \cdots, \min(p_1, p_2)\}$.*
- *Many other examples fall within this class of regularizers. For instance the ℓ_1-ℓ_2-norm to promote group-sparsity (Yuan and Lin, 2005), or the fused Lasso (Tibshirani et al., 2005). Yet another example is the total variation semi-norm $R(w) = \|Dw\|_1$ where D is a discrete approximation to the "gradient" operator (on a regular grid or on a graph); in this case, the strata are defined by piecewise constant vectors sharing the same jump set (edges in signals or images).*

Mirror-Stratifiable Regularizers. All the classical regularizers mentioned in Example 1 have moreover a strong relation between their primal and dual stratifications. These primal-dual relations are defined through the following correspondence operator $\mathcal{J}_R$ between subsets $S \subset \mathbb{R}^p$,

$$\mathcal{J}_R(S) := \bigcup_{x \in S} \mathrm{ri}(\partial R(x)),$$

where ri denotes the relative interior of a convex set. Following Fadili et al. (2017), we define mirror-stratifiabilty as follows.

Definition 1. *Let R be a proper lsc and convex function and R^* its Legendre-Fenchel conjugate. R is mirror-stratifiable with respect to a (primal) stratification $\mathcal{M} = \{M_i\}_{i\in I}$ of $\operatorname{dom}(\partial R)$ and a (dual) stratification $\mathcal{M}^* = \{M_i^*\}_{i\in I}$ of $\operatorname{dom}(\partial R^*)$ if $\mathcal{J}_R : \mathcal{M} \to \mathcal{M}^*$ is invertible with inverse $\mathcal{J}_{R^*}$ and $\mathcal{J}_R$ is decreasing for the relation $\leqslant$ defined by (1).*

This structure finds its roots in (Daniilidis et al., 2014), which introduces the tools to show that polyhedral functions, as well as spectral lifting of polyhedral functions, are mirror-stratifiable. In particular, all popular regularizers mentioned above (ℓ_1 norm, ℓ_1-ℓ_2 mixed norms, nuclear norm, total variation semi-norm) are mirror-stratifiable; see (Fadili et al., 2017).

Example 2. *Let us illustrate this notion in the case $R = \|\cdot\|_1$. As mentioned in Example 1, the strata M_I of $\operatorname{dom}(\partial R) = \mathbb{R}^p$ are sets of sparse vectors, with prescribed support. In the dual, $\operatorname{dom}(\partial R)^*$ is the unit ℓ_∞-ball, which can be naturally stratified by sets of vectors in $[-1,1]^p$ with a prescribed active set. More precisely, if we define*

$$\operatorname{active}(\eta) := \{i \in \{1,\cdots,p\} \ : \ |\eta_i| = 1\},$$

then these strata are of the form $M_I^* = \{\eta \in [-1,1]^p \ : \ \operatorname{active}(\eta) = I\}$. It is then an easy exercise to verify that the following correspondence operators $\mathcal{J}_R$ and $\mathcal{J}_{R^*}$ induce a decreasing bijection between the dual strata M_I^* and the primal strata M_I, meaning that:

$$(\forall I, J \subset \{1,\cdots,p\}) \quad \mathcal{J}_R(M_I) = M_I^*,\ \mathcal{J}_{R^*}(M_I^*) = M_I \\ \text{and} \quad I \subset J \Leftrightarrow M_I \leqslant M_J \Leftrightarrow M_I^* \geqslant M_J^*.$$

All the regularizers in Example 1 work in the same way. For instance, for the nuclear norm, the strata M_r made of rank-r matrices are in correspondence with strata M_r^* made of matrices having exactly r singular values equal to 1, and the others being of smaller amplitude.

3 Main results

We study model consistency by bypassing unrealistic assumptions (e.g., irrepresentable-type condition) and thus obtain flexible theoretical results. Throughout this paper, we only assume the following hypotheses:

$$\begin{cases} R \text{ is mirror-stratifiable,} \\ R \text{ is bounded from below,} \\ w_0 \text{ is the unique solution of } (\mathrm{P}_0). \end{cases} \tag{$\mathrm{H_M}$}$$

Under ($\mathrm{H_M}$), we establish general model consistency results of optimal solutions of the regularized ERM problem ($\mathrm{P}_{\lambda,n}$) (in Section 3.1), and of iterates of stochastic proximal algorithms to solve it (in Section 3.2). We also discuss how these results encompass the existing model consistency results (in Section 3.3).

Our analysis leverages the strong primal-dual structure of mirror-stratifiable regularizers, which is our key tool to localize the active strata at the solution of ($\mathrm{P}_{\lambda,n}$), even in the case where the irrepresentable condition is violated. We show that an enlarged model consistency holds, where the identified structure lies between the ideal one (the structure of w_0) and a worst-case one controlled by a particular dual element (the so-called dual vector/certificate)

$$\eta_0 \in \partial R(w_0),$$

defined as the optimal solution[1]

$$\eta_0 = \operatorname{Argmin}\left\{\langle C^\dagger \eta, \eta\rangle : \eta \in \partial R(w_0) \cap \operatorname{Im} C\right\} \tag{D_0}$$

where $C := \mathbb{E}_\rho\left[\mathbf{x}\mathbf{x}^\top\right] \in \mathbb{R}^{p\times p}$ is the expected (non-centered) covariance matrix, and $C^\dagger$ denotes its Moore-Penrose pseudo-inverse. The role of η_0 in sensitivity analysis of regularized ERM problems is well-known, but has been always done under a non-degeneracy assumption (see forthcoming discussions in Remark 1 and Section 3.3).

3.1 Model consistency for regularized ERM

Our first contribution, Theorem 1 below, states that for an appropriate regime of (λ_n, n), one can precisely localize with probability 1 the active stratum at $\widehat{w}_{\lambda_n,n}$ between a minimal active set associated to w_0 and a maximal one controlled by the dual vector η_0. In the special case of ℓ_1 minimization, this means that, almost surely, the support of $\widehat{w}_{\lambda_n,n}$ can be larger than that of w_0 but cannot be larger than the extended support characterized by $\operatorname{active}(\eta_0)$. This holds provided that λ_n decreases to 0 with n, but not too fast to account for errors stemming from the finite sampling.

Theorem 1. *Assume that ($\mathrm{H_M}$) holds, and suppose that $\mathbb{E}_\rho\left[\|\mathbf{x}\|^4\right] < +\infty$ and $\mathbb{E}_\rho\left[|\mathbf{y}|^4\right] < +\infty$. Let $(\lambda_n)_{n\in\mathbb{N}} \subset]0,+\infty[$ be such that*

$$\lambda_n \to 0 \quad \text{with} \quad \lambda_n \sqrt{n/(\log\log n)} \to +\infty.$$

Then, for n large enough, the following holds with probability 1:

$$M_{w_0} \leqslant M_{\widehat{w}_{\lambda_n,n}} \leqslant \mathcal{J}_{R^*}(M^*_{\eta_0}). \tag{2}$$

[1] Though we do not assume C to be invertible, η_0 is indeed unique since $\operatorname{Ker} C^\dagger = \operatorname{Im} C^\perp$. In the case where C is invertible, η_0 coincides with the element of $\partial R(w_0)$ having minimal norm, in the metric induced by C^{-1}.

Example 3. *Using the notations of Example 1 and 2, the enlarged consistency (2) specializes to*

$$\begin{aligned} &\mathrm{supp}(w_0) \subset \mathrm{supp}(\widehat{w}_{\lambda_n,n}) \subset \mathrm{active}(\eta_0), \\ &\mathrm{rank}(w_0) \leqslant \mathrm{rank}(\widehat{w}_{\lambda_n,n}) \leqslant \#\left\{s \in \sigma(\eta_0) \ : \ |s| = 1\right\}, \end{aligned}$$

for the ℓ_1 norm and the nuclear norm, respectively. $\sigma(\eta_0)$ denotes the vector of singular values of η_0.

The theorem guarantees that we have an enlarged model consistency, as soon as enough data is sampled. The first interest of this result is the finite identification, compared to the existing asymptotic results (even if the level of generality does not allow us to provide a bound on n); we discuss this in Section 3.3. The second and main advantage of our result is that it does not require any unrealistic non-degeneracy assumption. We explain this point in the next two remarks, by looking at the usual assumption and how it often fails to hold in high dimension.

Remark 1 (Irrepresentable condition and exact model consistency)**.** *If it is furthermore assumed that*

$$\eta_0 \in \mathrm{ri}(\partial R(w_0)), \tag{IC}$$

then it follows from Definition 1 that $M_{w_0} = \mathcal{J}_{R^*}(M^*_{\eta_0})$. In that setting, the consistency (2) just gives exact model consistency

$$M_{w_0} = M_{\widehat{w}_{\lambda_n,n}}.$$

This relative interiority assumption (IC) corresponds exactly to the "irrepresentable condition" which is classical in the learning literature (Zhao and Yu, 2006),(Bach, 2008a),(Bach, 2008b). Without this non-degeneracy hypothesis, we cannot expect to have exact model consistency (this is for instance illustrated in Section 5). The above theorem shows that there is still an approximate optimal model consistency, with two extreme strata fully characterized by the primal-dual pair (w_0, η_0). Our result is thus able to explain what is going on in the intricate situation where (IC) is violated.

Remark 2 (When the irrepresentable condition fails)**.** *The originality and interest of our model consistency result is that condition (IC) is not required to hold, since it is usually not valid in the context of large-scale learning. Let us give some insights on this condition in the specific case of ℓ_1-regularized problems. For instance, if the x_i's are drawn from a standard Gaussian i.i.d. distribution, the compressed sensing literature provides sample thresholds depending on the dimension p and the sparsity level $s = \|w_0\|_0$. In this scenario, it is known that uniqueness in ($\mathrm{H_M}$) holds for $n > 2s\log(p/s)$ (Amelunxen et al., 2014), while the irrepresentable condition holds only for $n > 2s\log(p)$ (Candes and Recht, 2013): the gap between these thresholds corresponds to the case where (IC) fails. Observe nevertheless that these results rely on the assumption that the features are incoherent (here Gaussian i.i.d.), which is not likely to be verified in a learning scenario, where they are typically highly correlated. A setting with a coherent operator C is that of deconvolution, where C is a (discerete) convolution operator associated to a smooth kernel, which is widely studied in the signal/image processing literature (in particular for the super-resolution). In this case, one can exactly determine the largest manifold $\mathcal{J}_{R^*}(M^*)$ involved in (2), see (Duval and Peyré, 2017).*

3.2 Model consistency for stochastic proximal-gradient algorithms

Our second main result describes model consistency for the iterates generated by a stochastic algorithm. In our situation, the general (relaxed) stochastic proximal gradient algorithm for solving ($\mathrm{P}_{\lambda,n}$) reads, starting from any initialization $\widehat{w}^0$, at iteration k:

$$\begin{cases} \widehat{d}^k = (\langle \widehat{w}^k, x_{i(k)}\rangle - y_{i(k)})x_{i(k)} + \widehat{\varepsilon}^k, \\ \widehat{z}^k = \mathrm{prox}_{\gamma_k \lambda R}(\widehat{w}^k - \gamma_k \widehat{d}^k), \\ \widehat{w}^{k+1} = (1-\alpha_k)\widehat{w}^k + \alpha_k \widehat{z}^k, \end{cases} \tag{RSPG}$$

where $(x_{i(k)}, y_{i(k)})$ are independent random variables drawn among $(x_i, y_i)_{i=1}^n$, $\gamma_k \in]0, +\infty[$ and $\alpha_k \in]0,1]$ are respectively deterministic stepsize and relaxation parameters. As it is, the iteration is written in an abstract way, since we do not specify how to define the random $\mathbb{R}^p$-valued variables $\widehat{\varepsilon}^k$. But as we explain below, several known stochastic methods can be written under the form of (RSPG) when $\alpha_k \equiv 1$.

Example 4. *If one takes $\widehat{\varepsilon}^k \equiv 0$, then (RSPG) becomes simply the proximal stochastic gradient method (Prox-SGD). Variance-reduced methods, like the SAGA algorithm (A. Defazio and Lacoste-Julien, 2014), or the Prox-SVRG algorithm (with option I) (Xiao and Zhang, 2014), also fall into this scheme. For these algorithms the idea is to take $\widehat{\varepsilon}^k$ as a combination of previously computed estimates of the gradient, in order to reduce the variance of $\widehat{d}^k$. For instance, SAGA corresponds to the choice:*

$$\widehat{\varepsilon}^k = \frac{1}{n}\sum_{i=1}^{n} g_{k,i} - g_{k,i(k)}$$

where the stored gradients are updated as

$$g_{i,k} := \begin{cases} (\langle \widehat{w}^k, x_{i(k)}\rangle - y_{i(k)})x_{i(k)} & \text{if } i = i(k) \\ g_{k-1,i} & \text{else.} \end{cases}$$

We show in Theorem 2 that for n large enough and λ_n appropiately chosen, we can identify after a finite number of iterations of (RSPG) an active stratum, which is again localized between two strata controlled by w_0 and η_0, respectively. For this result to hold, we have to make some reasonable assumptions on algorithm (RSPG). We need first to make hypotheses on the parameters α_k, γ_k, $\widehat{\varepsilon}^k$, to ensure that the iterates of (RSPG) converge to a solution of ($\mathrm{P}_{\lambda,n}$). Such hypotheses have been investigated in (Combettes and Pesquet, 2016; Rosasco et al., 2016; Atchadé et al., 2017) to establish useful convergence results. Beyond convergence, we study *structure identification* of these algorithms. It is known that convergence is not enough for model consistency of iterates. For instance, the classical proximal stochastic gradient method (corresponding to the case $\widehat{\varepsilon}^k \equiv 0$) is known to fail at generating sparse iterates for the case $R = \|\cdot\|_1$; see below Example 5 for discussions and references. To ensure the identification of low-dimensional strata, we require some control on the variance of the descent direction, by acting either on the parameters γ_k and α_k, or by wisely controlling $\widehat{\varepsilon}^k$. Before stating formally this set of hypotheses, we introduce $L_n := (1/n)\|\sum_{i=1}^n x_i x_i^*\|$ and the σ-algebra $\mathcal{F}_k := \sigma(\widehat{w}^1, \dots, \widehat{w}^k)$ generated by the first k iterates.

$$\begin{cases} \sigma_k \in [0,+\infty[,\ \alpha_k \in]0,1],\ \gamma_k \in]0, 2/L_n[\\ \mathbb{E}\left[\widehat{\varepsilon}^k | \mathcal{F}_k\right] = 0,\ \mathrm{Var}\left[\widehat{d}^k | \mathcal{F}_k\right] \leqslant \sigma_k^2 \\ \widehat{d}^k - \mathbb{E}\left[\widehat{d}^k | \mathcal{F}_k\right] \text{ converges a.s. to } 0 \\ \sum_{k=1}^{\infty} \alpha_k \gamma_k^2 \sigma_k^2 < +\infty \\ \|\widehat{w}^{k+1} - \widehat{w}^k\| = o(\alpha_k \gamma_k) \text{ a.s.} \end{cases} \tag{$\mathrm{H_A}$}$$

Let us briefly discuss these hypotheses. The second line in ($\mathrm{H_A}$) imposes some control on the variance of $\widehat{d}^k$. The fourth line asks for a fine balance between the parameters α_k, γ_k and σ_k. For instance, one could take α_k and γ_k to be constant, and work essentially on $\widehat{\varepsilon}^k$ to ensure that $\sigma_k \in \ell^2(\mathbb{N})$ and $\widehat{w}^k$ is a.s. asymptotically regular. Instead, one could consider an algorithm where σ_k does not vanish, but with appropriately decreasing step-sizes: γ_k and α_k should be carefully chosen to guarantee that the fourth row of ($\mathrm{H_A}$) holds.

Theorem 2. *Assume that ($\mathrm{H_M}$) holds, and suppose that $\mathbb{E}\left[\|\mathbf{x}\|^4\right] < +\infty$ and $\mathbb{E}\left[|\mathbf{y}|^4\right] < +\infty$. Let $(\lambda_n)_{n\in\mathbb{N}} \subset]0,+\infty[$ be such that*

$$\lambda_n \to 0 \quad \textit{with } \lambda_n \sqrt{n/(\log\log n)} \to +\infty.$$

Then, for n large enough, if $(\widehat{w}^k)_{k\in\mathbb{N}}$ is generated by (RSPG) under assumption ($\mathrm{H_A}$), then for k large enough:

$$M_{w_0} \leqslant M_{\widehat{z}^k} \leqslant \mathcal{J}_{R^*}(M^*_{\eta_0}) \quad \textit{holds almost surely.}$$

Example 5. *Let us look at two instances of (RSPG).*

- *The SAGA (resp. Prox-SVRG) algorithm is shown to verify ($\mathrm{H_A}$) in (Poon et al., 2018), provided that $\alpha_k \equiv 1$ and $\gamma_k \equiv \gamma = 1/(3L_n)$ (resp. $\gamma_k \equiv \gamma$ taken small enough).*
- *The proximal stochastic gradient method (Prox-SGD) is a specialization of (RSPG) with $\alpha_k \equiv 1$, $(\gamma_k)_{k\in\mathbb{N}} \in \ell^2(\mathbb{N}) \setminus \ell^1(\mathbb{N})$ and $\widehat{\varepsilon}^k \equiv 0$. If the iterates are bounded, the second line of ($\mathrm{H_A}$) automatically holds by the (strong) law of large numbers. Nevertheless, this algorithm does not satisfy the conclusions of Theorem 2: this was observed in (Xiao, 2010; Lee and Wright, 2012; Poon et al., 2018), and is illustrated in Section 5. A simple explanation is that for this algorithm, σ_k does not converge to 0, which is why we need to impose that the stepsize γ_k tends to zero. Even if it can be shown that $\|\widehat{w}^{k+1} - \widehat{w}^k\|/\gamma_k$ is bounded, it cannot be ensured that it is $o(1)$, which breaks the last hypothesis in ($\mathrm{H_A}$). Thus Theorem 2 does not apply in agreement with the observed behaviour of the SGD algorithm.*

3.3 Relation to previous results.

Model consistency of the regularized ERM has already been investigated for special cases (ℓ_1 (Zhao and Yu, 2006), ℓ_1-ℓ_2 (Bach, 2008a), or nuclear norm (Bach, 2008b)) and for the class of partly-smooth functions (Vaiter et al., 2014). The existing results hold asymptotically in probability, e.g. of the form

$$\lim_{n\to+\infty} \mathbb{P}\left(M_{w_0} = M_{\widehat{w}_{\lambda_n,n}}\right) = 1, \tag{3}$$

while we show that the consistency (2) holds almost surely, as soon as enough data is sampled. Nevertheless, our result lacks a quantitative estimation of how large n should be for the identification to hold. As a comparison, (Vaiter et al., 2014) shows that the probability in (3) converges as $1 - n^{-1/2}$, but the result heavily relies on the assumption that (IC) holds (which prevents the solution from "jumping" between the strata M_{w_0} and $\mathcal{J}_{R^*}(M^*_{\eta_0})$). Such qualitative estimates cannot be derived in our more general results without stronger assumptions and/or structure, which we want to avoid.

Compared to previous works, a chief advantage of our model consistency results is thus to avoid making an assumption which often fails to hold in high dimension. Indeed, as explained in Remark 2 and Section 5, the above-mentioned existing results hold under the irrepresentable condition (IC); and many of these also assume that the expected covariance matrix $C = \mathbb{E}_\rho\left[\mathbf{x}\mathbf{x}^\top\right]$ is invertible. The first work to deal with model consistency for a large class of functions without the irrepresentable condition assumption is (Fadili et al., 2017), which introduces of the notion of mirror-stratifiable functions, from which the authors derive identification properties of a *deterministic* penalized problem. Our Theorem 1 comes with a similar flavor, but extended to a supervised learning scenario and random sampling, which brought technical challenges as detailed in Section 4.

Finite activity identification for stochastic algorithms has been a topic of interest in the past years. (Xiao, 2010) made the observation that Prox-SGD has not the identification property for the ℓ_1 case. Instead, finite activity identification was proved by Lee and Wright (2012), for the regularized dual averaging (RDA) method, and by Poon et al. (2018), for the SAGA and Prox-SVRG algorithms. For these two papers, the regularizer R is assumed to be partly-smooth, and a non-degeneracy assumption is made. Again, Theorem 2 does not need such an assumption. We also propose a general set of hypotheses ($\mathrm{H_A}$) encompassing all these algorithms and beyond: this allows an explanation for why Prox-SGD fails (see Example 5), and could be used to analyze other algorithms than SAGA or Prox-SVRG.

4 Sketch of proofs

Our model consistency results follow from a sequence of results controlling the behaviour of optimal solutions and of iterates of algorithms. In this section, we sketch the rationale and the milestones of the proof; the proof of the two intermediate technical results are given in the supplementary material.

The core of the proofs rely on (Fadili et al., 2017, Theorem 1) about sensivity analysis of mirror-stratifiable functions. We state this result here in a modified form that is adapted to our analysis.

Proposition 1. *Let R be mirror-stratifiable. Then, there exists $\delta > 0$ such that for any pair $\eta \in \partial R(w)$,*

$$\max\{\|w-w_0\|, \|\eta-\eta_0\|\} \leqslant \delta \Rightarrow M_{w_0} \leqslant M_w \leqslant \mathcal{J}_{R^*}(M^*_{\eta_0}).$$

Proof. Suppose for contradiction that no such δ exists. Let $(\delta_k)_{k\in\mathbb{N}} \subset]0,+\infty[$ and $(w^k,\eta^k)_{k\in\mathbb{N}} \subset \mathrm{gph}(\partial R)$ be such that $\delta_k \downarrow 0$, $\max(\|w^k-w_0\|, \|\eta^k-\eta_0\|) \leqslant \delta_k$, but where M_{w^k} does not satisfy the claimed inequalities. Then $(w^k,\eta^k) \to (w_0,\eta_0)$ as $k \to \infty$, and $(w_0,\eta_0) \in \mathrm{gph}(\partial R)$ by definition in ($\mathrm{D_0}$). Upon applying (Fadili et al., 2017, Theorem 1), we have that $M_{w_0} \leqslant M_{w^k} \leqslant \mathcal{J}_{R^*}(M^*_{\eta_0})$ for k sufficiently large. This is a contradiction with the choice of w^k. □

Concerning Theorem 1, we introduce the notations

$$\widehat{C}_n := \frac{1}{n}\sum_i x_i x_i^\top \in \mathbb{R}^{p\times p},$$

$$\widehat{u}_n := \frac{1}{n}\sum_i y_i x_i \in \mathbb{R}^p, \quad u := \mathbb{E}_\rho\left[\mathbf{yx}\right],$$

which allows us to rewrite problems ($\mathrm{P_0}$) and ($\mathrm{P}_{\lambda,n}$) in a compact form:

$$\{w_0\} = \underset{w\in\mathbb{R}^p, Cw=u}{\mathrm{Argmin}} \; R(w),$$

$$\widehat{w}_{\lambda,n} \in \underset{w\in\mathbb{R}^p}{\mathrm{Argmin}} \; \lambda R(w) + \frac{1}{2}\langle \widehat{C}_n w, w\rangle - \langle \widehat{u}_n, w\rangle.$$

The optimality conditions for ($\mathrm{P}_{\lambda,n}$) allow to derive:

$$\widehat{\eta}_{\lambda_n,n} \in \partial R(\widehat{w}_{\lambda_n,n}), \quad \widehat{\eta}_{\lambda_n,n} := \frac{\widehat{u}_n - \widehat{C}_n \widehat{w}_{\lambda_n,n}}{\lambda_n}. \tag{4}$$

In view of Proposition 1, establishing Theorem 1 essentially boils down to showing the following proposition. The proof of this proposition requires technical lemmas to control the interlaced effects of convergence and sampling; see the supplementary material for details.

Proposition 2. *Under the assumptions of Theorem 1, $(\widehat{w}_{\lambda_n,n}, \widehat{\eta}_{\lambda_n,n}) \underset{n\to+\infty}{\longrightarrow} (w_0,\eta_0)$ almost surely.*

From Proposition 2, we deduce that there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$:

$$\max\{\|\widehat{w}_{\lambda_n,n} - w_0\|, \|\widehat{\eta}_{\lambda_n,n} - \eta_0\|\} \leqslant \delta/2 \quad a.s. \tag{5}$$

Using Proposition 1, we deduce that (2) holds a.s. for all $n \geqslant N$. To prove Theorem 2, we keep $n \geqslant N$ fixed, and consider $(\widehat{w}^k)_{k\in\mathbb{N}}$ to be generated by the (RSPG) algorithm. Using the definition of $\widehat{w}^{k+1}$, we can write

$$\widehat{z}^k = \widehat{w}^k + \frac{\widehat{w}^{k+1} - \widehat{w}^k}{\alpha_k}, \tag{6}$$

$$\widehat{w}^k - \gamma_k \widehat{d}^k \in \widehat{z}^k + \gamma_k \lambda_n \partial R(\widehat{z}^k). \tag{7}$$

Let us introduce

$$h_n(w) := (1/2n)\sum_{i=1}^n (\langle w, x_i\rangle - y_i)^2$$

$$\widehat{\xi}^k := \widehat{\varepsilon}^k - \nabla h_n(\widehat{w}^k) + (\langle w, x_{i(k)}\rangle - y_{i(k)}) x_{i(k)},$$

so that (6) and (7) can be rewritten as

$$\widehat{v}^k := \frac{\widehat{w}^k - \widehat{w}^{k+1}}{\alpha_k\gamma_k} - \xi_k - \nabla h_n(\widehat{w}^k) \in \lambda_n \partial R(\widehat{z}^k). \tag{8}$$

The missing block to conclude the proof of Theorem 2 is then the next proposition whose proof is in the supplementary material.

Proposition 3. *Let $n \in \mathbb{N}$, $\lambda_n \in]0,+\infty[$, and let $(\widehat{w}^k)_{k\in\mathbb{N}}$ be generated by the (RSPG) algorithm under assumption ($\mathrm{H_A}$). Then $(\widehat{z}^k, \widehat{v}^k)$ converges almost surely to $(\widehat{w}_{\lambda_n,n}, \widehat{\eta}_{\lambda_n,n})$, as $k \to +\infty$.*

We can now complete the proof of Theorem 2 as follows. In light of Proposition 3, we deduce that there exists $K \in \mathbb{N}$ such that for all $k \geqslant K$,

$$\max\{\|\widehat{w}_{\lambda_n,n} - \widehat{z}^k\|, \|\widehat{\eta}_{\lambda_n,n} - \widehat{v}^k\|\} \leqslant \delta/2 \quad a.s.$$

Without loss of generality, we can assume that the limit of the algorithm is the $\widehat{w}_{\lambda_n,n}$ appearing in (5). The above inequality, combined with (5), allows us to use Proposition 1, and this proves Theorem 2.

5 Numerical illustrations for sparse/low-rank regularization

We give some numerical illustrations of our model consistency results for two popular regularizers: the ℓ_1-norm and the nuclear norm. We generate random problem instances and control the low-complexity of the primal-dual pair of strata $(M_{w_0}, \mathcal{J}_{R^*}(M^*_{\eta_0}))$. The low-complexity of a strata M_w (i.e., the level of low-complexity of w) is measured by

$$\begin{aligned} R_0(M_w) &:= \|w\|_0 \quad \text{for } R = \|\cdot\|_1, \\ R_0(M_w) &:= \operatorname{rank}(w) \quad \text{for } R = \|\cdot\|_*. \end{aligned}$$

Observe that R_0 is well defined, since it does not depend of the choice of w in the strata (see Example 1).

Setup. The instances are randomly generated as follows. For $R = \|\cdot\|_1$, w_0 is drawn randomly among sparse vectors with sparsity level $R_0(M_{w_0}) = \|w_0\|_0 = s$, and we take $(p, n, s, \lambda) = (100, 50, 10, 0.2)$. For $R = \|\cdot\|_*$, w_0 is drawn randomly among low-rank matrices with rank $R_0(M_{w_0}) = \operatorname{rank}(w_0) = s$, and we take $(p, n, s, \lambda) = (20 \times 20, 300, 4, 0.03)$. The features $(x_i)_{i=1}^n$ are drawn at random in $\mathbb{R}^p$ with i.i.d. entries from a zero-mean standard Gaussian distribution. We take y_i as $\langle w_0, x_i \rangle$, to which we add a zero-mean white Gaussian noise with standard deviation 10^{-2}. We compute η_0 with an interior point solver, from which we deduce the upper-bound $R_0(\mathcal{J}_{R^*}(M^*_{\eta_0}))$.

FB vs. Prox-SGD vs. SAGA. First, we compare the deterministic Forward-Backward (FB) algorithm, the Prox-SGD method and SAGA on a simple instance of ($\mathrm{P}_{\lambda,n}$). All algorithms are run with $\alpha_k \equiv 1$, and we take $\gamma_k \equiv 1.8/L_n$ for the FB algorithm, $\gamma_k = 10/(k + 3 \times 10^4)$, and $\gamma_k \equiv 1/(3L'_n)$ for SAGA, where L_n is defined in (H_A) and $L'_n := \max_i \|x_i\|^2$. Figure 1 depicts the evolution of $R_0(M_{\widehat{w}^k})$ while running these three algorithms on ($\mathrm{P}_{\lambda,n}$). At each iteration, FB visits all the data at once, while Prox-SGD and SAGA need only one data. To fairly compare these three algorithms, we plot only the iterates at every batch (i.e. all iterates for FB, and one every n iterates for the stochastic algorithms).

As expected, the two stochastic algorithms exhibit an oscillating behaviour. But for SAGA, these oscillations are damped quickly, and the support of $\widehat{w}^k$ stabilizes after a finite number of iterations. On the contrary, Prox-SGD suffers from constant variations of the support, and is unable to generate iterates with a sparse support. Another observation is that FB and SAGA identify a support which is larger than the one of w_0 but below the extended one governed by η_0, which is in agreement with Theorem 2. A natural question is then: if we replace w_0 by another low-complexity vector, and consider other data, what can be said about the complexity of the obtained solution? This is discussed next.

Randomized experiments for SAGA. We now focus on SAGA, and look at the strata that its iterates can identify. For the ℓ_1 norm (resp. nuclear norm), we draw 1000 (resp. 200) realizations of $(w_0, (x_i, y_i)_{i=1}^n)$ exactly as before. For each realization, we compute η_0 with high precision by using a solver. We then select among the realizations those for which $R_0(\mathcal{J}_{R^*}(M^*_{\eta_0}))$ belongs exactly to $\{10, 20\}$ for the ℓ_1 norm (resp. to $\{4, 7\}$ for the nuclear norm), and we apply the SAGA algorithm to these. The evolution of $R_0(M_{\widehat{w}^k})$ in these cases are plotted in Figure 2.

We see that for the realizations for which $R_0(\mathcal{J}_{R^*}(M^*_{\eta_0})) = R_0(M_{w_0})$ (the blue curves), the algorithm indeed identifies in finite time the stratum where w_0 belongs. Otherwise, we see that the algorithm often identifies a stratum of the same dimension as that of $R_0(\mathcal{J}_{R^*}(M^*_{\eta_0}))$, or sometimes smaller, but which is always larger than M_{w_0}. These observations are consistent with the predictions of Theorem 2.

6 Conclusion

In this paper, we provided a fine and unified analysis for studying model stability/consistency, when considering empirical risk minimization with a mirror-stratifiable regularizer, and solving it with a stochastic algorithm. We showed that, even in the absence of the irrepresentable condition, the low-complexity of an approximate empirical solution remains controlled by a dual certificate. Moreover, we proposed a general algorithmic framework in which stochastic algorithms inherit almost surely finite activity identification.

Acknowledgements

J. Fadili was partly supported by Institut Universitaire de France. The work of G. Garrigos and G. Peyré have been supported by the European Research Council (ERC project NORIA). Part of this work was realized while G. Garrigos was a postdoc at the ENS Paris.

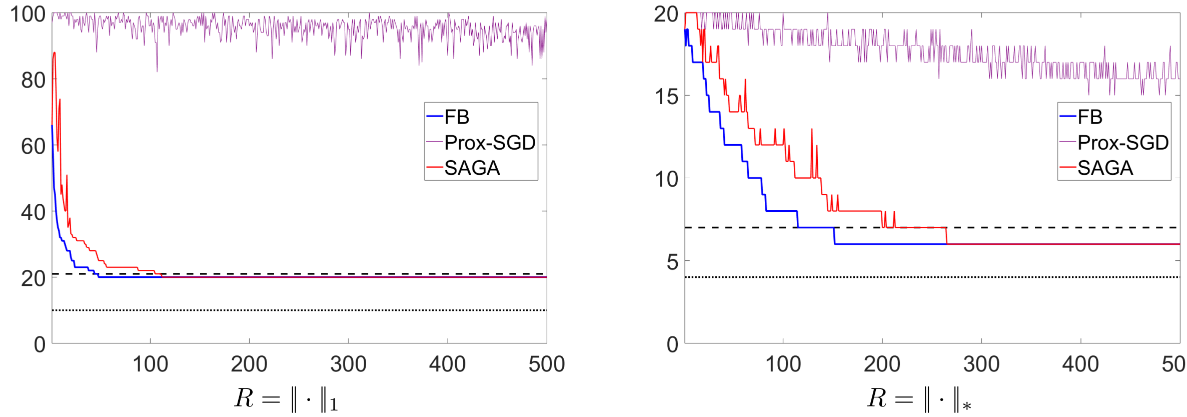

Figure 1: Evolution of $R_0(M_{\widehat{w}^k})$ along the batches of the FB (blue), Prox-SGD (purple) and SAGA (red) algorithms, applied to solve $(\mathrm{P}_{\lambda,n})$. The black dotted line (resp. black dashed line) indicates the value $R_0(M_{w_0})$ (resp. the value of $R_0(\mathcal{J}_{R^*}(M^*_{\eta_0}))$); these are the dimensions of the two extreme strata.

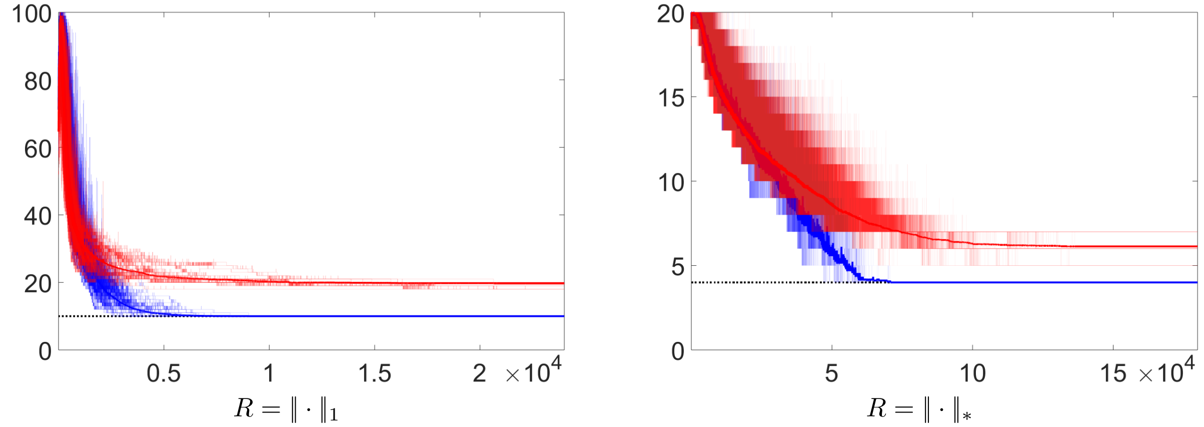

Figure 2: Evolution of $R_0(M_{\widehat{w}^k})$ along the iterations of SAGA, for problems where $R_0(\mathcal{J}_{R^*}(M^*_{\eta_0})) = R_0(M_{w_0}) + \delta$. Blue trajectories correspond to problems for which $\delta = 0$, and for red trajectories $\delta = 10$ (left) or 3 (right). The thick lines correspond to averaged trajectories. The black dotted line indicates the value $R_0(M_{w_0})$.

References

A. Defazio, F. B. and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*.

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

Atchadé, Y. F., Fort, G., and Moulines, E. (2017). On perturbed proximal gradient algorithms. *J. Mach. Learn. Res*, 18(1):310–342.

Auslender, A. and Teboulle, M. (2003). *Asymptotic Cones and Functions in Optimization and Variational Inequalities.* Springer.

Bach, F. (2008a). Consistency of the group Lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9(Jun):1179–1225.

Bach, F. (2008b). Consistency of trace norm minimization. *The Journal of Machine Learning Research*, 9(Jun):1019–1048.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces.* Springer.

Candes, E. and Recht, B. (2013). Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589.

Combettes, P. L. and Pesquet, J.-C. (2015). Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248.

Combettes, P. L. and Pesquet, J.-C. (2016). Stochastic approximations and perturbations in forward-backward splitting for monotone operators. *Pure and Applied Functional Analysis*, 1(1):13–37.

Daniilidis, A., Drusvyatskiy, D., and Lewis, A. S.

(2014). Orthogonal invariance and identifiability. *SIAM Journal on Matrix Analysis and Applications*, 35(2):580–598.

Duval, V. and Peyré, G. (2017). Sparse regularization on thin grids i: the lasso. *Inverse Problems*, 33(5):055008.

Fadili, J., Malick, J., and Peyré, G. (2017). Sensitivity analysis for mirror-stratifiable convex functions. *arXiv preprint arXiv:1707.03194*.

Fazel, M. (2002). *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University.

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

Poon, C., Liang, J., and Schönlieb, C.-B. (2018). Local convergence properties of saga/prox-svrg and acceleration. *arXiv:1802.02554*.

Rosasco, L., Villa, S., and Vũ, B. C. (2016). A stochastic inertial forward–backward splitting algorithm for multivariate monotone inclusions. *Optimization*, 65(6):1293–1314.

Stewart, G. (1977). On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM review*, 19(4):634–662.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Vaiter, S., Peyré, G., and Fadili, J. (2014). Model consistency of partly smooth regularizers. Preprint 00987293, HAL. to appear in IEEE Trans. Inf. Theory.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

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

Xiao, L. and Zhang, T. (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075.

Yuan, M. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563.