Regularized Learning Schemes in Feature Banach Spaces*†

Patrick L. Combettes[1], Saverio Salzo[2], and Silvia Villa[3]

[1]North Carolina State University
Department of Mathematics
Raleigh, NC 27695-8205, USA
plc@math.ncsu.edu

[2]Massachusetts Institute of Technology and Istituto Italiano di Tecnologia
Laboratory for Computational and Statistical Learning
Cambridge, MA 02139, USA
saverio.salzo@iit.it

[3] Politecnico di Milano
Dipartimento di Matematica
20133 Milano, Italy
silvia.villa@polimi.it

Abstract

This paper proposes a unified framework for the investigation of constrained learning theory in reflexive Banach spaces of features via regularized empirical risk minimization. The focus is placed on Tikhonov-like regularization with totally convex functions. This broad class of regularizers provides a flexible model for various priors on the features, including in particular hard constraints and powers of Banach norms. In such context, the main results establish a new general form of the representer theorem and the consistency of the corresponding learning schemes under general conditions on the loss function, the geometry of the feature space, and the modulus of total convexity of the regularizer. In addition, the proposed analysis gives new insight into basic tools such as reproducing Banach spaces, feature maps, and universality. Even when specialized to Hilbert spaces, this framework yields new results that extend the state of the art.

Keywords. consistency, Banach spaces, empirical risk, feature map, reproducing kernel, regularization, representer theorem, statistical learning, totally convex function.

MSC 2010 subject classifications. 62G08, 46E22, 46N30, 68T05, 60B11.

*The work of P. L. Combettes was supported by the CNRS MASTODONS project under grant 2016TABASCO.

†Contact author: P. L. Combettes, plc@math.ncsu.edu, phone: +1 (919) 515 2671.

1 Introduction

A common problem arising in decision sciences is to infer a functional relation from the observation of a finite number of realizations $(x_i, y_i)_{1\leqslant i\leqslant n}$ of random input/output samples from an unknown common distribution P [28, 41, 65, 68]. Given a loss function ℓ and a set $\mathcal{C}$ of functions from the input set $\mathcal{X}$ to the output set $\mathcal{Y}$, the problem is formalized as follows

$$\underset{f\in\mathcal{C}}{\text{infimize}}\;\; R(f), \qquad R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, y, f(x))P(\mathrm{d}(x, y)). \tag{1.1}$$

Since P is not known, the goal is to devise a *consistent learning scheme*, that is, a rule that assigns to each sample $(x_i, y_i)_{1\leqslant i\leqslant n}$ an estimator $f_n \in \mathcal{C}$ such that, $R(f_n) \to \inf R(\mathcal{C})$ as n becomes arbitrarily large.

In this paper, we consider estimators defined by Tikhonov-like regularization. Given an empirical approximation R_n of the *risk* R and a parameter space $\mathcal{F}$, we consider a hypothesis space of functions from $\mathcal{X}$ to $\mathcal{Y}$ described through a linear operator $A\colon \mathcal{F} \to \mathcal{Y}^{\mathcal{X}}$. An estimator is defined by the problem

$$\underset{u\in\mathcal{F}}{\text{minimize}}\;\; R_n(Au) + \lambda_n G(u), \tag{1.2}$$

where $(\lambda_n)_{n\in\mathbb{N}}$ in $\mathbb{R}_{++}$ is a vanishing sequence and $G\colon \mathcal{F} \to [0, +\infty]$ is a *regularizer*, that is, a function modeling some known properties of the target. The above approach is classical, and related to the theory of regularized M-estimators [66] and regularized empirical risk minimization [68]. Many popular learning algorithms are off-springs of this approach, including support vector machines, ridge regression, and sparsity based methods [68, 81], to name a few.

The goal of this paper is to study the theoretical properties of a large family of learning schemes of the form (1.2) designed for problem (1.1). In particular, we consider very general forms of constraint sets, parameterizations of the hypothesis space, and regularizers. Flexibility in the choice of these quantities plays a crucial role in the incorporation of the information potentially available on the problem at hand. More precisely, we assume $\mathcal{C}$ to be a large set of functions defined by pointwise constraints on the function values, e.g., the set of positive functions, the parameter (*feature*) space $\mathcal{F}$ to be a reflexive Banach space, and the regularizer to be a totally convex function. Moreover, we take $\mathcal{Y}$ as a subset of a Banach space, so as to deal with multi-task learning [35, 80] and regression with functional response [55, 37]. Within this context, our contribution is twofold: we analyze the variational problem (1.2), characterizing the form of its solutions, and establish sufficient conditions for the consistency of the corresponding estimators.

Problem in (1.2) is usually analyzed in reproducing kernel Hilbert spaces. Indeed, in this setting, the characterization of the form of the minimizers is well known and is typically referred to as representer theorem [43, 58]. It provides explicit expressions for the minimizers in terms of the corresponding reproducing kernel [31]. The case of hypothesis spaces which are Banach spaces is much less studied; see, e.g., [78, 79, 80]. A first contribution of our paper is to further develop these studies providing a refined analysis of the reproducing property in reflexive Banach spaces, considering also the question of universality [22, 23, 52] in the presence of constraints. A crucial difference with respect to the Hilbert space setting is that in Banach spaces, feature maps, rather than the kernel, become the natural quantities to study the problem, since the kernel may even not exist. Indeed, we prove a new form of the representer theorem for general probability measures and extended-valued convex regularizers that characterizes the minimizers in terms of the feature map, the subgradient of the loss, and the subgradient of the regularizer. Moreover, we show that the computation of the solution of (1.2) can be reduced to that of the dual optimization problem, which is finite dimensional and convex. This fact can be quite helpful in making Banach space problems more practical numerically, in contrast with, for instance, the results in [80] that lead to solving a nonlinear system of equations.

Regarding the statistical analysis, our primary concern is to provide minimal but explicit conditions on problem (1.2) to ensure consistency of its minimizers with respect to problem (1.1). For that purpose, a stability approach [30, 62] turns out to be natural. Indeed, while different strategies can be considered, e.g., based on covering numbers, fat-shattering dimensions [28, 71, 72], or Rademacher complexities [7, 49], these results provide conditions in terms of the complexity measures that need to be made explicit. As we comment later in the paper (see Remark 4.9), in the general setting considered here, this turns out to be a problem in its own right; moreover, stronger assumptions on the probability measure and on the loss are usually required. Finally, we note that approaches using Rademacher complexities do not seem to be applicable outside of the setting of Euclidean space-valued functions and separable losses, since no suitable comparison principle [45, Theorem 4.12] exists.

Our stability approach allows us to bypass these difficulties and directly obtain explicit conditions under the general assumptions outlined above. More precisely, our statistical analysis is based on a sensitivity theorem characterizing the dependence of the solution of problem (1.2) on the underlying probability measure. The analysis is conducted in terms of the feature map and it relies on various tools of convex analysis, geometry of Banach spaces, and probability in Banach spaces. The modulus of total convexity of the regularizer G and the Rademacher type of the dual of $\mathcal{F}$ play a key role, but we remark that the existence of the kernel is not required. Overall, we establish a non trivial extension of the approach in Hilbert spaces considered in [30, 62].

The contributions of the paper are the following.

- We consider a constrained risk minimization problem and a general form of learning schemes based on Tikhonov-like regularization with totally convex regularizers and Banach function spaces of Banach space-valued functions.
- We advance the theory of reproducing Banach spaces and study the problem of universality under constraints.
- We analyze the variational problem defining Tikhonov regularization, and provide a novel characterization of its solutions, generalizing previous forms of the representer theorem.
- We provide minimal explicit sufficient conditions for consistency using a stability argument.

Notation is provided in Section 2. Section 3 is devoted to the study of Banach spaces of vector-valued functions and their description by operator-valued feature maps; universality is studied in the presence of constraints and the representer and sensitivity theorems are established. In Section 4, the regularized learning scheme is formalized and the main consistency theorems are presented. Finally, the Appendix contains technical results on the Lipschitz continuity of convex functions, totally convex functions, Tikhonov-like regularization, and concentration inequalities in Banach spaces.

2 Notation and basic facts

We set $\mathbb{R}_+ = [0, +\infty[$ and $\mathbb{R}_{++} =]0, +\infty[$. Let $\mathcal{B} \neq \{0\}$ be a real Banach space. The closed ball of $\mathcal{B}$ of radius $\rho \in \mathbb{R}_{++}$ centered at the origin is denoted by $B(\rho)$. Let $p \in [1, +\infty]$. The conjugate of p is

$$p^* = \begin{cases} +\infty & \text{if } \ p = 1 \\ p/(p-1) & \text{if } \ 1 < p < +\infty \\ 1 & \text{if } \ p = +\infty. \end{cases} \tag{2.1}$$

Convex Analysis

Let $F \colon \mathcal{B} \to]-\infty, +\infty]$. The *domain of* F is $\operatorname{dom} F = \big\{u \in \mathcal{B} \;\big|\; F(u) < +\infty\big\}$ and F is *proper* if $\operatorname{dom} F \neq \varnothing$. Suppose that F is proper and convex. The *Moreau subdifferential* of F is the set-valued operator

$$\partial F \colon \mathcal{B} \to 2^{\mathcal{B}^*} \colon u \in \mathcal{B} \mapsto \big\{u^* \in \mathcal{B}^* \;\big|\; (\forall v \in \mathcal{B})\ F(u) + \langle v - u, u^* \rangle \leqslant F(v)\big\}, \tag{2.2}$$

and its domain is $\operatorname{dom} \partial F = \big\{u \in \mathcal{B} \;\big|\; \partial F(u) \neq \varnothing\big\}$. Moreover, for every $(u, v) \in \operatorname{dom} F \times \mathcal{B}$, we set $F'(u; v) = \lim_{t \to 0^+} (F(u + tv) - F(u))/t$. If F is proper and bounded from below and $\mathcal{C} \subset \mathcal{B}$ is such that $\mathcal{C} \cap \operatorname{dom} F \neq \varnothing$, we put $\operatorname{Argmin}_{\mathcal{C}} F = \big\{u \in \mathcal{C} \;\big|\; F(u) = \inf F(\mathcal{C})\big\}$, and when it is a singleton we denote by $\operatorname{argmin}_{\mathcal{C}} F$ its unique element. Moreover, we set

$$(\forall \epsilon \in \mathbb{R}_{++}) \quad \operatorname{Argmin}^{\epsilon}_{\mathcal{C}} F = \big\{u \in \mathcal{C} \;\big|\; F(u) \leqslant \inf F(\mathcal{C}) + \epsilon\big\}. \tag{2.3}$$

We denote by $\Gamma_0(\mathcal{B})$ the class of functions $F \colon \mathcal{B} \to]-\infty, +\infty]$ which are proper, convex, and lower semicontinuous. We set $\Gamma_0^+(\mathcal{B}) = \big\{F \in \Gamma_0(\mathcal{B}) \;\big|\; F \geqslant 0\big\}$.

Geometry of Banach spaces

We say that $\mathcal{B}$ is of Rademacher *type* $q \in [1, 2]$ [46, Definition 1.e.12] if there exists $T \in [1, +\infty[$, so that for every $n \in \mathbb{N} \smallsetminus \{0\}$ and $(u_i)_{1 \leqslant i \leqslant n}$ in $\mathcal{B}$,

$$\int_0^1 \Big\| \sum_{i=1}^n r_i(t) u_i \Big\|^q \mathrm{d}t \leqslant T \Big(\sum_{i=1}^n \|u_i\|^q \Big)^{1/q}, \tag{2.4}$$

where $(r_i)_{i \in \mathbb{N}}$ denote the Rademacher functions, that is, for every $i \in \mathbb{N}$, $r_i \colon [0, 1] \to \{-1, 1\} \colon t \mapsto \operatorname{sign}(\sin(2^i \pi t))$. The smallest T for which (2.4) holds is denoted by T_q. Since every Banach space is of Rademacher type 1, this notion is of interest for $q \in]1, 2]$. Moreover, a Banach space of Rademacher type $q \in]1, 2]$ is also of Rademacher type $p \in]1, q[$.

The Banach space $\mathcal{B}$ is called *smooth* [25] if, for every $u \in \mathcal{B}$ there exists a unique $u^* \in \mathcal{B}^*$ such that $\|u^*\| = 1$ and $\langle u, u^* \rangle = 1$. The smoothness property is equivalent to the Gâteaux differentiability of the norm on $\mathcal{B} \setminus \{0\}$. We say that $\mathcal{B}$ is *strictly convex* if, for every u and every v in $\mathcal{B}$ such that $\|u\| = \|v\| = 1$ and $u \neq v$, one has $\|(u + v)/2\| < 1$. The *modulus of convexity of* $\mathcal{B}$ is

$$\begin{aligned} \delta_{\mathcal{B}} \colon\ ]0, 2] &\to \mathbb{R}_+ \\ \varepsilon &\mapsto \inf \Big\{ 1 - \Big\| \frac{u + v}{2} \Big\| \;\Big|\; (u, v) \in \mathcal{B}^2, \|u\| = \|v\| = 1, \|u - v\| \geqslant \varepsilon \Big\}, \end{aligned} \tag{2.5}$$

and the *modulus of smoothness of $\mathcal{B}$* is

$$\begin{aligned}\rho_{\mathcal{B}}\colon \mathbb{R}_+ &\to \mathbb{R}_+\\ \tau &\mapsto \sup\Big\{\frac{1}{2}\big(\|u+v\|+\|u-v\|\big)-1 \,\Big|\, (u,v)\in\mathcal{B}^2, \|u\|=1,\ \|v\|\leqslant\tau\Big\}.\end{aligned} \tag{2.6}$$

We say that $\mathcal{B}$ is *uniformly convex* if $\delta_{\mathcal{B}}$ vanishes only at zero, and *uniformly smooth* if $\lim_{\tau\to 0}\rho_{\mathcal{B}}(\tau)/\tau = 0$ [11, 46]. Now let $q\in[1,+\infty[$. Then $\mathcal{B}$ *has modulus of convexity of power type* q if there exists $c\in\mathbb{R}_{++}$ such that, for every $\varepsilon\in\,]0,2]$, $\delta_{\mathcal{B}}(\varepsilon)\geqslant c\varepsilon^q$, and it *has modulus of smoothness of power type* q if there exists $c\in\mathbb{R}_{++}$ such that, for every $\tau\in\mathbb{R}_{++}$, $\rho_{\mathcal{B}}(\tau)\leqslant c\tau^q$ [11, 46]. A smooth Banach space with modulus of smoothness of power type q is of Rademacher type q [46, Theorem 1.e.16]. Therefore, the notion of Rademacher type is weaker than that of uniform smoothness of power type, in particular it does not imply reflexivity (see the discussion after [46, Theorem 1.e.16]).

If $p\in\,]1,+\infty[$, the *p-duality map* of $\mathcal{B}$ is $J_{\mathcal{B},p}=\partial(\|\cdot\|^p/p)$ [25], and hence

$$(\forall u\in\mathcal{B})\quad J_{\mathcal{B},p}(u)=\big\{u^*\in\mathcal{B}^* \,\big|\, \langle u,u^*\rangle=\|u\|^p \quad\text{and}\quad \|u^*\|=\|u\|^{p-1}\big\}. \tag{2.7}$$

For every $u\in\mathcal{B}$ and every $\lambda\in\mathbb{R}_+$, $J_{\mathcal{B},p}(\lambda u)=\lambda^{p-1}J_{\mathcal{B},p}(u)$ and $J_{\mathcal{B},p}(-u)=-J_{\mathcal{B},p}(u)$. For $p=2$ we obtain the *normalized duality map* $J_{\mathcal{B}}$. Moreover, if $\mathcal{B}$ is reflexive, strictly convex, and smooth, then $J_{\mathcal{B},p}$ is single-valued and its unique selection, which we denote also by $J_{\mathcal{B},p}$, is a bijection from $\mathcal{B}$ onto $\mathcal{B}^*$ and $J_{\mathcal{B}^*,p^*}=J_{\mathcal{B},p}^{-1}$.

Totally convex functions

Totally convex functions, were introduced in [17] and further studied in [18, 19, 77]. This notion lies between strict convexity and strong convexity. Suppose that $\mathcal{B}$ is reflexive and let $F\colon\mathcal{B}\to\,]-\infty,+\infty]$ be a proper convex function. The *modulus of total convexity of F* [20] is

$$\begin{aligned}&\psi\colon \operatorname{dom}F\times\mathbb{R}\to[0,+\infty]:\\ &\qquad (u,t)\mapsto \inf\big\{F(v)-F(u)-F'(u;v-u) \,\big|\, v\in\operatorname{dom}F,\ \|v-u\|=t\big\}\end{aligned} \tag{2.8}$$

and F is *totally convex at* $u\in\operatorname{dom}F$ if, for every $t\in\mathbb{R}_{++}$, $\psi(u,t)>0$. The function F is *totally convex* if it is totally convex at every point of its domain. Let ψ be the modulus of total convexity of F. For every $\rho\in\mathbb{R}_{++}$ such that $B(\rho)\cap\operatorname{dom}F\neq\varnothing$, the *modulus of total convexity of F on $B(\rho)$* is

$$\psi_\rho\colon\mathbb{R}\to[0,+\infty]:t\mapsto \inf_{u\in B(\rho)\cap\operatorname{dom}F}\psi(u,t), \tag{2.9}$$

and F is *totally convex on $B(\rho)$* if $\psi_\rho>0$ on $\mathbb{R}_{++}$. Moreover, F is *totally convex on bounded sets* if, for every $\rho\in\mathbb{R}_{++}$ such that $B(\rho)\cap\operatorname{dom}F\neq\varnothing$, it is totally convex on $B(\rho)$. Let $\phi\colon\mathbb{R}\to[0,+\infty]$ be such that $\phi(0)=0$ and $\operatorname{dom}\phi\subset\mathbb{R}_+$. We set

$$\widehat{\phi}\colon\mathbb{R}\to[0,+\infty]:t\mapsto\begin{cases}0 & \text{if } t=0\\ \phi(t)/|t| & \text{if } t\neq 0.\end{cases} \tag{2.10}$$

The upper-quasi inverse of ϕ is [53, 77]

$$\phi^\natural\colon\mathbb{R}\to[0,+\infty]:s\mapsto\begin{cases}\sup\big\{t\in\mathbb{R}_+ \,\big|\, \phi(t)\leqslant s\big\} & \text{if } s\geqslant 0\\ +\infty & \text{if } s<0.\end{cases} \tag{2.11}$$

Note that, for every $(t,s) \in \mathbb{R}_+^2$, $\phi(t) \leqslant s \Rightarrow t \leqslant \phi^\natural(s)$. We set

$$\mathcal{A}_0 = \big\{\phi\colon \mathbb{R} \to [0,+\infty] \,\big|\, \operatorname{dom}\phi \subset \mathbb{R}_+,\ \phi \text{ is increasing on } \mathbb{R}_+, \\ \phi(0) = 0, (\forall t \in \mathbb{R}_{++})\ \phi(t) > 0\big\} \quad (2.12)$$

and

$$\mathcal{A}_1 = \Big\{\phi \in \mathcal{A}_0 \,\Big|\, \widehat{\phi} \text{ is increasing on } \mathbb{R}_+,\ \lim_{t\to 0^+} \widehat{\phi}(t) = 0\Big\}. \quad (2.13)$$

Suppose that F is totally convex at $u \in \operatorname{dom} F$. Then $\psi(u,\cdot) \in \mathcal{A}_0$ and $\psi(u,\cdot)\widehat{\ } \in \mathcal{A}_0$. Moreover, if additionally $\partial F(u) \neq \varnothing$, then $\psi(u,\cdot)\widehat{\ } \in \mathcal{A}_1$. Suppose that $\mathcal{B}$ is uniformly convex with power type, then, for every $r \in \mathbb{R}_{++}$, $\|\cdot\|^r$ is totally convex on bounded sets (See Appendix A.2).

Lebesgue spaces of vector-valued and operator-valued functions

When a Banach space is regarded as a measurable space it is with respect to its Borel σ-algebra. Let $(\mathcal{Z},\mathfrak{A},\mu)$ be a σ-finite measure space and let Y be a separable real Banach space with norm $|\cdot|$. We denote by $\mathcal{M}(\mathcal{Z},\mathsf{Y})$ the set of measurable functions from $\mathcal{Z}$ into Y. If $p \neq +\infty$, $L^p(\mathcal{Z},\mu;\mathsf{Y})$ is the Banach space of all (equivalence classes of) measurable functions $f \in \mathcal{M}(\mathcal{Z},\mathsf{Y})$ such that $\int_{\mathcal{Z}} |f|^p \mathrm{d}\mu < +\infty$ and $L^\infty(\mathcal{Z},\mu;\mathsf{Y})$ is the Banach space of all (equivalence classes of) measurable functions $f \in \mathcal{M}(\mathcal{Z},\mathsf{Y})$ which are μ-essentially bounded. Let $f \in L^p(\mathcal{Z},\mu;\mathsf{Y})$. Then $\|f\|_p = \big(\int_{\mathcal{Z}} |f|^p \mathrm{d}\mu\big)^{1/p}$ if $p \neq +\infty$, and $\|f\|_\infty = \mu\text{-}\operatorname{ess\text{-}sup}_{z\in\mathcal{Z}} |f(z)|$ otherwise. If $p \in \left]1,+\infty\right[$, $L^p(\mathcal{Z},\mu;\mathbb{R})$ is uniformly convex and uniformly smooth, and it has modulus of convexity of power type $\max\{2,p\}$, and modulus of smoothness of power type $\min\{2,p\}$ [46, p. 63] and hence it is of Rademacher type $\min\{2,p\}$. If $\mathcal{Z}$ is countable, $\mathfrak{A} = 2^{\mathcal{Z}}$, and μ is the counting measure, we set $l^p(\mathcal{Z};\mathsf{Y}) = L^p(\mathcal{Z},\mu;\mathsf{Y})$ and $l^p(\mathcal{Z}) = L^p(\mathcal{Z},\mu;\mathbb{R})$. Let Y and Z be separable real Banach spaces. We denote by $\mathscr{L}(\mathsf{Y},\mathsf{Z})$ the Banach space of continuous linear operators from Y into Z endowed with the operator norm. A map $\Phi\colon \mathcal{Z} \to \mathscr{L}(\mathsf{Y},\mathsf{Z})$ is *strongly measurable* if, for every $y \in \mathsf{Y}$, the function $\mathcal{Z} \to \mathsf{Z}\colon z \mapsto \Phi(z)y$ is measurable. In such a case the function $\mathcal{Z} \to \mathbb{R}\colon z \mapsto \|\Phi(z)\|$ is measurable [34]. If $p \neq +\infty$, $L^p[\mathcal{Z},\mu;\mathscr{L}(\mathsf{Y},\mathsf{Z})]$ is the Banach space of all (equivalence classes of) strongly measurable functions $\Phi\colon \mathcal{Z} \to \mathscr{L}(\mathsf{Y},\mathsf{Z})$ such that $\int_{\mathcal{Z}} \|\Phi(z)\|^p \mu(\mathrm{d}z) < +\infty$ and $L^\infty[\mathcal{Z},\mu;\mathscr{L}(\mathsf{Y},\mathsf{Z})]$ is the Banach space of all (equivalence classes of) strongly measurable functions $\Phi\colon \mathcal{Z} \to \mathscr{L}(\mathsf{Y},\mathsf{Z})$ such that $\mu\text{-}\operatorname{ess\text{-}sup}_{z\in\mathcal{Z}} \|\Phi(z)\| < +\infty$ [10]. Let $\Phi \in L^p[\mathcal{Z},\mu;\mathscr{L}(\mathsf{Y},\mathsf{Z})]$. Then $\|\Phi\|_p = \big(\int_{\mathcal{Z}} \|\Phi(z)\|^p \mu(\mathrm{d}z)\big)^{1/p}$ if $p \neq +\infty$, and $\|\Phi\|_\infty = \mu\text{-}\operatorname{ess\text{-}sup}_{z\in\mathcal{Z}} \|\Phi(z)\|$ otherwise.

Probability

Let $(\Omega,\mathfrak{A},\mathsf{P})$ be a probability space, let P^* be the associated outer probability. For every $\xi\colon \Omega \to \mathbb{R}$ and $t \in \mathbb{R}$, we set

$$[\xi > t] = \big\{\omega \in \Omega \,\big|\, \xi(\omega) > t\big\}; \quad (2.14)$$

the sets $[\xi < t]$, $[\xi \geqslant t]$, and $[\xi \leqslant t]$ are defined analogously. Let $(U_n)_{n\in\mathbb{N}}$ and U be functions from Ω to $\mathcal{B}$. The sequence $(U_n)_{n\in\mathbb{N}}$ converges *in* P-*outer probability* to U, in symbols $U_n \xrightarrow{\mathsf{P}^*} U$, if [67]

$$(\forall \varepsilon \in \mathbb{R}_{++}) \quad \mathsf{P}^*\big[\|U_n - U\| > \varepsilon\big] \to 0, \quad (2.15)$$

and it converges P^*-almost surely (a.s.) to U if

$$(\exists\, \Omega_0 \subset \Omega) \quad \mathsf{P}^*\Omega_0 = 0 \quad \text{and} \quad (\forall \omega \in \Omega \smallsetminus \Omega_0) \quad U_n(\omega) \to U(\omega). \quad (2.16)$$

The probability space $(\Omega,\mathfrak{A},\mathsf{P})$ is *complete* if, for every $A \in \mathfrak{A}$ such that $\mathsf{P}(A) = 0$, and every $B \subset A$, we have $B \in \mathfrak{A}$.

3 Learning in Banach spaces

Basic tools such as feature maps, reproducing kernel Hilbert spaces, and representer theorems have played an instrumental role in the development of Hilbertian learning theory [42, 58, 62]. In recent years, there has been a marked interest in extending these tools to Banach spaces; see for instance [36, 78, 80] and references therein. The primary objective of this section is to further develop the theory on these topics.

3.1 Banach spaces of vector-valued functions and feature map representations

Sampling based nonparametric estimation naturally calls for formulations involving spaces of functions for which the pointwise evaluation operator is continuous. In the Hilbert space setting, this framework hinges on the notions of a reproducing kernel Hilbert space and of a feature map, which have been extensively investigated, e.g., in [21, 62]. On the other hand, the study of reproducing kernel Banach spaces has been developed primarily in [78, 80]. However, in the Banach space setting, the continuity of the pointwise evaluation operators, the existence of a kernel, and the existence of a feature map may no longer be equivalent and further investigation is in order. Towards this goal, we start with the following proposition which extends [21, Proposition 2.4].

Proposition 3.1 *Let $\mathcal{X}$ be a nonempty set, let Y and $\mathcal{F}$ be separable real Banach spaces, and let $A\colon \mathcal{F} \to \mathsf{Y}^{\mathcal{X}}$ be a linear operator. Then the following are equivalent:*

(i) *$A\colon \mathcal{F} \to \mathsf{Y}^{\mathcal{X}}$ is continuous for the topology of pointwise convergence on $\mathsf{Y}^{\mathcal{X}}$.*

(ii) *There exists a map $\Phi\colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)$ such that*

$$(\forall u \in \mathcal{F})(\forall x \in \mathcal{X}) \quad (Au)(x) = \Phi(x)^* u. \tag{3.1}$$

(iii) *$\operatorname{ran} A$ can be endowed with a Banach space structure such that the point-evaluation operators on $\operatorname{ran} A$ are continuous, $A\colon \mathcal{F} \to \operatorname{ran} A$ is continuous, and the quotient operator of A is a Banach space isometry from $\mathcal{F}/\ker A$ onto $\operatorname{ran} A$.*

Proof. Set $\mathcal{W} = \operatorname{ran} A$ and $\mathcal{N} = \ker A$. Let $\pi_{\mathcal{N}}\colon \mathcal{F} \to \mathcal{F}/\mathcal{N}\colon u \mapsto u + \mathcal{N}$ be the canonical projection operator and let $\tilde{A}\colon \mathcal{F}/\mathcal{N} \to \mathsf{Y}^{\mathcal{X}}$ be the unique linear map such that $A = \tilde{A} \circ \pi_{\mathcal{N}}$. Then $\tilde{A}$ is injective and $\operatorname{ran} \tilde{A} = \operatorname{ran} A$. Moreover, for every $x \in \mathcal{X}$, we define the point-evaluation operator $\mathrm{ev}_x\colon \mathcal{W} \to \mathsf{Y}\colon f \mapsto f(x)$. We recall that A is continuous for the topology of pointwise convergence on $\mathsf{Y}^{\mathcal{X}}$ if and only if, for every $x \in \mathcal{X}$, $\mathrm{ev}_x \circ A\colon \mathcal{F} \to \mathsf{Y}$ is continuous.

(i)⇒(ii): Set $\Phi\colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)\colon x \mapsto (\mathrm{ev}_x \circ A)^*$.

(ii)⇒(i): Let $x \in \mathcal{X}$. Then, by (3.1), $\mathrm{ev}_x \circ A = \Phi(x)^*$ is continuous.

(i)⇒(iii): Since $\mathcal{N}$ is a closed vector subspace of $\mathcal{F}$, the quotient space $\mathcal{F}/\mathcal{N}$ is a Banach space with the quotient norm $\pi_{\mathcal{N}} u \mapsto \|\pi_{\mathcal{N}} u\|_{\mathcal{F}/\mathcal{N}} = \inf_{v \in \mathcal{N}} \|u - v\|$. Thus, we endow $\mathcal{W}$ with the Banach space structure transported from $\mathcal{F}/\mathcal{N}$ by $\tilde{A}$, i.e., for every $u \in \mathcal{F}$, $\|Au\| = \|\tilde{A}\pi_{\mathcal{N}} u\| = \|\pi_{\mathcal{N}} u\|_{\mathcal{F}/\mathcal{N}}$. Denote by $|\cdot|$ the norm of Y. Let $x \in \mathcal{X}$ and $f \in \mathcal{W}$. Then there exists $u \in \mathcal{F}$ such that $f = Au$, and hence $(\forall v \in \mathcal{N})$ $|f(x)| = |(\mathrm{ev}_x \circ A)(u + v)| \leqslant \|\mathrm{ev}_x \circ A\| \, \|u + v\|$. Taking the infimum over $\mathcal{N}$, and recalling the definition of the quotient norm, we get $|f(x)| \leqslant \|\mathrm{ev}_x \circ A\| \|\pi_{\mathcal{N}} u\|_{\mathcal{F}/\mathcal{N}} = \|\mathrm{ev}_x \circ A\| \|f\|$. Hence, $\mathrm{ev}_x\colon \mathcal{W} \to \mathsf{Y}$ is continuous. Finally, $A\colon \mathcal{F} \to \mathcal{W}$ is continuous since $A = \tilde{A} \circ \pi_{\mathcal{N}}$.

(iii)⇒(i): Let $x \in \mathcal{X}$. Since $A\colon \mathcal{F} \to \mathcal{W}$ is continuous and $\mathrm{ev}_x\colon \mathcal{W} \to \mathsf{Y}$ is continuous, $\mathrm{ev}_x \circ A\colon \mathcal{F} \to \mathsf{Y}$ is likewise. □

Definition 3.2 In the setting of Proposition 3.1, if A is continuous for the topology of pointwise convergence on $\mathsf{Y}^{\mathcal{X}}$, then the unique map Φ defined in (ii) is the *feature map* associated with A and $\mathcal{F}$ is the feature space.

Definition 3.3 Let $\mathcal{X}$ be a nonempty set and let Y be a separable real Banach space. Let $\mathcal{W}$ be a real Banach space of functions from $\mathcal{X}$ to Y. Then

(i) $\mathcal{W}$ is a *pre-reproducing kernel Banach space* if, for every $x \in \mathcal{X}$, the point-evaluation operator $\mathrm{ev}_x\colon \mathcal{W} \to \mathsf{Y}\colon f \mapsto f(x)$ is continuous [60].

(ii) $\mathcal{W}$ is a *reproducing kernel Banach space* if it is a reflexive, strictly convex, and smooth pre-reproducing kernel Banach space.

Remark 3.4

(i) Proposition 3.1 establishes that pre-reproducing kernel Banach spaces can always be built via feature map representations. We note that pre-reproducing kernel Banach spaces are called *function Banach spaces* in [16].

(ii) Equation (3.1) is equivalent to

$$(\forall u \in \mathcal{F})(\forall x \in \mathcal{X})(\forall \mathsf{w}^* \in \mathsf{Y}^*) \quad \langle u, \Phi(x)\mathsf{w}^* \rangle = \langle (Au)(x), \mathsf{w}^* \rangle, \tag{3.2}$$

which shows that A is injective if and only if $\{\Phi(x)\mathsf{w}^* \mid x \in \mathcal{X},\, \mathsf{w}^* \in \mathsf{Y}^*\}$ is dense in $\mathcal{F}^*$. Note that this last denseness condition (hence the injectivity of A) is usually required in the the current literature on reproducing kernel Banach spaces [78, 79, 80]. We do not need this assumption.

Proposition 3.5 *Let $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}}, \mu)$ be a σ-finite measure space, let Y and $\mathcal{F}$ be separable real Banach spaces, let $A\colon \mathcal{F} \to \mathsf{Y}^{\mathcal{X}}$ be linear and continuous for the topology of pointwise convergence on $\mathsf{Y}^{\mathcal{X}}$, and let $\Phi\colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)$ be the associated feature map. Then the following hold:*

(i) *$\Phi\colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)$ is strongly measurable if and only if* $\operatorname{ran} A \subset \mathcal{M}(\mathcal{X}, \mathsf{Y})$.

(ii) *Let $p \in [1, +\infty]$ and suppose that $\Phi \in L^p[\mathcal{X}, \mu; \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)]$. Then* $\operatorname{ran} A \subset L^p(\mathcal{X}, \mu; \mathsf{Y})$ *and, for every $u \in \mathcal{F}$, $\|Au\|_p \leqslant \|\Phi\|_p \|u\|$.*

Proof. (i): It follows from Pettis' theorem [32, Theorem II.2] and (3.2) that $\Phi\colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)$ is strongly measurable if and only if, for every $u \in \mathcal{F}$, Au is measurable.

(ii): Let $u \in \mathcal{F}$ and note that, by (i), Au is measurable. Moreover, by (3.1), $(\forall x \in \mathcal{X})\ |(Au)(x)| = |\Phi(x)^* u| \leqslant \|\Phi(x)\|\, \|u\|$. □

We now define a notion of universality for spaces of vector-valued functions [22, 23] with respect to a constraint set.

Definition 3.6 Let $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}})$ be a measurable space, let Y be a separable uniformly convex real Banach space, and let $\mathcal{W}$ be a vector space of bounded measurable functions from $\mathcal{X}$ to Y. Let $\mathcal{C} \subset \mathcal{M}(\mathcal{X}; \mathsf{Y})$ be a convex set.

(i) $\mathcal{W}$ is ∞*-universal relative to* $\mathcal{C}$ if, for every probability measure μ on $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}})$ and for every $f \in \mathcal{C} \cap L^\infty(\mathcal{X}, \mu; \mathsf{Y})$, there exists $(f_n)_{n\in\mathbb{N}} \in (\mathcal{C} \cap \mathcal{W})^{\mathbb{N}}$ such that $\sup_{n\in\mathbb{N}} \|f_n\|_\infty < +\infty$ and $f_n \to f$ μ-a.e.

(ii) Let $p \in [1, +\infty[$. The space $\mathcal{W}$ is p*-universal relative to* $\mathcal{C}$ if, for every probability measure μ on $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}})$, $\mathcal{C} \cap \mathcal{W}$ is dense in $\mathcal{C} \cap L^p(\mathcal{X}, \mu; \mathsf{Y})$.

When $\mathcal{C} = \mathcal{M}(\mathcal{X}; \mathsf{Y})$ the reference to the set $\mathcal{C}$ is omitted.

Definition 3.7 Let $(\mathsf{Y}, |\cdot|)$ be a real normed vector space. The Attouch-Wets topology [6, 12] on the class $\mathscr{C}_{\mathsf{Y}}$ of nonempty closed subsets of Y is that induced by the following family of pseudometrics

$$(\forall\, \rho \in \mathbb{R}_{++})(\forall\, (\mathsf{C}_1, \mathsf{C}_2) \in \mathscr{C}_{\mathsf{Y}}^2) \quad \mathrm{dist}_\rho(\mathsf{C}_1, \mathsf{C}_2) = \sup_{|\mathsf{w}| \leqslant \rho} |d_{\mathsf{C}_1}(\mathsf{w}) - d_{\mathsf{C}_2}(\mathsf{w})|, \tag{3.3}$$

where $d_{\mathsf{C}}(\mathsf{w}) = \inf_{\mathsf{y}\in\mathsf{C}} |\mathsf{y} - \mathsf{w}|$ is the distance function to the set C.

The following proposition shows that Definition 3.6 is an extension of the standard notion of universality in the context of reproducing kernel Hilbert spaces [23, 52, 62].

Theorem 3.8 *Let $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}})$ be a measurable space, let Y be a separable uniformly convex real Banach space, and let $\mathcal{W}$ be a vector space of bounded measurable functions from $\mathcal{X}$ to Y. Let $(\mathsf{C}(x))_{x\in\mathcal{X}}$ be a family of closed convex subsets of Y containing 0, let $\mathcal{C} = \big\{f \in \mathcal{M}(\mathcal{X}, \mathsf{Y}) \;\big|\; (\forall x \in \mathcal{X})\ f(x) \in \mathsf{C}(x)\big\}$, and let $p \in [1, +\infty[$. Consider the following properties:*

(a) $\mathcal{W}$ is ∞-universal relative to $\mathcal{C}$.

(b) $\mathcal{W}$ is p-universal relative to $\mathcal{C}$.

Then the following hold:

(i) *Suppose that $x \mapsto \mathsf{C}(x)$ is measurable [24]. Then (a)⇒(b).*

(ii) *Suppose that $\mathcal{X}$ is a locally compact Hausdorff space and let $\mathscr{C}_0(\mathcal{X}; \mathsf{Y})$ be the space of continuous functions from $\mathcal{X}$ to Y vanishing at infinity [14]. Suppose that $\mathcal{W} \subset \mathscr{C}_0(\mathcal{X}; \mathsf{Y})$ and that $x \mapsto \mathsf{C}(x)$ is continuous with respect to the Attouch-Wets topology. Consider the following property:*

(c) $\mathcal{C} \cap \mathcal{W}$ is dense in $\mathcal{C} \cap \mathscr{C}_0(\mathcal{X}; \mathsf{Y})$ for the uniform topology.

Then (a)⇔(b)⇔(c).

Proof. (i): Suppose that (a) holds and let μ be a probability measure on $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}})$. We have $\mathcal{W} \subset L^\infty(\mathcal{X}, \mu; \mathsf{Y})$. We derive from (a) and the dominated convergence theorem that $\mathcal{C} \cap \mathcal{W}$ is dense in $\mathcal{C} \cap L^\infty(\mathcal{X}, \mu; \mathsf{Y})$ for the topology of $L^p(\mathcal{X}, \mu; \mathsf{Y})$. Next, let $f \in \mathcal{C} \cap L^p(\mathcal{X}, \mu; \mathsf{Y})$ and let $\epsilon \in \mathbb{R}_{++}$. Since $L^\infty(\mathcal{X}, \mu; \mathsf{Y})$ is dense in $L^p(\mathcal{X}, \mu; \mathsf{Y})$ for the topology of $L^p(\mathcal{X}, \mu; \mathsf{Y})$, there exists $g \in L^\infty(\mathcal{X}, \mu; \mathsf{Y})$ such that $\|f - g\|_p \leqslant \epsilon/2$. The function

$$P_{\mathsf{C}}(g) \colon \mathcal{X} \to \mathsf{Y} \colon x \mapsto P_{\mathsf{C}(x)}(g(x)) \tag{3.4}$$

is well defined [40, Proposition 3.2] and its measurability follows from the application of [24, Lemma III.39] with $\varphi \colon \mathcal{X} \times \mathsf{Y} \to \mathbb{R} \colon (x, y) \mapsto -|y - g(x)|$ and $\Sigma = \mathsf{C} \colon \mathcal{X} \to 2^{\mathsf{Y}}$. Then $P_{\mathsf{C}}(g) \in \mathcal{C}$ and, for every $x \in \mathcal{X}$, since $\{0, f(x)\} \subset \mathsf{C}(x)$,

$$\begin{cases} |P_{\mathsf{C}(x)}(g(x))| \leqslant |P_{\mathsf{C}(x)}(g(x)) - g(x)| + |g(x)| \leqslant 2|g(x)| \\ |P_{\mathsf{C}(x)}(g(x)) - f(x)| \leqslant |P_{\mathsf{C}(x)}(g(x)) - g(x)| + |g(x) - f(x)| \leqslant 2|g(x) - f(x)|. \end{cases} \tag{3.5}$$

Therefore $P_{\mathsf{C}}(g) \in L^\infty(\mathcal{X},\mu;\mathsf{Y})$ and $\|P_{\mathsf{C}}(g) - f\|_p \leqslant 2\|f-g\|_p \leqslant \epsilon$.

(ii): (c)⇒(a): Let μ be a probability measure on $(\mathcal{X},\mathfrak{A}_{\mathcal{X}})$ and let $f \in \mathcal{C} \cap L^\infty(\mathcal{X},\mu;\mathsf{Y})$. We denote by $\mathscr{K}(\mathcal{X};\mathsf{Y})$ the space of continuous functions from $\mathcal{X}$ to Y with compact support. Since $\mathcal{X}$ is completely regular, we derive from Lusin's theorem [33, Corollary 1 in III.§15.8] and Urysohn's lemma, that there exists a sequence $(g_n)_{n\in\mathbb{N}}$ in $\mathscr{K}(\mathcal{X};\mathsf{Y})$ such that $g_n \to f$ μ-a.e. and $\sup_{n\in\mathbb{N}} \|g_n\|_\infty \leqslant \|f\|_\infty$. Let $n \in \mathbb{N}$ and define the function $P_{\mathsf{C}}(g_n)\colon \mathcal{X} \to \mathsf{Y}\colon x \mapsto P_{\mathsf{C}(x)}(g_n(x))$. Let us prove that $P_{\mathsf{C}}(g_n)$ is continuous. Let $x_0 \in \mathcal{X}$. Since $\lim_{x\to x_0} \mathsf{C}(x) = \mathsf{C}(x_0)$ in the Attouch-Wets topology, there exist a neighborhood U_1 of x_0 and $t \in \mathbb{R}_{++}$ such that, for every $x \in U_1$, $\inf|\mathsf{C}(x)| < t$. Moreover there exist a neighborhood U_2 of x_0 and $q \in \mathbb{R}_{++}$ such that, for every $x \in U_2$, $g_n(x) \in B(q)$. Now, fix $r \in [3q+t,+\infty[$. Then, for every $x \in U_1 \cap U_2$, since $r \geqslant 3q + \inf|\mathsf{C}(x)|$, it follows from [53, Corollary 3.3 and Theorem 4.1] that

$$\begin{aligned}
&|P_{\mathsf{C}}(g_n)(x) - P_{\mathsf{C}}(g_n)(x_0)| \\
&\quad\leqslant |P_{\mathsf{C}(x)}(g_n(x)) - P_{\mathsf{C}(x)}(g_n(x_0))| + |P_{\mathsf{C}(x)}(g_n(x_0)) - P_{\mathsf{C}(x_0)}(g_n(x_0))| \\
&\quad\leqslant \phi^\natural\big(2r|g_n(x)-g_n(x_0)|\big) + |g_n(x)-g_n(x_0)| + \phi^\natural(2r\,\mathrm{dist}_{2q+t}(\mathsf{C}(x),\mathsf{C}(x_0))),
\end{aligned} \tag{3.6}$$

where $\phi \in \mathcal{A}_0$ is the modulus of uniform monotonicity of the normalized duality map of Y on $B(r)$, and, for every $\rho \in \mathbb{R}_{++}$, dist_ρ is as in Definition 3.7. Hence, since $\lim_{x\to x_0} \mathrm{dist}_{2q+t}(\mathsf{C}(x),\mathsf{C}(x_0)) = 0$, $\lim_{x\to x_0} |g_n(x)-g_n(x_0)| = 0$, and $\lim_{s\to 0^+} \phi^\natural(s) = 0$ by Proposition A.5(v), the continuity of $P_{\mathsf{C}}(g_n)$ at x_0 follows. In addition, since $0 \in \bigcap_{x\in\mathcal{X}} \mathsf{C}(x)$, the support of $P_{\mathsf{C}}(g_n)$ is contained in that of g_n. Therefore, for every $n \in \mathbb{N}$, $P_{\mathsf{C}}(g_n) \in \mathcal{C} \cap \mathscr{K}(\mathcal{X};\mathsf{Y})$, $\|P_{\mathsf{C}}(g_n)\|_\infty \leqslant 2\|g_n\|_\infty$ and, $(\forall x \in \mathcal{X})$ $|P_{\mathsf{C}(x)}(g_n(x)) - f(x)| \leqslant 2|g_n(x) - f(x)|$. Hence $P_{\mathsf{C}}(g_n) \to f$ μ-a.e. It follows from (c) that, for every $n \in \mathbb{N}$, there exists $f_n \in \mathcal{C} \cap \mathcal{W}$ such that $\|f_n - P_{\mathsf{C}}(g_n)\|_\infty \leqslant 1/(n+1)$. Therefore $\sup_{n\in\mathbb{N}} \|f_n\|_\infty \leqslant \sup_{n\in\mathbb{N}}(1 + \|P_{\mathsf{C}}(g_n)\|_\infty) \leqslant 1 + 2\|f\|_\infty$ and $f_n \to f$ μ-a.e.

(b)⇒(c): We follow the same reasoning as in the proof of [23, Theorem 4.1]. By contradiction, suppose that $\mathcal{C} \cap \mathcal{W}$ is not dense in $\mathcal{C} \cap \mathscr{C}_0(\mathcal{X};\mathsf{Y})$. Since $\mathcal{C} \cap \mathcal{W}$ is nonempty and convex, by the Hahn-Banach theorem, there exists $f_0 \in \mathcal{C} \cap \mathscr{C}_0(\mathcal{X};\mathsf{Y})$ and $\varphi \in \mathscr{C}_0(\mathcal{X};\mathsf{Y})^*$, and $\alpha \in \mathbb{R}$ such that

$$(\forall f \in \mathcal{C} \cap \mathcal{W}) \quad \varphi(f) < \alpha < \varphi(f_0). \tag{3.7}$$

Now, by [33, Corollary 2 and Theorem 5 in III.§19.3] there is a probability measure μ on $\mathcal{X}$ and a function $h \in L^\infty(\mathcal{X},\mu;\mathsf{Y}^*)$ such that

$$(\forall\, f \in \mathcal{C}_0(\mathcal{X};\mathsf{Y})) \quad \varphi(f) = \int_{\mathcal{X}} \langle f(x), h(x)\rangle \mathrm{d}\mu(x)\,. \tag{3.8}$$

Since $\varphi \neq 0$, we have $h \neq 0$. Moreover $h \in L^{p^*}(\mathcal{X},\mu;\mathsf{Y}^*)$. Therefore

$$(\forall\, f \in \mathcal{C} \cap \mathcal{W}) \quad \langle f, h\rangle_{p,p^*} < \alpha < \langle f_0, h\rangle_{p,p^*}. \tag{3.9}$$

Let $H^\alpha_- = \{f \in L^p(\mathcal{X},\mu;\mathsf{Y}) \,|\, \langle f,h\rangle_{p,p^*} \leqslant \alpha\}$. Then H^α_- is a closed half-space of $L^p(\mathcal{X},\mu;\mathsf{Y})$. Therefore, by (3.9), $\overline{\mathcal{C}\cap\mathcal{W}} \subset H^\alpha_-$ and $f_0 \notin H^\alpha_-$. Hence, $\mathcal{C} \cap \mathcal{W}$ is not dense in $\mathcal{C} \cap L^p(\mathcal{X},\mu;\mathsf{Y})$. □

Remark 3.9 The Attouch-Wets topology considered in the statement of Theorem 3.8 is also called bounded Hausdorff topology and is in fact a generalization of the Hausdorff topology to non-compact sets.

In the next proposition we show that in the Banach space setting, the duality map (see Section 2) is instrumental to properly define a kernel. This will require the involved Banach spaces to satisfy additional geometric properties.

Proposition 3.10 *Under the assumptions of Proposition 3.1, let $\Phi\colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)$ be defined by (3.1) and set $\mathcal{W} = \operatorname{ran} A$. Let $\mathscr{B}(\mathsf{Y}^*, \mathsf{Y})$ be the set of operators mapping bounded subsets of Y^* into bounded subsets of Y. Suppose that $\mathcal{F}$ is reflexive, strictly convex, and smooth, and let $p \in\,]1, +\infty[$. Then $\mathcal{W}$ is a reproducing kernel Banach space and there exists a unique $K_p\colon \mathcal{X} \times \mathcal{X} \to \mathscr{B}(\mathsf{Y}^*, \mathsf{Y})$, called kernel, such that*

$$(\forall\, u \in \mathcal{F})(\forall\, x \in \mathcal{X})(\forall\, y^* \in \mathsf{Y}^*) \quad \begin{cases} K_p(x, \cdot)y^* \in \mathcal{W} \\ \langle Au, J_{\mathcal{W},p}(K_p(x, \cdot)y^*)\rangle = \langle (Au)(x), y^*\rangle. \end{cases} \tag{3.10}$$

Moreover, we have

$$(\forall\, x \in \mathcal{X})(\forall\, x' \in \mathcal{X}) \quad K_p(x, x') = \Phi(x')^* \circ J_{\mathcal{F},p}^{-1} \circ \Phi(x). \tag{3.11}$$

Proof. Let $\mathcal{N} = \ker A$. Proposition 3.1 implies that $\mathcal{W}$ is isometrically isomorphic to $\mathcal{F}/\mathcal{N}$. Define

$$K_p\colon \mathcal{X} \times \mathcal{X} \to \mathscr{B}(\mathsf{Y}^*, \mathsf{Y})\colon (x, x') \mapsto \Phi(x')^* \circ J_{\mathcal{F},p}^{-1} \circ \Phi(x)\,. \tag{3.12}$$

Then (3.1) yields

$$(\forall\, x \in \mathcal{X})(\forall\, y^* \in \mathsf{Y}^*) \quad K_p(x, \cdot)y^* = AJ_{\mathcal{F},p}^{-1}(\Phi(x)y^*). \tag{3.13}$$

Since $\mathcal{F}$ is reflexive, strictly convex, and smooth, $\mathcal{F}/\mathcal{N}$ and $\mathcal{W}$ are likewise. Defining $\tilde{A}$ and $\pi_{\mathcal{N}}$ as in the proof of Proposition 3.1, we have $\tilde{A}^* \circ J_{\mathcal{W},p} \circ \tilde{A} = J_{\mathcal{F}/\mathcal{N},p}$ and $J_{\mathcal{F},p} = \pi_{\mathcal{N}}^* \circ J_{\mathcal{F}/\mathcal{N},p} \circ \pi_{\mathcal{N}}$. Hence, $A^* \circ J_{\mathcal{W},p} \circ A = J_{\mathcal{F},p}$. Therefore, it follows from (3.13) and (3.1) that, for every $(x, u) \in \mathcal{X} \times \mathcal{F}$,

$$\begin{aligned} (\forall\, y^* \in \mathsf{Y}^*) \quad \langle Au, J_{\mathcal{W},p}(K_p(x, \cdot)y^*)\rangle &= \big\langle Au, J_{\mathcal{W},p}(AJ_{\mathcal{F},p}^{-1}(\Phi(x)y^*))\big\rangle & (3.14)\\ &= \langle u, \Phi(x)y^*\rangle \\ &= \langle (Au)(x), y^*\rangle. & (3.15) \end{aligned}$$

Finally if a kernel satisfies (3.10), it satisfies (3.14) and hence (3.13), and thus coincides with K_p. ☐

Remark 3.11

(i) Equation (3.10) is a representation formula, meaning that the values of the functions in $\mathcal{W}$ can be computed in terms of the kernel K_p, which is said to be associated with the feature map Φ.

(ii) Definition 3.3(ii) is more general than [80, Definition 2.2], since the latter requires that both $\mathcal{F}$ and Y be uniformly convex and uniformly smooth. Thus, Proposition 3.10 extends [80, Theorems 2.3 and 3.1]. To this respect, we note also that what is essential to properly define a kernel is that the duality map is single valued and bijective, and this is equivalent to require strict convexity and smoothness only. Moreover, in Proposition 3.10, the kernel is built from a feature map, a general p-duality map, and without any density assumption (see Remark 3.4(ii)), which results in a more general setting than that of [78, 80]. Finally, we emphasize that, when dealing with kernels in Banach spaces, there is no reason to restrict oneself to the normalized duality map. Rather, allowing general p-duality maps usually makes the computation of the kernel easier, as the following two examples show.

Remark 3.12 In the setting of Proposition 3.10, consider the scalar case $\mathsf{Y} = \mathbb{R}$ [78]. Then, for every $x \in \mathcal{X}$, $\Phi(x)^* \in \mathcal{F}^*$ and the kernel becomes

$$K_p \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R} \colon (x, x') \mapsto \big\langle J_{\mathcal{F},p}^{-1}(\Phi(x)^*), \Phi(x')^* \big\rangle. \tag{3.16}$$

Moreover, for every $x \in \mathcal{X}$, $K_p(x, \cdot) = A[J_{\mathcal{F},p}^{-1}(\Phi(x)^*)]$, and formula (3.10) turns into

$$(\forall\, u \in \mathcal{F})(\forall\, x \in \mathcal{X}) \quad \langle Au, J_{\mathcal{W},p}(K(x, \cdot))\rangle = (Au)(x)\,. \tag{3.17}$$

It follows from the definitions of K_p and $J_{\mathcal{F},p}$ that

$$(\forall\,(x, x') \in \mathcal{X} \times \mathcal{X}) \quad K_p(x, x) = \|\Phi(x)\|^{p^*} \quad \text{and} \quad |K_p(x, x')| \leqslant K_p(x, x)^{1/p} K_p(x', x')^{1/p^*}. \tag{3.18}$$

Example 3.13 (generalized linear model) Let $\mathcal{X}$ be a nonempty set, let Y be a separable real Banach space with norm $|\cdot|$, let $\mathbb{K}$ be a nonempty countable set, let $r \in [1, +\infty[$. Let $(\phi_k)_{k\in\mathbb{K}}$ be a family of functions from $\mathcal{X}$ to Y, which, in this context, is usually called a *dictionary* [30, 61]. Assume that for every $x \in \mathcal{X}$, $(\phi_k(x))_{k\in\mathbb{K}} \in l^{r^*}(\mathbb{K}; \mathsf{Y})$ and denote by $\|(\phi_k(x))_{k\in\mathbb{K}}\|_{r^*}$ its norm in $l^{r^*}(\mathbb{K}; \mathsf{Y})$. Set

$$A \colon l^r(\mathbb{K}) \to \mathsf{Y}^{\mathcal{X}} \colon u = (\mu_k)_{k\in\mathbb{K}} \mapsto \sum_{k\in\mathbb{K}} \mu_k \phi_k \text{ (pointwise)}. \tag{3.19}$$

Let $x \in \mathcal{X}$. By Hölder's inequality we derive that, for every $u \in \mathcal{F}$, $|(Au)(x)| \leqslant \|u\|_r \|(\phi_k(x))_{k\in\mathbb{K}}\|_{r^*}$, which implies that $\mathrm{ev}_x \circ A$ is continuous. Therefore, Proposition 3.1 ensures that

$$\operatorname{ran} A = \Big\{ f \in \mathsf{Y}^{\mathcal{X}} \;\Big|\; (\exists\, u \in l^r(\mathbb{K}))(\forall\, x \in \mathcal{X}) \quad f(x) = \sum_{k\in\mathbb{K}} \mu_k \phi_k(x) \Big\} \tag{3.20}$$

can be endowed with a Banach space structure for which the point-evaluation operators are continuous. Moreover

$$\ker A = \Big\{ u \in l^r(\mathbb{K}) \;\Big|\; (\forall\, x \in \mathcal{X}) \quad \sum_{k\in\mathbb{K}} \mu_k \phi_k(x) = 0 \Big\} \tag{3.21}$$

and, for every $u \in l^r(\mathbb{K})$, $\|Au\| = \inf_{v\in\ker A} \|u - v\|_r$. Hence, for every $f \in \operatorname{ran} A$,

$$\|f\| = \inf \Big\{ \|u\|_r \;\Big|\; u \in l^r(\mathbb{K}) \text{ and } (\forall\, x \in \mathcal{X}) \quad f(x) = \sum_{k\in\mathbb{K}} \mu_k \phi_k(x) \Big\}. \tag{3.22}$$

Let us compute the feature map $\Phi \colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*, l^{r^*}(\mathbb{K}))$. Let $x \in \mathcal{X}$, let $y^* \in \mathsf{Y}^*$, and denote by $\langle \cdot, \cdot \rangle_{r,r^*}$ the canonical pairing between $l^r(\mathbb{K})$ and $l^{r^*}(\mathbb{K})$. Then, for every $u \in l^r(\mathbb{K})$,

$$\langle u, \Phi(x) y^* \rangle_{r,r^*} = \langle \Phi(x)^* u, y^* \rangle = \langle (Au)(x), y^* \rangle = \sum_{k\in\mathbb{K}} \mu_k \langle \phi_k(x), y^* \rangle, \tag{3.23}$$

which gives $\Phi(x) y^* = (\langle \phi_k(x), y^* \rangle)_{k\in\mathbb{K}}$. Since $\mathscr{L}(\mathsf{Y}^*, l^{r^*}(\mathbb{K}))$ and $l^{r^*}(\mathbb{K}; \mathsf{Y})$ are isomorphic Banach spaces, the feature map can be identified with

$$\Phi \colon \mathcal{X} \to l^{r^*}(\mathbb{K}; \mathsf{Y}) \colon x \mapsto (\phi_k(x))_{k\in\mathbb{K}}. \tag{3.24}$$

We remark that $\operatorname{ran} A$ is p-universal if, for every probability measure μ on $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}})$, the span of $(\phi_k)_{k\in\mathbb{K}}$ is dense in $L^p(\mathcal{X}, \mu; \mathsf{Y})$. Now suppose that $r > 1$. Since $l^r(\mathbb{K})$ is reflexive, strictly convex, and smooth, Proposition 3.10 asserts that $\operatorname{ran} A$ is a reproducing kernel Banach space and that the underlying kernel $K_r\colon \mathcal{X}\times\mathcal{X}\to\mathscr{B}(\mathsf{Y}^*,\mathsf{Y})$ can be computed explicitly. Indeed, [25, Proposition 4.9] implies that the r-duality map of $l^r(\mathbb{K})$ is

$$J_r\colon l^r(\mathbb{K})\to l^{r^*}(\mathbb{K})\colon u=(\mu_k)_{k\in\mathbb{K}}\mapsto(|\mu_k|^{r-1}\operatorname{sign}(\mu_k))_{k\in\mathbb{K}} \tag{3.25}$$

Moreover, $J_r^{-1}\colon l^{r^*}(\mathbb{K})\to l^r(\mathbb{K})$ is the r^*-duality map of $l^{r^*}(\mathbb{K})$ (hence it has the same form as (3.25) with r replaced by r^*). Thus, for every $(x,x')\in\mathcal{X}\times\mathcal{X}$ and every $y^*\in\mathsf{Y}$

$$K_r(x,x')y^*=\Phi(x')^*\big(J_r^{-1}(\Phi(x)y^*)\big)=\sum_{k\in\mathbb{K}}|\langle\phi_k(x),y^*\rangle|^{r^*-1}\operatorname{sign}(\langle\phi_k(x),y^*\rangle)\phi_k(x'). \tag{3.26}$$

In the scalar case $\mathsf{Y}=\mathbb{R}$, this becomes

$$K_r(x,x')=\big\langle J_r^{-1}(\Phi(x)),\Phi(x')\big\rangle_{r,r^*}=\sum_{k\in\mathbb{K}}|\phi_k(x)|^{r^*-1}\operatorname{sign}(\phi_k(x))\phi_k(x'). \tag{3.27}$$

Example 3.14 (Sobolev spaces) Let $(d,k,m)\in(\mathbb{N}\smallsetminus\{0\})^3$ and let $p\in\,]1,+\infty[$. Let $\mathcal{X}\subset\mathbb{R}^d$ be a nonempty open bounded set with regular boundary and consider the Sobolev space $W^{m,p}(\mathcal{X};\mathbb{R}^k)$, normed with $\|\cdot\|_{m,p}\colon f\mapsto\big(\sum_{\alpha\in\mathbb{N}^d,|\alpha|\leqslant m}\|D^\alpha f\|_p^p\big)^{1/p}$. Recall that, if $mp>d$, then $W^{m,p}(\mathcal{X};\mathbb{R}^k)$ is continuously embedded in $\mathcal{C}(\overline{\mathcal{X}};\mathbb{R}^k)$ [1]. Therefore

$$(\exists\,\beta\in\mathbb{R}_{++})(\forall x\in\mathcal{X})(\forall f\in W^{m,p}(\mathcal{X};\mathbb{R}^k))\quad|f(x)|\leqslant\|f\|_\infty\leqslant\beta\|f\|_{m,p}. \tag{3.28}$$

Moreover $W^{m,p}(\mathcal{X};\mathbb{R}^k)$ is isometrically isomorphic to a closed vector subspace of $[L^p(\mathcal{X};\mathbb{R}^k)]^n$, for a suitable $n\in\mathbb{N}$, normed with $\|\cdot\|_p\colon(f_1,\ldots,f_n)\mapsto\big(\sum_{i=1}^n\|f_i\|_p^p\big)^{1/p}$. Therefore, $W^{m,p}(\mathcal{X};\mathbb{R}^k)$ is uniformly convex and smooth (with the same moduli of convexity and smoothness as L^p). This shows that $W^{m,p}(\mathcal{X};\mathbb{R}^k)$ is a reproducing kernel Banach space and also that the associated feature map Φ is bounded. Likewise, $W_0^{m,p}(\mathcal{X};\mathbb{R}^k)$ is a reproducing kernel Banach space endowed with the norm $\|\nabla\cdot\|_p$, where this time $\nabla\colon W_0^{m,p}(\mathcal{X};\mathbb{R}^k)\to L^p(\mathcal{X};\mathbb{R}^{k\times d})$ is an isometry. For simplicity, we address the computation of the kernel for the space $W_0^{1,p}(\mathcal{X};\mathbb{R})$. In this case, the p-duality map is

$$\frac{1}{p}\partial\|\nabla\cdot\|_p^p=-\Delta_p\colon W_0^{1,p}(\mathcal{X};\mathbb{R})\to\big(W_0^{1,p}(\mathcal{X};\mathbb{R})\big)^*, \tag{3.29}$$

where Δ_p is the p-Laplacian operator [5, Section 6.6]. Therefore, it follows from (3.16) that

$$(\forall\,(x,x')\in\mathcal{X}^2)\quad K_p(x,x')=u(x'),\quad\text{where}\quad u\neq0\quad\text{and}\quad-\Delta_p u=\mathrm{ev}_x\,. \tag{3.30}$$

In the case when $\mathcal{X}=[0,1]$, the kernel can be computed explicitly as follows

$$(\forall\,(x,x')\in\mathcal{X}^2)\quad K_p(x,x')=\begin{cases}\dfrac{(1-x)x'}{\big(x^{p-1}+(1-x)^{p-1}\big)^{1/(p-1)}} & \text{if } x'\leqslant x\\[2ex] \dfrac{(1-x')x}{\big(x^{p-1}+(1-x)^{p-1}\big)^{1/(p-1)}} & \text{if } x'\geqslant x\end{cases} \tag{3.31}$$

Finally, using a mollifier argument [1, Theorem 2.29], $W_0^{m,p}(\mathcal{X};\mathbb{R})_+$ is dense in $\mathscr{C}_0(\mathcal{X};\mathbb{R})_+$. Hence, by Theorem 3.8, $W_0^{m,p}(\mathcal{X};\mathbb{R})$ is universal relative to the cone of $\mathbb{R}_+$-valued functions.

Remark 3.15 Proposition 3.10 and the results pertaining to the computation of the kernel are of interest in their own right. Note, however, that they will not be directly exploited subsequently since in the main results of Section 4.1 knowledge of a kernel will turn out not to be indispensable.

3.2 Representer and sensitivity theorems in Banach spaces

In the classical setting, a representer theorem states that a minimizer of a Tikhonov regularized empirical risk function defined over a reproducing kernel Hilbert space can be represented as a finite linear combination of the feature map values on the training points [58]. The investigation in Banach spaces was initiated in [50] and continued in [79, 80]. In this section representer theorems are established in the general context of Banach spaces, totally convex regularizers, vector-valued functions, and approximate minimization. These contributions capture and extend existing results. Moreover, we study the sensitivity of such representations with respect to perturbations of the probability distribution on $\mathcal{X} \times \mathcal{Y}$.

Definition 3.16 Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets, let $(\mathcal{X} \times \mathcal{Y}, \mathfrak{A}, P)$ be a complete probability space, and let $P_{\mathcal{X}}$ be the marginal probability measure of P on $\mathcal{X}$. Let Y be a separable reflexive real Banach space with norm $|\cdot|$ and Borel σ-algebra $\mathfrak{B}_{\mathsf{Y}}$. $\Upsilon(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ is the set of functions $\ell\colon \mathcal{X} \times \mathcal{Y} \times \mathsf{Y} \to \mathbb{R}_+$ such that ℓ is measurable with respect to the tensor product σ-algebra $\mathfrak{A} \otimes \mathfrak{B}_{\mathsf{Y}}$ and, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(x, y, \cdot)\colon \mathsf{Y} \to \mathbb{R}$ is continuous and convex. A function in $\Upsilon(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ is a *loss*. The *risk* associated with $\ell \in \Upsilon(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ and P is

$$R\colon \mathcal{M}(\mathcal{X}, \mathsf{Y}) \to [0, +\infty] : f \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \ell\big(x, y, f(x)\big) P(\mathrm{d}(x, y)). \tag{3.32}$$

In addition,

(i) given $p \in [1, +\infty[$, $\Upsilon_p(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$ is the set of functions $\ell \in \Upsilon(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ such that

$$(\exists\, b \in L^1(\mathcal{X}\times\mathcal{Y}, P; \mathbb{R}))(\exists\, c \in \mathbb{R}_+)(\forall (x, y, \mathsf{w}) \in \mathcal{X}\times\mathcal{Y}\times\mathsf{Y}) \quad \ell(x, y, \mathsf{w}) \leqslant b(x, y) + c|\mathsf{w}|^p; \tag{3.33}$$

(ii) $\Upsilon_\infty(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$ is the set of functions $\ell \in \Upsilon(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ such that

$$(\forall \rho \in \mathbb{R}_{++})(\exists\, g_\rho \in L^1(\mathcal{X} \times \mathcal{Y}, P; \mathbb{R})) \\ (\forall\, (x, y) \in \mathcal{X} \times \mathcal{Y})(\forall \mathsf{w} \in B(\rho)) \quad \ell(x, y, \mathsf{w}) \leqslant g_\rho(x, y); \tag{3.34}$$

(iii) $\Upsilon_{\mathsf{Y},\mathrm{loc}}(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ is the set of functions $\ell \in \Upsilon(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ such that

$$(\forall \rho \in \mathbb{R}_{++})(\exists\, \mathrm{Lip}(\ell; \rho) \in \mathbb{R}_{++})(\forall (x, y) \in \mathcal{X} \times \mathcal{Y})(\forall (\mathsf{w}, \mathsf{w}') \in B(\rho)^2) \\ |\ell(x, y, \mathsf{w}) - \ell(x, y, \mathsf{w}')| \leqslant \mathrm{Lip}(\ell; \rho)|\mathsf{w} - \mathsf{w}'|. \tag{3.35}$$

Remark 3.17

(i) The properties defining the classes of losses introduced in Definition 3.16 arise in the calculus of variations [38]. Let $p \in [1, +\infty]$ and suppose that $\ell \in \Upsilon_p(\mathcal{X}\times\mathcal{Y}\times\mathsf{Y}, P)$. Then the risk (3.32) is real-valued on $L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$. Moreover, since for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(x, y, \cdot)$ is convex and continuous, $R\colon L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y}) \to \mathbb{R}_+$ is convex and continuous [38, Corollaries 6.51 and 6.53].

(ii) If $\ell \in \Upsilon_p(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$ then $\ell(x, y, \cdot)$ is bounded on bounded sets. Hence, by Proposition A.1(ii), $\ell(x, y, \cdot)$ is Lipschitz continuous relative to bounded sets.

(iii) If $q \in [p, +\infty]$, then $\Upsilon_p(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P) \subset \Upsilon_q(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$.

(iv) Suppose that $\ell \in \Upsilon_{\mathsf{Y},\mathrm{loc}}(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ and that there exists $f \in L^\infty(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$ such that $R(f) < +\infty$. Then $\ell \in \Upsilon_\infty(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$ and (i) implies that $R\colon L^\infty(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y}) \to \mathbb{R}_+$ is convex and continuous.

(v) The following are consequences of Propositions A.1(ii) and A.2(ii):

(a) Suppose that $\ell \in \Upsilon_1(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$ and let $c \in \mathbb{R}_+$ be as in Definition 3.16(i). Then $\ell \in \Upsilon_{\mathsf{Y},\mathrm{loc}}(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ and $\sup_{\rho \in \mathbb{R}_{++}} \mathrm{Lip}(\ell; \rho) \leqslant c$. Hence ℓ is Lipschitz continuous in the third variable, uniformly with respect to the first two. Moreover, in this case, the inequality in (3.33) is true with $b = \ell(\cdot, \cdot, 0)$.

(b) Let $p \in \left]1, +\infty\right[$, let $\ell \in \Upsilon_p(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$, and suppose that the inequality in (3.33) holds with b bounded and some $c \in \mathbb{R}_+$. Then $\ell \in \Upsilon_{\mathsf{Y},\mathrm{loc}}(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ and $\ell(\cdot, \cdot, 0)$ is bounded. Moreover, for every $\rho \in \mathbb{R}_{++}$, $\mathrm{Lip}(\ell; \rho) \leqslant (p-1)\|b\|_\infty + 3cp \max\{1, \rho^{p-1}\}$.

(c) Let $\ell \in \Upsilon_\infty(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$. Then the functions $(g_\rho)_{\rho \in \mathbb{R}_{++}}$ in (3.34) belong to $L^\infty(P)$ if and only if $\ell \in \Upsilon_{\mathsf{Y},\mathrm{loc}}(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ and $\ell(\cdot, \cdot, 0)$ is bounded. In this case, for every $\rho \in \mathbb{R}_{++}$, $\mathrm{Lip}(\ell; \rho) \leqslant 2\|g_{\rho+1}\|_\infty$.

Example 3.18 (L^p-loss) Consider the setting of Definition 3.16 and let $p \in \left[1, +\infty\right[$. Suppose that $\mathcal{Y} \subset \mathsf{Y}$, that $\int_{\mathcal{X} \times \mathcal{Y}} |y|^p P(\mathrm{d}(x, y)) < +\infty$, and that

$$(\forall (x, y, \mathsf{w}) \in \mathcal{X} \times \mathcal{Y} \times \mathsf{Y}) \quad \ell(x, y, \mathsf{w}) = |y - \mathsf{w}|^p. \tag{3.36}$$

Then $\ell \in \Upsilon_p(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$. Moreover, suppose that $\mathcal{Y}$ is bounded and set $\beta = \sup_{y \in \mathcal{Y}} |y|$. Then $\ell \in \Upsilon_{\mathsf{Y},\mathrm{loc}}(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ and $(\forall \rho \in \mathbb{R}_{++})$ $\mathrm{Lip}(\ell; \rho) \leqslant p(\rho + \beta)^{p-1}$. Indeed, the case $p = 1$ is straightforward. If $p > 1$, it follows from (A.7) that, for every $y \in \mathcal{Y}$ and every $(\mathsf{w}, \mathsf{w}') \in \mathsf{Y}^2$, $\big||\mathsf{w} - y|^p - |\mathsf{w}' - y|^p\big| \leqslant p \max\{|y - \mathsf{w}|^{p-1}, |y - \mathsf{w}'|^{p-1}\}|\mathsf{w} - \mathsf{w}'|$. Therefore, for every $(\mathsf{w}, \mathsf{w}') \in B(\rho)^2$ and every $y \in \mathcal{Y}$, $\big||\mathsf{w} - y|^p - |\mathsf{w}' - y|^p\big| \leqslant p(\rho + \beta)^{p-1}|\mathsf{w} - \mathsf{w}'|$.

Now we propose a general representer theorem which involves the feature map from Definition 3.2.

Theorem 3.19 (Representer) *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets, let $(\mathcal{X} \times \mathcal{Y}, \mathfrak{A}, P)$ be a complete probability space, and let $P_{\mathcal{X}}$ be the marginal probability measure of P on $\mathcal{X}$. Let Y be a separable reflexive real Banach space with norm $|\cdot|$, let $\mathcal{F}$ be a separable reflexive real Banach space, let $A\colon \mathcal{F} \to \mathcal{M}(\mathcal{X}, \mathsf{Y})$ be linear and continuous with respect to pointwise convergence on $\mathsf{Y}^{\mathcal{X}}$, and let Φ be the associated feature map. Let $p \in [1, +\infty]$, let $\ell \in \Upsilon_p(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$, let R be the risk associated with ℓ and P, and suppose that $\Phi \in L^p[\mathcal{X}, P_{\mathcal{X}}; \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)]$. Set $F = R \circ A$, let $G \in \Gamma_0^+(\mathcal{F})$, let $\lambda \in \mathbb{R}_{++}$, let $\epsilon \in \mathbb{R}_+$, and suppose that $u_\lambda \in \mathcal{F}$ satisfies*

$$\inf \|\partial(F + \lambda G)(u_\lambda)\| \leqslant \epsilon. \tag{3.37}$$

Then there exists $h_\lambda \in L^{p^*}(\mathcal{X} \times \mathcal{Y}, P; \mathsf{Y}^*)$ such that

$$(\forall (x, y) \in \mathcal{X} \times \mathcal{Y}) \quad h_\lambda(x, y) \in \partial_{\mathsf{Y}} \ell\big(x, y, (Au_\lambda)(x)\big) \tag{3.38}$$

and

$$(\exists\, e^* \in \mathcal{F}^*) \quad \|e^*\| \leqslant \epsilon \quad \text{and} \quad e^* - \mathsf{E}_P(\Phi h_\lambda) \in \lambda \partial G(u_\lambda), \tag{3.39}$$

where $\Phi h_\lambda\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{F}^*\colon (x, y) \mapsto \Phi(x) h_\lambda(x, y)$ and, for every $(x, y, \mathsf{w}) \in \mathcal{X} \times \mathcal{Y} \times \mathsf{Y}$, $\partial_{\mathsf{Y}} \ell(x, y, \mathsf{w}) = \partial \ell(x, y, \cdot)(\mathsf{w})$. Moreover, the following hold:

(i) *Suppose that $p \neq +\infty$. Let (b, c) be as in Definition 3.16(i). If $p = 1$, then $\|h_\lambda\|_\infty \leqslant c$; if $p > 1$, then $\|h_\lambda\|_1 \leqslant (p-1)\|b\|_1 + 3pc(1 + \|\Phi\|_p^{p-1}\|u_\lambda\|^{p-1})$.*

(ii) *Suppose that $p = +\infty$, that $\ell \in \Upsilon_{\mathsf{Y},loc}(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y})$ and let $\rho \in]\|u_\lambda\|, +\infty[$. Then $h_\lambda \in L^\infty(\mathcal{X} \times \mathcal{Y}, P; \mathsf{Y}^*)$ and $\|h_\lambda\|_\infty \leqslant \operatorname{Lip}(\ell; \rho\|\Phi\|_\infty)$.*

Proof. Set

$$\Psi \colon L^p(\mathcal{X} \times \mathcal{Y}, P; \mathsf{Y}) \to [0, +\infty] : g \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \ell(z, g(z)) P(\mathrm{d}z). \tag{3.40}$$

Since $\ell \in \Upsilon_p(\mathcal{X}\times\mathcal{Y}\times\mathsf{Y}, P)$, Ψ is real-valued and convex. Place $L^p(\mathcal{X}\times\mathcal{Y}, P; \mathsf{Y})$ and $L^{p^*}(\mathcal{X}\times\mathcal{Y}, P; \mathsf{Y}^*)$ in duality by means of the pairing

$$\langle \cdot, \cdot \rangle_{p,p^*} \colon (g, h) \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \langle g(z), h(z) \rangle \, P(\mathrm{d}z). \tag{3.41}$$

From now on, we denote by L^p and L^{p^*} the above cited Lebesgue spaces, endowed with the weak topologies $\sigma(L^p, L^{p^*})$ and $\sigma(L^{p^*}, L^p)$, derived from the duality (3.41). Moreover, since $\ell \geqslant 0$, it follows from [57, Theorem 21(c)-(d)] that $\Psi \colon L^p \to \mathbb{R}$ is lower semicontinuous and

$$(\forall g \in L^p) \quad \partial\Psi(g) = \big\{h \in L^{p^*} \;\big|\; h(z) \in \partial_{\mathsf{Y}}\ell(z, g(z)) \text{ for } P\text{-a.a. } z \in \mathcal{X} \times \mathcal{Y}\big\}. \tag{3.42}$$

Next, since $\Phi \in L^p[\mathcal{X}, P_{\mathcal{X}}; \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)]$, it follows from Proposition 3.5(ii), that $A \colon \mathcal{F} \to L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$ is continuous. Therefore the map $\widehat{A} \colon \mathcal{F} \to L^p$ defined by

$$(\forall\, u \in \mathcal{F}) \quad \widehat{A}u \colon \mathcal{X} \times \mathcal{Y} \to \mathsf{Y} \colon (x, y) \mapsto (Au)(x) \tag{3.43}$$

is linear and continuous. Moreover,

$$(\forall\, u \in \mathcal{F})(\forall\, h \in L^{p^*}) \quad \langle \widehat{A}u, h \rangle_{p,p^*} = \int_{\mathcal{X}\times\mathcal{Y}} \langle u, \Phi(x)h(x, y) \rangle \, P(\mathrm{d}(x, y)) = \langle u, \mathsf{E}_P(\Phi h) \rangle. \tag{3.44}$$

Note that, in (3.44), $\mathsf{E}_P(\Phi h)$ is well defined, since Φh is measurable [34, Proposition 1.7], and, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\|\Phi(x)h(x, y)\| \leqslant \|\Phi(x)\||h(x, y)|$. Hence, by Hölder's inequality $\int_{\mathcal{X}\times\mathcal{Y}} \|\Phi(x)h(x, y)\| P(\mathrm{d}(x, y)) < +\infty$, and (3.44) implies that $\widehat{A}^* : L^{p^*} \to \mathcal{F}^* \colon h \mapsto \mathsf{E}_P(\Phi h)$. Now, since $F = \Psi \circ \widehat{A}$, applying [77, Theorem 2.8.3(vi)] to $\Psi \colon L^p \to \mathbb{R}$ and $\widehat{A} \colon \mathcal{F} \to L^p$ and, taking into account (3.42), we get

$$\begin{aligned} \partial F(u_\lambda) &= \widehat{A}^*(\partial\Psi(\widehat{A}u_\lambda)) \\ &= \big\{\mathsf{E}_P(\Phi h) \;\big|\; h \in L^{p^*}, h(x, y) \in \partial_{\mathsf{Y}}\ell(x, y, (Au_\lambda)(x)) \text{ for } P\text{-a.a. } (x, y) \in \mathcal{X} \times \mathcal{Y}\big\}. \end{aligned} \tag{3.45}$$

Using (3.37) and [77, Theorem 2.8.3(vii)], there exists $e^* \in B(\varepsilon)$ such that $e^* \in \partial(F + \lambda G)(u_\lambda) = \partial F(u_\lambda) + \lambda\partial G(u_\lambda)$. Hence, in view of (3.45), there exists $h_\lambda \in L^{p^*}$ satisfying $h_\lambda(x, y) \in \partial_{\mathsf{Y}}\ell(x, y, (Au_\lambda)(x))$ for P-a.a. $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $e^* - \mathsf{E}_P[\Phi h_\lambda] \in \lambda\partial G(u_\lambda)$. Since P is complete, and for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\operatorname{dom} \partial_{\mathsf{Y}}\ell(x, y, \cdot) \neq \varnothing$, we can modify h_λ so that $h_\lambda(x, y) \in \partial_{\mathsf{Y}}\ell(x, y, (Au_\lambda)(x))$ holds for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

(i): Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Since $h_\lambda(x, y) \in \partial_{\mathsf{Y}}\ell(x, y, (Au_\lambda)(x))$,

$$|(Au_\lambda)(x)| = |\Phi(x)^* u_\lambda| \leqslant \|\Phi(x)\|\|u_\lambda\|. \tag{3.46}$$

By Definition 3.16(i), there exists $b \in L^1(\mathcal{X}\times\mathcal{Y},P;\mathbb{R})_+$ and $c \in \mathbb{R}_{++}$ such that, for every $\mathsf{w} \in \mathsf{Y}$, $\ell(x,y,\mathsf{w}) \leqslant b(x,y) + c|\mathsf{w}|^p$. Therefore, it follows from Proposition A.2 and (3.46) that, if $p=1$, we have $|h_\lambda(x,y)| \leqslant c$ and, if $p>1$, we have $|h_\lambda(x,y)| \leqslant (p-1)b(x,y) + 3pc(\|\Phi(x)\|^{p-1}\|u_\lambda\|^{p-1}+1)$. Hence, using Jensen's inequality, $\|h_\lambda\|_1 \leqslant (p-1)\|b\|_1 + 3cp(1+\|\Phi\|_p^{p-1}\|u_\lambda\|^{p-1})$.

(ii): Let $(x,y) \in \mathcal{X}\times\mathcal{Y}$ be such that $\|\Phi(x)\| \leqslant \|\Phi\|_\infty$, and set $\tau = \rho\|\Phi\|_\infty$. We assume $\tau > 0$. Then (3.46) yields $|(Au_\lambda)(x)| < \tau$. Thus, since $B(\tau)$ is a neighborhood of $(Au_\lambda)(x)$ in Y, $\ell(x,y,\cdot)$ is Lipschitz continuous relative to $B(\tau)$, with Lipschitz constant $\mathrm{Lip}(\ell;\tau)$ and $h_\lambda(x,y) \in \partial_{\mathsf{Y}}\ell(x,y,(Au_\lambda)(x))$, Proposition A.1(i) gives $|h_\lambda(x,y)| \leqslant \mathrm{Lip}(\ell;\tau)$. ☐

Remark 3.20

(i) Condition (3.37) is a relaxation of the characterization of u_λ as an exact minimizer of $F+\lambda G$ via Fermat's rule, namely $0 \in \partial(F+\lambda G)(u_\lambda)$.

(ii) Using different methods, [80, Theorem 5.7] gives a representer theorem which holds only for reproducing kernel Banach spaces of vector-valued functions, discrete probabilities, and $\epsilon = 0$ (see the following Remark 3.23). By contrast, Theorem 3.19 is formulated for general probability measures and in terms of the feature map. This underlines the fact that the kernel plays no role in the representation and does not even need to exist.

(iii) Theorem 3.19 is sufficiently general to deal with an offset space [31]. To see this, let $\mathcal{F}_1$ and $\mathcal{F}_2$ be separable reflexive real Banach spaces, let $A_1\colon \mathcal{F}_1 \to \mathcal{M}(\mathcal{X},\mathsf{Y})$ and $A_2\colon \mathcal{F}_2 \to \mathcal{M}(\mathcal{X},\mathsf{Y})$ be linear operators which are continuous with respect to pointwise convergence on $\mathsf{Y}^{\mathcal{X}}$, let $\Phi_1\colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*,\mathcal{F}_1^*)$ and $\Phi_2\colon \mathcal{X} \to \mathscr{L}(\mathsf{Y}^*,\mathcal{F}_2^*)$ be the feature maps associated with A_1 and A_2 respectively, and let $G_1 \in \Gamma_0^+(\mathcal{F}_1)$. Suppose that, in Theorem 3.19, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, $\epsilon = 0$, and

$$(\forall u = (u_1,u_2) \in \mathcal{F}_1\times\mathcal{F}_2) \quad Au = A_1u_1 + A_2u_2 \quad \text{and} \quad G(u) = G_1(u_1). \tag{3.47}$$

Then, setting $u_\lambda = (u_{1,\lambda}, u_{2,\lambda})$, (3.38) and (3.39) yield

$$(\forall\,(x,y)\in\mathcal{X}\times\mathcal{Y}) \quad h_\lambda(x,y) \in \partial_{\mathsf{Y}}\ell\big(x,y,(A_1u_{1,\lambda})(x) + (A_2u_{2,\lambda})(x)\big) \tag{3.48}$$

and

$$-\mathsf{E}_P(\Phi_1 h_\lambda) \in \lambda\partial G_1(u_{1,\lambda}) \quad \text{and} \quad \mathsf{E}_P(\Phi_2 h_\lambda) = 0. \tag{3.49}$$

This gives a representer theorem with offset space $\mathcal{F}_2$. If we assume further that $\mathcal{F}_1$ and $\mathcal{F}_2$ are reproducing kernel Hilbert spaces of scalar functions, that $G_1 = \|\cdot\|^2$, and that $p < +\infty$, the resulting special case of (3.48) and (3.49) appears in [31, Theorem 2].

Corollary 3.21 *In Theorem 3.19, make the additional assumption that $\mathcal{F}$ is strictly convex and smooth, that there exists a convex even function $\varphi\colon \mathbb{R} \to \mathbb{R}_+$ vanishing only at 0 such that*

$$G = \varphi \circ \|\cdot\|, \tag{3.50}$$

and that $u_\lambda \neq 0$. Let $r \in \,]1,+\infty[$. Then there exist $e^* \in \mathcal{F}^*$, $h_\lambda \in L^{p^*}(\mathcal{X}\times\mathcal{Y},P;\mathsf{Y}^*)$, and $\xi(u_\lambda) \in \partial\varphi(\|u_\lambda\|)$ such that $\|e^*\| \leqslant \epsilon$, (3.38) holds, and

$$J_{\mathcal{F},r}(u_\lambda) = \frac{\|u_\lambda\|^{r-1}}{\lambda\xi(u_\lambda)}(e^* - \mathsf{E}_P[\Phi h_\lambda]). \tag{3.51}$$

Proof. Note $\partial\varphi(\mathbb{R}_{++}) \subset \mathbb{R}_{++}$ since φ is strictly increasing on $\mathbb{R}_{++}$. It follows from Theorem 3.19 that there exist $h_\lambda \in L^{p^*}(\mathcal{X}\times\mathcal{Y}, P; \mathsf{Y}^*)$ and $e^* \in \mathcal{F}^*$ such that (3.38) and (3.39) hold. Next, we prove that

$$(\forall\, u \in \mathcal{F}) \quad \partial G(u) = \big\{u^* \in \mathcal{F}^* \;\big|\; \langle u, u^*\rangle = \|u\|\,\|u^*\| \text{ and } \|u^*\| \in \partial\varphi(\|u\|)\big\}. \tag{3.52}$$

It follows from [9, Example 13.7] that, for every $u^* \in \mathcal{F}^*$, $G^*(u^*) = \varphi^*(\|u^*\|)$. Moreover, the Fenchel-Young identity entails that, for every $(u, u^*) \in \mathcal{F}\times\mathcal{F}^*$, we have

$$\begin{aligned} u^* \in \partial G(u) &\Leftrightarrow \varphi(\|u\|) + \varphi^*(\|u^*\|) = \langle u, u^*\rangle \\ &\Leftrightarrow \langle u, u^*\rangle = \|u\|\|u^*\| \text{ and } \|u^*\| \in \partial\varphi(\|u\|)\,. \end{aligned} \tag{3.53}$$

Set $u^*_\lambda = \big(e^* - \mathsf{E}_P(\Phi h_\lambda)\big)/\lambda$. Since $u_\lambda \notin \{0\} = \mathrm{Argmin}_{\mathcal{F}}\, G = \big\{u \in \mathcal{F} \;\big|\; 0 \in \partial G(u)\big\}$ and $u^*_\lambda \in \partial G(u_\lambda)$, then $u^*_\lambda \neq 0$. Now put $v^*_\lambda = \|u_\lambda\|^{r-1} u^*_\lambda/\|u^*_\lambda\|$, then (3.52) yields $\langle u_\lambda, v^*_\lambda\rangle = \|u_\lambda\|^r$ and $\|u^*_\lambda\| \in \partial\varphi(\|u_\lambda\|)$. Moreover, $\|v^*_\lambda\| = \|u_\lambda\|^{r-1}$. Hence, (2.7) yields $v^*_\lambda = J_{\mathcal{F},r}(u_\lambda)$ and (3.51) follows. ☐

Remark 3.22 In Corollary 3.21 let $\varphi = |\cdot|^r$. Then (3.51) specializes to

$$J_{\mathcal{F},r}(u_\lambda) = \frac{1}{r\lambda}\big(e^* - \mathsf{E}_P(\Phi h_\lambda)\big). \tag{3.54}$$

If $\mathcal{F}$ is a Hilbert space, $r = 2$, and $\epsilon = 0$, we obtain the representation $u_\lambda = -(2\lambda)^{-1}\mathsf{E}_P(\Phi h_\lambda)$, which was first obtained in [31, Corollary 3].

Remark 3.23 Let $\epsilon = 0$ and let $P = n^{-1}\sum_{i=1}^n \delta_{(x_i,y_i)}$ be the empirical probability measure associated with the sample $(x_i, y_i)_{1\leqslant i\leqslant n} \in (\mathcal{X}\times\mathcal{Y})^n$. In this context, we obtain a representation for the solution u_λ to the regularized empirical risk minimization problem

$$\underset{u\in\mathcal{F}}{\text{minimize}}\;\; \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, Au(x_i)) + \lambda G(u). \tag{3.55}$$

Indeed (3.39) implies that there exists $(\mathsf{w}^*_i)_{1\leqslant i\leqslant n} \in (\mathsf{Y}^*)^n$ such that

$$\frac{1}{\lambda}\sum_{i=1}^n \Phi(x_i)\mathsf{w}^*_i \in \partial G(u_\lambda) \tag{3.56}$$

We observe that the coefficients $(\mathsf{w}^*_i)_{1\leqslant i\leqslant n}$, solve the dual problem

$$\min_{(\mathsf{w}^*_i)_{1\leqslant i\leqslant n}\in(\mathsf{Y}^*)^n} \lambda G^*\Big(\frac{1}{\lambda}\sum_{i=1}^n \Phi(x_i)\mathsf{w}^*_i\Big) + \frac{1}{n}\sum_{i=1}^n \ell^*(x_i, y_i, -n\mathsf{w}^*_i), \tag{3.57}$$

of (3.55), where $\ell^*(x_i, y_i, \cdot)$ is the conjugate of $\ell(x_i, y_i, \cdot)$. Thus, if G^* is differentiable and Y is finite dimensional, (3.55) can be solved via the finite dimensional convex problem (3.57), by inverting (3.56), which yields

$$u_\lambda = \nabla G^*\Bigg(\frac{1}{\lambda}\sum_{i=1}^n \Phi(x_i)\mathsf{w}^*_i\Bigg). \tag{3.58}$$

If G is as in Corollary 3.21, then (3.58) gives $u_\lambda = J^{-1}_{\mathcal{F},r}\big(\sum_{i=1}^n \Phi(x_i)\mathsf{w}^*_i\big)$. Thus, u_λ, can be expressed in terms of the feature vectors $(\Phi(x_i))_{1\leqslant i\leqslant n}$, for some vector coefficients $(\mathsf{w}^*_i)_{1\leqslant i\leqslant n} \in (\mathsf{Y}^*)^n$. This covers the classical setting of representer theorems in scalar-valued Banach spaces of functions [79, Theorem 3] and improves the vector-valued case of [80, Theorem 5.7]. The dual variational framework (3.57) requires less restrictions and offers more flexibility in terms of solution methods than the fixed point approach proposed in [36], [78, Theorem 23], and [80, Section 5.3].

Example 3.24 We recover a case-study of [50]. Let $\phi\colon \mathbb{R}_+ \to \mathbb{R}_+$ be strictly increasing, continuous, and such that $\phi(0) = 0$ and $\lim_{t\to+\infty} \phi(t) = +\infty$. Define $\varphi\colon \mathbb{R} \to \mathbb{R}_+ \colon t \mapsto \int_0^{|t|} \phi(s)\mathrm{d}s$, which is strictly convex, even, and vanishes only at 0. Assume that $\overline{\lim}_{t\to 0}\, \varphi(2t)/\varphi(t) < +\infty$, let $(\Omega, \mathfrak{S}, \mu)$ be a measure space, and let $\mathcal{F} = L_\varphi(\Omega, \mu; \mathbb{R})$ be the associated Orlicz space endowed with the Luxemburg norm induced by φ. We recall that $\mathcal{F}^* = L_{\varphi^*}(\Omega, \mu; \mathbb{R})$, the Orlicz space endowed with the Orlicz norm associated to φ^* [56]. Moreover, in this case the normalized duality map $J_{\mathcal{F}^*} = J_{\mathcal{F}}^{-1}\colon \mathcal{F}^* \to \mathcal{F}$ can be computed. Indeed, by [56, Theorem 7.2.5], we obtain that, for every $g \in \mathcal{F}^*$, there exists $\kappa_g \in \mathbb{R}_{++}$ such that $J_{\mathcal{F}^*}(g) = \|g\|\phi^{-1}(\kappa_g|g|)\operatorname{sign}(g)$. Given $(g_i)_{1\leqslant i\leqslant n} \in (\mathcal{F}^*)^n$, $(y_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^n$, and $\lambda \in \mathbb{R}_{++}$, the problem considered in [50] is to solve

$$\underset{u\in\mathcal{F}}{\text{minimize}}\ \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \langle u, g_i\rangle) + \lambda\varphi(\|u\|). \tag{3.59}$$

This corresponds to the framework considered in Corollary 3.21 and Remark 3.23, with $\mathcal{X} = \mathcal{F}^*$, $\mathcal{Y} = \mathsf{Y} = \mathbb{R}$, $P = n^{-1}\sum_{i=1}^n \delta_{(g_i,y_i)}$, and $(\forall g \in \mathcal{X})(\forall u \in \mathcal{F})\ (Au)(g) = \langle u, g\rangle$. Since, in this case, for every $g \in \mathcal{X}$, $\Phi(g) = g$, we derive from (3.58) that there exist $\kappa \in \mathbb{R}_{++}$ and $(\alpha_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^n$ such that

$$u_\lambda = \|u_\lambda\|\phi^{-1}\bigg(\kappa\Big|\sum_{i=1}^{n}\alpha_i g_i\Big|\bigg)\operatorname{sign}\bigg(\sum_{i=1}^{n}\alpha_i g_i\bigg) \text{ and } -n\lambda\phi(\|u_\lambda\|)\alpha_i \in \|u_\lambda\|\partial\ell(y_i,\cdot)(\langle u_\lambda, g_i\rangle). \tag{3.60}$$

We conclude this section with a sensitivity result in terms of a perturbation on the underlying probability measure.

Theorem 3.25 (Sensitivity) *In Theorem 3.19, make the additional assumption that G is totally convex at every point of* dom G *and let ψ be its modulus of total convexity. Take $h_\lambda \in L^{p^*}(\mathcal{X}\times\mathcal{Y}, P; \mathsf{Y}^*)$ such that conditions (3.38)-(3.39) hold. Let $\widetilde{P}$ be a probability measure on $(\mathcal{X}\times\mathcal{Y}, \mathfrak{A})$ such that $\ell \in \Upsilon_\infty(\mathcal{X}\times\mathcal{Y}\times\mathsf{Y}, \widetilde{P})$ and Φ is $\widetilde{P}_{\mathcal{X}}$-essentially bounded. Define*

$$\widetilde{R}\colon \mathcal{M}(\mathcal{X},\mathcal{Y}) \to [0,+\infty] \colon f \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \ell(x,y,f(x))\widetilde{P}(\mathrm{d}(x,y)) \quad \textit{and} \quad \widetilde{F} = \widetilde{R}\circ A. \tag{3.61}$$

Let $\tilde{\epsilon} \in \mathbb{R}_{++}$ and let $\tilde{u}_\lambda \in \mathcal{F}$ be such that $\inf \|\partial(\widetilde{F} + \lambda G)(\tilde{u}_\lambda)\| \leqslant \tilde{\epsilon}$. *Then the following hold:*

(i) $h_\lambda \in L^1(\mathcal{X}\times\mathcal{Y}, \widetilde{P}; \mathsf{Y}^*)$.

(ii) $\psi(u_\lambda,\cdot)^\frown(\|\tilde{u}_\lambda - u_\lambda\|) \leqslant \big(\|\mathsf{E}_{\widetilde{P}}(\Phi h_\lambda) - \mathsf{E}_P(\Phi h_\lambda)\| + \epsilon + \tilde{\epsilon}\big)/\lambda$.

Proof. (i): Let γ be the norm of Φ in $L^\infty[\mathcal{X}, \widetilde{P}_{\mathcal{X}}; \mathscr{L}(\mathsf{Y},\mathsf{Z})]$ and let $\rho \in\,]\gamma\|u_\lambda\|, +\infty[$. Since $\ell \in \Upsilon_\infty(\mathcal{X}\times\mathcal{Y}\times\mathsf{Y}, \widetilde{P})$, there exists $g \in L^1(\mathcal{X}\times\mathcal{Y}, \widetilde{P}; \mathbb{R})$ such that

$$(\forall\,(x,y) \in \mathcal{X}\times\mathcal{Y})(\forall\,\mathsf{w} \in B(\rho+1)) \quad \ell(x,y,\mathsf{w}) \leqslant g(x,y). \tag{3.62}$$

Let $(x,y) \in \mathcal{X}\times\mathcal{Y}$ be such that $\|\Phi(x)\| \leqslant \gamma$. Then $|(Au_\lambda)(x)| \leqslant \|\Phi(x)\|\|u_\lambda\| \leqslant \gamma\|u_\lambda\| < \rho$. Therefore, since $h_\lambda(x,y) \in \partial_{\mathsf{Y}}\ell(x,y,(Au_\lambda)(x))$, it follows from Proposition A.1(i)-(ii) and (3.62) that $|h_\lambda(x,y)| \leqslant 2\sup \ell(x,y,B(\rho+1)) \leqslant 2g(x,y)$. Hence $h_\lambda \in L^1(\mathcal{X}\times\mathcal{Y}, \widetilde{P}; \mathsf{Y}^*)$.

(ii): Let $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Since $h_\lambda(x,y) \in \partial_{\mathsf{Y}}\ell(x,y,(Au_\lambda)(x))$, we have

$$\begin{aligned}\langle \tilde{u}_\lambda - u_\lambda, \Phi(x)h_\lambda(x,y)\rangle &= \langle (A\tilde{u}_\lambda)(x) - (Au_\lambda)(x), h_\lambda(x,y)\rangle \\ &\leqslant \ell(x,y,(A\tilde{u}_\lambda)(x)) - \ell(x,y,(Au_\lambda)(x)).\end{aligned} \tag{3.63}$$

Since Φ is $\widetilde{P}_{\mathcal{X}}$-essentially bounded and $h_\lambda \in L^1(\mathcal{X}\times\mathcal{Y}, \widetilde{P}; \mathsf{Y}^*)$, Φh_λ is $\widetilde{P}$-integrable. Integrating (3.63) with respect to $\widetilde{P}$ yields

$$\big\langle \tilde{u}_\lambda - u_\lambda, \mathsf{E}_{\widetilde{P}}(\Phi h_\lambda)\big\rangle \leqslant \widetilde{R}(A\tilde{u}_\lambda) - \widetilde{R}(Au_\lambda). \tag{3.64}$$

Moreover, (3.39) and (A.9) yield

$$\langle \tilde{u}_\lambda - u_\lambda, e^* - \mathsf{E}_P(\Phi h_\lambda)\rangle + \lambda\psi(u_\lambda, \|\tilde{u}_\lambda - u_\lambda\|) \leqslant \lambda G(\tilde{u}_\lambda) - \lambda G(u_\lambda). \tag{3.65}$$

Summing the last two inequalities we obtain

$$\begin{aligned}\big\langle \tilde{u}_\lambda - u_\lambda, \mathsf{E}_{\widetilde{P}}(\Phi h_\lambda) - \mathsf{E}_P(\Phi h_\lambda) + e^*\big\rangle &+ \lambda\psi(u_\lambda, \|\tilde{u}_\lambda - u_\lambda\|)\\ &\leqslant (\widetilde{F} + \lambda G)(\tilde{u}_\lambda) - (\widetilde{F} + \lambda G)(u_\lambda).\end{aligned} \tag{3.66}$$

Since there exists $\tilde{e}^* \in \mathcal{F}^*$ such that $\|\tilde{e}^*\| \leqslant \tilde{\epsilon}$ and $\langle u_\lambda - \tilde{u}_\lambda, \tilde{e}^*\rangle \leqslant (\widetilde{F} + \lambda G)(u_\lambda) - (\widetilde{F} + \lambda G)(\tilde{u}_\lambda)$, we have $(\widetilde{F} + \lambda G)(\tilde{u}_\lambda) - (\widetilde{F} + \lambda G)(u_\lambda) \leqslant \tilde{\epsilon}\|u_\lambda - \tilde{u}_\lambda\|$. This, together with (3.66), yields

$$\lambda\psi(u_\lambda, \|\tilde{u}_\lambda - u_\lambda\|) \leqslant (\epsilon + \tilde{\epsilon})\|\tilde{u}_\lambda - u_\lambda\| + \|\mathsf{E}_{\widetilde{P}}(\Phi h_\lambda) - \mathsf{E}_P(\Phi h_\lambda)\|\|\tilde{u}_\lambda - u_\lambda\| \tag{3.67}$$

and the statement follows. ☐

4 Learning via regularization

We study statistical learning in Banach spaces and present the main results of the paper.

4.1 Consistency theorems

We first formulate our assumptions. They involve the feature map from Definition 3.2, as well as the loss and the risk introduced in Definition 3.16.

Assumption 4.1

(i) $(\Omega, \mathfrak{S}, \mathsf{P})$ is a complete probability space, $\mathcal{X}$ and $\mathcal{Y}$ are two nonempty sets, $\mathfrak{A}$ is a sigma algebra on $\mathcal{X}\times\mathcal{Y}$ containing the singletons, $(X, Y)\colon (\Omega, \mathfrak{S}, \mathsf{P}) \to (\mathcal{X}\times\mathcal{Y}, \mathfrak{A})$ is a random variable with distribution P on $\mathcal{X}\times\mathcal{Y}$, and P has marginal $P_{\mathcal{X}}$ on $\mathcal{X}$.

(ii) Y is a separable reflexive real Banach space, $\ell \in \Upsilon_{\mathsf{Y},\mathrm{loc}}(\mathcal{X}\times\mathcal{Y}\times\mathsf{Y})$, $R\colon \mathcal{M}(\mathcal{X}, \mathsf{Y}) \to [0, +\infty]$ is the risk associated with ℓ and P, and there exists $f \in L^\infty(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$ such that $R(f) < +\infty$. For every $\rho \in \mathbb{R}_{++}$, $\mathrm{Lip}(\ell; \rho)$ is as in (3.35).

(iii) $\mathcal{C}$ is a nonempty convex subset of $\mathcal{M}(\mathcal{X}, \mathsf{Y})$.

(iv) $\mathcal{F}$ is a separable reflexive real Banach space, $q \in [2, +\infty[$, $\mathcal{F}^*$ is of Rademacher type q^* with Rademacher type constant T_{q^*}.

(v) $A\colon \mathcal{F} \to \mathcal{M}(\mathcal{X}, \mathsf{Y})$ is linear and continuous with respect to pointwise convergence on $\mathsf{Y}^{\mathcal{X}}$, Φ is the feature map associated with A, $\Phi \in L^\infty[\mathcal{X}, P_{\mathcal{X}}; \mathscr{L}(\mathsf{Y}^*, \mathcal{F}^*)]$.

(vi) $G \in \Gamma_0^+(\mathcal{F})$, $G(0) = 0$, the modulus of total convexity of G is ψ, $\psi_0 = \psi(0, \cdot)$, and G is totally convex on bounded sets.

(vii) $(\lambda_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{R}_{++}$ such that $\lambda_n \to 0$.

(viii) $(X_i, Y_i)_{i\in\mathbb{N}}$ is a sequence of independent copies of (X, Y). For every $n \in \mathbb{N} \smallsetminus \{0\}$, $Z_n = (X_i, Y_i)_{1\leqslant i\leqslant n}$ and

$$R_n \colon \mathcal{M}(\mathcal{X}, \mathsf{Y}) \times (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}_+ \colon (f, (x_1, y_1), \ldots, (x_n, y_n)) \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i, f(x_i)). \tag{4.1}$$

The function $\varepsilon \colon \mathbb{R}_{++} \to [0, 1]$ satisfies $\lim_{\lambda\to 0^+} \varepsilon(\lambda) = 0$. For every $n \in \mathbb{N} \smallsetminus \{0\}$ and every $\lambda \in \mathbb{R}_{++}$, the function $u_{n,\lambda} \colon (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$ satisfies

$$(\forall z \in (\mathcal{X} \times \mathcal{Y})^n) \quad u_{n,\lambda}(z) \in \operatorname{Argmin}^{\varepsilon(\lambda)}_{\mathcal{F}}(R_n(A\cdot, z) + \lambda G). \tag{4.2}$$

In the context of learning theory, $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space, which can be considered to be embedded in the ambient space Y. The probability distribution P describes a functional relation from $\mathcal{X}$ into $\mathcal{Y}$ and R quantifies the expected loss of a function $f \colon \mathcal{X} \to \mathsf{Y}$ with respect to the underlying distribution P. The set $\mathcal{C}$ models a priori constraints. Since $\mathcal{M}(\mathcal{X}, \mathsf{Y})$ is poorly structured, measurable functions are handled via the Banach feature space $\mathcal{F}$ and the feature map Φ. Note that the resulting space of functions is only a pre-reproducing kernel Banach space in the sense of [60], since a kernel is not required. Under the provision that the range of A is universal relative to $\mathcal{C}$ (see Definition 3.6) every function $f \in \mathcal{C}$ can be approximately represented by a feature $u \in \mathcal{F}$ via $f \approx Au$. Since the true risk R depends on P, which is unknown, the empirical risk R_n is constructed from the available data, namely a realization of Z_n. In (4.2), $u_{n,\lambda}$ is obtained by approximately minimizing a regularized empirical risk. Regularization is achieved by the addition of the convex function G, which will be asked to fulfill certain compatibility conditions with the constraint set $\mathcal{C}$, e.g., $\overline{\operatorname{dom} G} = A^{-1}(\mathcal{C})$. The objective of our analysis can be stated as follows.

Problem 4.2 (consistency) Consider the setting of Assumption 4.1. The problem is to approach the infimum of the risk R on $\mathcal{C}$ by means of approximate solutions

$$u_{n,\lambda_n}(Z_n) \in \operatorname{Argmin}^{\varepsilon(\lambda_n)}_{\mathcal{F}}(R_n(A\cdot, Z_n) + \lambda_n G) \tag{4.3}$$

to the empirical regularized problems

$$\underset{u\in\mathcal{F}}{\text{minimize}} \;\; R_n(Au, Z_n) + \lambda_n G(u), \tag{4.4}$$

in the sense that $R(Au_{n,\lambda_n}(Z_n)) \to \inf R(\mathcal{C})$ in probability (*weak consistency*) or almost surely (*strong consistency*), under suitable conditions on $(\lambda_n)_{n\in\mathbb{N}}$.

Definition 4.3 Let $p \in [1, +\infty]$. Then $\mathcal{C}$ in Assumption 4.1 is *p-admissible* if $\mathcal{C} \subset L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$, or if $\mathcal{C} \cap L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y}) \neq \varnothing$ and there exists a family $(\mathsf{C}(x))_{x\in\mathcal{X}}$ of closed convex subsets of Y such that $\mathcal{C} = \big\{f \in \mathcal{M}(\mathcal{X}, \mathsf{Y}) \;\big|\; (\forall x \in \mathcal{X})\; f(x) \in \mathsf{C}(x)\big\}$.

We are now ready to state the two main results of the paper (see Section 4.2 for proofs).

Theorem 4.4 *Suppose that Assumption 4.1 holds, set $\varsigma = \|\Phi\|_\infty$, and write $\varepsilon = \varepsilon_1\varepsilon_2$, where ε_1 and ε_2 are functions from $\mathbb{R}_{++}$ to $[0, 1]$. Let $p \in [1, +\infty]$ and suppose that $\ell \in \Upsilon_p(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}, P)$, that $\mathcal{C}$ is p-admissible, that $\operatorname{ran} A$ is p-universal relative to $\mathcal{C}$, and that $A(\operatorname{dom} G) \subset \mathcal{C} \cap \operatorname{ran} A \subset \overline{A(\operatorname{dom} G)}$, where the closure is in $L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$. Then the following hold:*

(i) *Assume that* $\ell(\cdot,\cdot,0)$ *is bounded and let* $(\forall n \in \mathbb{N})$ $\rho_n \in \big[\psi_0^\natural\big((\|\ell(\cdot,\cdot,0)\|_\infty + 1)/\lambda_n\big), +\infty\big[$. *Suppose that*

$$\operatorname{Lip}(\ell; \varsigma\rho_n)\varepsilon_1(\lambda_n) \to 0 \quad \textit{and} \quad \varepsilon_2(\lambda_n) = O\bigg(\frac{\operatorname{Lip}(\ell;\varsigma\rho_n)}{n^{1/q}}\bigg), \tag{4.5}$$

and that

$$(\forall \tau \in \mathbb{R}_{++}) \quad \operatorname{Lip}(\ell;\varsigma\rho_n)(\widehat{\psi}_{\rho_n})^\natural\bigg(\frac{\tau \operatorname{Lip}(\ell;\varsigma\rho_n)}{\lambda_n n^{1/q}}\bigg) \to 0. \tag{4.6}$$

Then $R(Au_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf R(\mathcal{C})$. *Moreover, if*

$$(\forall \tau \in \mathbb{R}_{++}) \quad \operatorname{Lip}(\ell;\varsigma\rho_n)(\widehat{\psi}_{\rho_n})^\natural\bigg(\frac{\tau \operatorname{Lip}(\ell;\varsigma\rho_n)\log n}{\lambda_n n^{1/q}}\bigg) \to 0, \tag{4.7}$$

then $R(Au_{n,\lambda_n}(Z_n)) \to \inf R(\mathcal{C})$ P^**–a.s.*

(ii) *Assume that* $p \in \,]1,+\infty[$ *and that the function* b *associated with* ℓ *in Definition 3.16(i) is bounded, and let* $(\forall n \in \mathbb{N})$ $\rho_n \in \big[\psi_0^\natural\big((\|\ell(\cdot,\cdot,0)\|_\infty + 1)/\lambda_n\big), +\infty\big[$. *Suppose that*

$$\rho_n^{p-1}\varepsilon_1(\lambda_n) \to 0,\ \varepsilon_2(\lambda_n) = O\Big(\frac{\rho_n^{p-1}}{n^{1/q}}\Big),\ \textit{and}\ (\forall \tau \in \mathbb{R}_{++})\ \rho_n^{p-1}(\widehat{\psi}_{\rho_n})^\natural\bigg(\frac{\tau\rho_n^{p-1}}{\lambda_n n^{1/q}}\bigg) \to 0. \tag{4.8}$$

Then $R(Au_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf R(\mathcal{C})$. *Moreover, if*

$$(\forall \tau \in \mathbb{R}_{++}) \quad \rho_n^{p-1}(\widehat{\psi}_{\rho_n})^\natural\bigg(\frac{\tau\rho_n^{p-1}\log n}{\lambda_n n^{1/q}}\bigg) \to 0, \tag{4.9}$$

then $R(Au_{n,\lambda_n}(Z_n)) \to \inf R(\mathcal{C})$ P^**–a.s.*

(iii) *Assume that* $p = 1$ *and let* $(\forall n \in \mathbb{N})$ $\rho_n \in \big[\psi_0^\natural((R(0)+1)/\lambda_n), +\infty\big[$. *Suppose that*

$$\varepsilon_1(\lambda_n) \to 0,\ \ \varepsilon_2(\lambda_n) = O\Big(\frac{1}{n^{1/q}}\Big), \quad \textit{and} \quad (\forall \tau \in \mathbb{R}_{++})\ \ (\widehat{\psi}_{\rho_n})^\natural\bigg(\frac{\tau}{\lambda_n n^{1/q}}\bigg) \to 0. \tag{4.10}$$

Then $R(Au_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf R(\mathcal{C})$. *Moreover, if*

$$(\forall \tau \in \mathbb{R}_{++}) \quad (\widehat{\psi}_{\rho_n})^\natural\bigg(\frac{\tau\log n}{\lambda_n n^{1/q}}\bigg) \to 0, \tag{4.11}$$

then $R(Au_{n,\lambda_n}(Z_n)) \to \inf R(\mathcal{C})$ P^**–a.s.*

(iv) *Suppose that* $S = \operatorname{Argmin}_{\operatorname{dom} G}(R \circ A) \neq \varnothing$. *Then there exists a unique* $u^\dagger \in S$ *which minimizes* G *on* S; *moreover,* $Au^\dagger \in \mathcal{C}$ *and* $R(Au^\dagger) = \inf R(\mathcal{C})$. *Furthermore, suppose that the following conditions are satisfied:*

$$\varepsilon_1(\lambda_n) \to 0, \quad \frac{\varepsilon_2(\lambda_n)}{\lambda_n} \to 0, \quad \textit{and} \quad \frac{1}{\lambda_n n^{1/q}} \to 0\,. \tag{4.12}$$

Then $\|u_{n,\lambda_n}(Z_n) - u^\dagger\| \xrightarrow{\mathsf{P}^*} 0$ *and* $R(Au_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} R(Au^\dagger)$. *Finally, suppose in addition that*

$$(\log n)/(n^{1/q}\lambda_n) \to 0\,. \tag{4.13}$$

Then $\|u_{n,\lambda_n}(Z_n) - u^\dagger\| \to 0$ P^**–a.s. and* $R(Au_{n,\lambda_n}(Z_n)) \to R(Au^\dagger)$ P^**–a.s.*

Remark 4.5

(i) In the setting of Example 3.18, $\ell(\cdot,\cdot,0)$ is bounded if $\mathcal{Y}$ is a bounded subset of Y.

(ii) $A(\operatorname{dom} G) \subset \mathcal{C} \cap \operatorname{ran} A \subset \overline{A(\operatorname{dom} G)}$ is a compatibility condition between G and $\mathcal{C}$. It is satisfied in particular when $\overline{\operatorname{dom} G} = A^{-1}(\mathcal{C})$, since $A(A^{-1}(\mathcal{C})) = \mathcal{C} \cap \operatorname{ran} A$. On the other hand, $\operatorname{ran} A$ is trivially ∞-universal relative to $\mathcal{C}$ when $\mathcal{C} \subset \operatorname{ran} A$, or $\operatorname{ran} A \subset \mathcal{C}$ and $\operatorname{ran} A$ is ∞-universal.

(iii) For every $\rho \in \mathbb{R}_{++}$, $\operatorname{dom} (\widehat{\psi}_\rho)^\natural$ is an interval containing 0 with nonempty interior. Indeed, it follows from Assumption 4.1(vi) that $\operatorname{Argmin}_{\mathcal{F}} G \neq \varnothing$, hence $0 \in \operatorname{dom} \partial G$. Therefore, Proposition A.6(viii) ensures that, for every $\rho \in \mathbb{R}_+$, $\psi_\rho \in \mathcal{A}_1$. Thus, Proposition A.5(vii) yields $(\widehat{\psi}_\rho)^\natural \in \mathcal{A}_0$ and the statement follows from Proposition A.5(ii).

(iv) Let $(s_n)_{n\in\mathbb{N}}$ and $(\rho_n)_{n\in\mathbb{N}}$ be sequences in $\mathbb{R}_{++}$ and suppose that $\rho = \inf_{n\in\mathbb{N}} \rho_n > 0$. Then $(\widehat{\psi}_{\rho_n})^\natural(s_n) \to 0 \Rightarrow s_n \to 0$. Indeed, for every $n \in \mathbb{N}$, $\rho \leqslant \rho_n \Rightarrow \psi_{\rho_n} \leqslant \psi_\rho \Rightarrow \widehat{\psi}_{\rho_n} \leqslant \widehat{\psi}_\rho \Rightarrow (\widehat{\psi}_\rho)^\natural \leqslant (\widehat{\psi}_{\rho_n})^\natural$. Therefore $(\widehat{\psi}_{\rho_n})^\natural(s_n) \to 0 \Rightarrow (\widehat{\psi}_\rho)^\natural(s_n) \to 0 \Rightarrow s_n \to 0$ by Proposition A.5(iv).

Next we consider an important special case, in which the consistency conditions can be made explicit.

Corollary 4.6 *Suppose that Assumption 4.1 holds, set $\varsigma = \|\Phi\|_\infty$, and write $\varepsilon = \varepsilon_1\varepsilon_2$, where ε_1 and ε_2 are functions from $\mathbb{R}_{++}$ to $[0,1]$. Let $p \in [1,+\infty]$ and suppose that $\ell \in \Upsilon_p(\mathcal{X}\times\mathcal{Y}\times\mathsf{Y}, P)$, that $\mathcal{C}$ is p-admissible, that $\operatorname{ran} A$ is p-universal relative to $\mathcal{C}$, that $A(\operatorname{dom} G) \subset \mathcal{C} \cap \operatorname{ran} A \subset \overline{A(\operatorname{dom} G)}$, where the closure is in $L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$. In addition, assume that*

$$\begin{cases} \mathcal{F} \text{ is uniformly convex with modulus of convexity of power type } q \\ G = \eta\|\cdot\|^r + H, \quad \text{where } \eta \in \mathbb{R}_{++},\ r \in \left]1,+\infty\right[, \text{ and } H \in \Gamma_0^+(\mathcal{F}). \end{cases} \tag{4.14}$$

Let β be the constant defined in Proposition A.9, and set $m = \max\{r, q\}$. Then the following holds:

(i) *Assume that $\ell(\cdot,\cdot,0)$ is bounded and set $(\forall n \in \mathbb{N})\ \rho_n = \big((\|\ell(\cdot,\cdot,0)\|_\infty + 1)/(\eta\beta\lambda_n)\big)^{1/r}$. Suppose that*

$$\operatorname{Lip}(\ell;\varsigma\rho_n)\varepsilon_1(\lambda_n) \to 0, \qquad \varepsilon_2(\lambda_n) = O\left(\frac{\operatorname{Lip}(\ell;\varsigma\rho_n)}{n^{1/q}}\right), \quad \text{and} \quad \frac{\operatorname{Lip}(\ell;\varsigma\rho_n)^m}{\lambda_n^{m/r} n^{1/q}} \to 0. \tag{4.15}$$

Then $R(Au_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf R(\mathcal{C})$. Moreover, if

$$\frac{\operatorname{Lip}(\ell;\varsigma\rho_n)^m \log n}{\lambda_n^{m/r} n^{1/q}} \to 0, \tag{4.16}$$

then $R(Au_{n,\lambda_n}(Z_n)) \to \inf R(\mathcal{C})$ P^*-a.s.

(ii) *Assume that $p \in \left]1,+\infty\right[$, that the function b associated with ℓ in Definition 3.16(i) is bounded, and that*

$$\frac{\varepsilon_1(\lambda_n)}{\lambda_n^{(p-1)/r}} \to 0, \quad \varepsilon_2(\lambda_n) = O\left(\frac{1}{n^{1/q}\lambda_n^{(p-1)/r}}\right), \quad \text{and} \quad \frac{1}{\lambda_n^{pm/r} n^{1/q}} \to 0. \tag{4.17}$$

Then $R(Au_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf R(\mathcal{C})$. Moreover, if $(\log n)/(\lambda_n^{pm/r} n^{1/q}) \to 0$, then $R(Au_{n,\lambda_n}(Z_n)) \to \inf R(\mathcal{C})$ P^*-a.s.

(iii) *Assume that* $p = 1$ *and that*

$$\varepsilon_1(\lambda_n) \to 0, \quad \varepsilon_2(\lambda_n) = O\Big(\frac{1}{n^{1/q}}\Big), \quad \textit{and} \quad \frac{1}{\lambda_n^{m/r} n^{1/q}} \to 0. \tag{4.18}$$

Then $R(Au_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf R(\mathcal{C})$. *Moreover, if* $(\log n)/(\lambda_n^{m/r} n^{1/q}) \to 0$, *then* $R(Au_{n,\lambda_n}(Z_n)) \to \inf R(\mathcal{C})$ P^**–a.s.*

Remark 4.7 Corollary 4.6 shows that consistency is achieved when the sequence of regularization parameters $(\lambda_n)_{n\in\mathbb{N}}$ converges to zero not too fast. The upper bound depends on the power type of the modulus of convexity of the feature space, the exponent of the norm in the regularizer, and the Lipschitz behavior of the loss. Note that a faster decay of $(\lambda_n)_{n\in\mathbb{N}}$ is allowed when $q = 2$.

Remark 4.8 The class of regularizers considered in Corollary 4.6 includes the elastic-net penalty both in the setting of generalized linear models [30] and multiple kernel learning [63]. The proofs of Theorem 4.4 and Corollary 4.6 are based on a stability analysis, which combines the sensitivity Theorem 3.25 with a Banach space-valued Hoeffding's inequality (Theorem A.17). The strength of such a method is that it can be applied in very general situations, since it does not require any structure on the input space and no hypotheses on the probability measure. We highlight that the analysis can be applied even to unbounded output spaces if Hoeffding's inequality is replaced by Markov's inequality.

Remark 4.9 In the setting of general regularizers and/or Banach feature spaces, the literature on consistency of regularized empirical risk minimizers is scarce.

(i) In [62, Theorem 7.20] only continuous, real-valued regularizers are considered and consistency is established under the provision that local Rademacher complexities can be suitably bounded and an appropriate variance bound holds [8]. However, it is not clear whether this result is useful for other regularizers apart from the squared norm.

(ii) A well-studied method to prove consistency of regularized empirical risk minimization is based on covering numbers [28, 71, 72]. However, it should be stressed that the application of such method in the vector-valued setting would require the following additional assumptions: (a) the input space is a locally compact topological space and the feature map is continuous with respect to the uniform operator topology and takes compact operators as values (this implies the finiteness of the related covering numbers); (b) the covering numbers decay polynomially (this usually requires smooth kernels); and (c) an appropriate variance bound for the loss is available.

(iii) In [59], the consistency of an ℓ^1-regularized empirical risk minimization scheme is studied in a particular type of Banach spaces of functions, in which a linear representer theorem is shown to hold. Note that, in general reproducing kernel Banach spaces, the representation is not linear; see Corollary 3.21 and [79, 80]. In [61], consistency and learning rates are provided for classification problems and $G = \|\cdot\|$, under appropriate growth assumptions on the average empirical entropy numbers.

(iv) In [48] a class of regularizers inducing structured sparsity is considered and associated statistical bounds are provided.

We complete this section by providing an illustration of the above consistency theorems to learning with dictionaries in the context of Example 3.13. The setting will be a specialization of Assumption 4.1 to specific types of feature maps and regularizers. Our analysis extends in several directions that of [30].

Example 4.10 (Generalized linear model) Suppose that Assumption 4.1(i)-(iii) hold. Let $\mathbb{K}$ be a nonempty at most countable set, let $r \in]1, +\infty[$, and let $\mathcal{F} = l^r(\mathbb{K})$. Let $\varsigma \in \mathbb{R}_{++}$, let $(\phi_k)_{k\in\mathbb{K}}$ be a dictionary of functions in $\mathcal{M}(\mathcal{X}, \mathsf{Y})$ such that, for $P_{\mathcal{X}}$-a.a. $x \in \mathcal{X}$, $\sum_{k\in\mathbb{K}} |\phi_k(x)|^{r^*} \leqslant \varsigma^{r^*}$, and set

$$A\colon \mathcal{F} \to \mathsf{Y}^{\mathcal{X}} \colon u = (\mu_k)_{k\in\mathbb{K}} \mapsto \sum_{k\in\mathbb{K}} \mu_k \phi_k \quad \text{(pointwise)}. \tag{4.19}$$

Let $\Phi\colon \mathcal{X} \to l^{r^*}(\mathbb{K}; \mathsf{Y})\colon x \mapsto (\phi_k(x))_{k\in\mathbb{K}}$ be the associated feature map. For every $k \in \mathbb{K}$, let $\eta_k \in \mathbb{R}_+$ and let $h_k \in \Gamma_0^+(\mathbb{R})$ be such that $h_k(0) = 0$. Define

$$G\colon \mathcal{F} \to [0, +\infty] \colon u = (\mu_k)_{k\in\mathbb{K}} \mapsto \sum_{k\in\mathbb{K}} g_k(\mu_k), \quad \text{where} \quad (\forall k \in \mathbb{K}) \quad g_k = h_k + \eta_k |\cdot|^r. \tag{4.20}$$

Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}_{++}$ such that $\lambda_n \to 0$ and let $(X_i, Y_i)_{i\in\mathbb{N}}$ be a sequence of independent copies of (X, Y). For every $n \in \mathbb{N}\smallsetminus\{0\}$, let $Z_n = (X_i, Y_i)_{1\leqslant i\leqslant n}$, and let $u_{n,\lambda_n}(Z_n)$ be defined according to (4.2) as an approximate minimizer of the regularized empirical risk

$$\frac{1}{n}\sum_{i=1}^{n} \ell\big(X_i, Y_i, (Au)(X_i)\big) + \lambda_n G(u). \tag{4.21}$$

The above model covers several classical regularization schemes, such as the Tikhonov (ridge regression) model [41], the ℓ_1 or lasso model [64], the elastic net model [30, 81], the bridge regression model [39, 44], as well as generalized Gaussian models [2]. Furthermore the following hold:

(i) $\mathcal{F}$ is uniformly convex with modulus of convexity of power type $\max\{2, r\}$ (see Section 2). Moreover, $\operatorname{ran} A \subset \mathcal{M}(\mathcal{X}, \mathsf{Y})$,

$$(\forall x \in \mathcal{X})(\forall u \in \mathcal{F}) \quad |\Phi(x)^* u| = |(Au)(x)| \leqslant \|u\|_r \|(\phi_k(x))_{k\in\mathbb{K}}\|_{r^*} \leqslant \varsigma \|u\|_r, \tag{4.22}$$

and therefore $\|\Phi\|_\infty \leqslant \varsigma$. Now suppose that $\inf_{k\in\mathbb{K}} \eta_k > 0$. Then, in view of Proposition A.9, G is totally convex on bounded sets. Altogether, Assumption 4.1 holds with $q = \max\{2, r\}$.

(ii) Let $p \in [1, +\infty]$ and suppose that one of the following holds:

(a) $\mathcal{C} = A\big(l^r(\mathbb{K}) \cap \times_{k\in\mathbb{K}} \operatorname{dom} h_k\big)$.

(b) $\mathcal{C} = \mathcal{M}(\mathcal{X}, \mathsf{Y})$ and $\operatorname{span}\{\phi_k\}_{k\in\mathbb{K}}$ is p-universal (Definition 3.6).

Then $\mathcal{C}$ is p-admissible (Definition 4.3), $A(\operatorname{dom} G) \subset \mathcal{C} \cap \operatorname{ran} A \subset \overline{A(\operatorname{dom} G)}$ (where the closure is in $L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$), and $\operatorname{ran} A$ is p-universal relative to $\mathcal{C}$. Indeed, as for (ii)(a), $\mathcal{C} \subset \operatorname{ran} A \subset L^p(\mathcal{X}, P_{\mathcal{X}}; \mathsf{Y})$, hence $\mathcal{C}$ is p-admissible and $\operatorname{ran} A$ is p-universal relative to $\mathcal{C}$. Moreover, $A(\operatorname{dom} G) \subset \mathcal{C} \subset \overline{A(\operatorname{dom} G)}$ since, for every $u \in l^r(\mathbb{K}) \cap \times_{k\in\mathbb{K}} \operatorname{dom} h_k$ and every $\epsilon \in \mathbb{R}_{++}$, there exists $\bar{u} \in \mathbb{R}^{\mathbb{K}}$ with finite support, such that $\|u - \bar{u}\|_r \leqslant \epsilon$ and $\|Au - A\bar{u}\|_p \leqslant \varsigma \|u - \bar{u}\|_r \leqslant \varsigma\epsilon$. On the other hand, it follows from Theorem 3.8(ii) that, if $\mathcal{C} = \mathcal{M}(\mathcal{X}, \mathsf{Y})$, (ii)(b) is satisfied when $\mathcal{X}$ is a locally compact topological space and $\operatorname{span}\{\phi_k\}_{k\in\mathbb{K}}$ is dense in $\mathscr{C}_0(\mathcal{X}, \mathsf{Y})$ endowed with the uniform topology.

(iii) Let $\mathcal{C}$ be as in item (ii)(a) or (ii)(b), let $\eta \in \mathbb{R}_{++}$, and suppose that $(\forall k \in \mathbb{K})\ \eta_k \geqslant \eta$. Then consistency can be obtained in the setting of Corollary 4.6, where $q = \max\{2, r\} = m$. In particular, in items (ii) and (iii) of Corollary 4.6, we have $\lambda_n^{pm/r} n^{1/q} = \lambda_n^p n^{1/r}$, if $r \geqslant 2$; and $\lambda_n^{pm/r} n^{1/q} = \lambda_n^{2p/r} n^{1/2}$, if $r \leqslant 2$. Moreover, by Theorem 4.4(iv), weak consistency holds if $1/(\lambda_n n^{1/\max\{2,r\}}) \to 0$, and strong consistency holds if $(\log n)/(\lambda_n n^{1/\max\{2,r\}}) \to 0$.

(iv) Suppose that $r \in \,]1, 2]$ and that the loss function is differentiable with respect to the third variable. Then, by exploiting the separability of G, for a given sample size n, an estimate $u_{n,\lambda_n}(z_n)$ can be constructed in $l^2(\mathbb{K})$ using proximal splitting algorithms such as those described in [27, 69].

Remark 4.11 Let us compare the results of Example 4.10 to the existing literature on generalized linear models.

(i) In the special case when $\mathbb{K}$ is finite, $r > 1$, and $G = \|\cdot\|_r^r$, [44] provides an excess risk bound which depends on the dimension of the dictionary (the cardinality of $\mathbb{K}$) and the level of sparsity of the regularized risk minimizer; see [15] for a recent account of the role of sparsity in regression.

(ii) In the special case when $r = 2$ and, for every $k \in \mathbb{K}$, $h_k = w_k|\cdot|$ with $w_k \in \mathbb{R}_{++}$ in (4.20), we recover the elastic net framework of [30]. This special case yields a strongly convex problem in a Hilbert space. In our general setting, the exponent r may take any value in $]1, +\infty[$. Note also that our framework allows for the enforcement of hard constraints on the coefficients since the functions $(h_k)_{k\in\mathbb{K}}$ are not required to be real-valued. We highlight that, when specialized to the elastic net regularizer, Theorem 4.4(iv) guarantees consistency under the same conditions as in [30, Theorem 2].

4.2 Proofs of the main results

We start with a few properties of the functions underlying our construct. To this end, throughout this subsection, the following notation will used.

Notation 4.12 In the setting of Assumption 4.1,

$$F = R \circ A \quad \text{and} \quad (\forall n \in \mathbb{N} \smallsetminus \{0\}) \quad F_n \colon \mathcal{F} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}_+ \colon (u, z) \mapsto R_n(Au, z). \tag{4.23}$$

In addition, $\varsigma = \|\Phi\|_\infty$, and, for every $n \in \mathbb{N} \smallsetminus \{0\}$ and $\lambda \in \mathbb{R}_{++}$,

$$\alpha_{n,\lambda} \colon \mathbb{R}_{++} \times \mathbb{R}_{++} \to \mathbb{R}_+$$
$$(\tau, \rho) \mapsto \frac{\varsigma \mathrm{Lip}(\ell; \varsigma\rho)}{\lambda} \left(\frac{4T_{q^*}}{n^{1/q}} + 2\sqrt{\frac{2\tau}{n}} + \frac{4\tau}{3n} \right). \tag{4.24}$$

Now let $\tau \in [1, +\infty[$ and $n \in \mathbb{N} \smallsetminus \{0\}$. Then, since $2\sqrt{2\tau} \leqslant 1 + 2\tau \leqslant 3\tau$ and $n^{1/q} \leqslant n^{1/2} \leqslant n$, we have

$$(\forall\, \rho \in \mathbb{R}_{++}) \quad \alpha_{n,\lambda}(\tau, \rho) \leqslant \frac{\tau\varsigma(4T_{q^*} + 5)\mathrm{Lip}(\ell; \varsigma\rho)}{\lambda n^{1/q}} \tag{4.25}$$

Proposition 4.13 *Suppose that Assumption 4.1 is satisfied. Then the following hold:*

(i) $F\colon \mathcal{F} \to \mathbb{R}_+$ *is convex and continuous.*

(ii) *Let* $n \in \mathbb{N} \smallsetminus \{0\}$ *and* $z \in (\mathcal{X} \times \mathcal{Y})^n$. *Then* $F_n(\cdot, z)\colon \mathcal{F} \to \mathbb{R}_+$ *is convex and continuous.*

(iii) G *is coercive and strictly convex.*

(iv) *For every* $\lambda \in \mathbb{R}_{++}$, $F + \lambda G$ *admits a unique minimizer.*

Proof. (i): Remark 3.17(iv) ensures that $R\colon L^\infty(\mathcal{X}, P_\mathcal{X}; \mathsf{Y}) \to \mathbb{R}_+$ is convex and continuous. In turn, Proposition 3.5(ii) implies that $A\colon \mathcal{F} \to L^\infty(\mathcal{X}, P_\mathcal{X}; \mathsf{Y})$ is continuous.

(ii): The argument is the same as above, except that P is replaced by the empirical measure $(1/n)\sum_{i=1}^n \delta_{(x_i,y_i)}$, where $z = (x_i, y_i)_{1 \leqslant i \leqslant n}$.

(iii): It follows from Assumption 4.1(vi) and Proposition A.6(ix) that G is coercive; its strict convexity follows from the definition in (2.8).

(iv): By (i) and (iii), $F + \lambda G$ is a strictly convex coercive function in $\Gamma_0^+(\mathcal{F})$. It therefore admits a unique minimizer [77, Theorem 2.5.1(ii) and Proposition 2.5.6]. ☐

The strategy of the proof of Theorem 4.4 is to split the error in three parts, i.e.,

$$\begin{aligned} &R(Au_{n,\lambda}(Z_n)) - \inf R(\mathcal{C}) \\ &= (F(u_{n,\lambda}(Z_n)) - F(u_\lambda)) + (F(u_\lambda) - \inf F(\operatorname{dom} G)) + (\inf F(\operatorname{dom} G) - \inf R(\mathcal{C})), \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{where} \quad u_\lambda = \operatorname{argmin}_{\mathcal{F}}(F + \lambda G). \end{aligned} \tag{4.26}$$

Note that Proposition 4.13(iv) ensures that u_λ is uniquely defined. The first term on the right-hand side of (4.26) is known as the sample error and the second term as the approximation error. Proposition A.12(ii) ensures that the approximation error goes to zero as $\lambda \to 0$. Below, we start by showing that $\inf R(\mathcal{C}) - \inf F(\operatorname{dom} G) = 0$, if $\operatorname{ran} A$ is universal with respect to $\mathcal{C}$ and some compatibility conditions between G and $\mathcal{C}$ hold. Next, we study the sample error. Note that $F(u_{n,\lambda}(Z_n)) - F(u_\lambda)$ may not be measurable, hence the convergence results are given with respect to the outer probability P^*.

Proposition 4.14 *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be nonempty sets, let* $(\mathcal{X} \times \mathcal{Y}, \mathfrak{A}, P)$ *be a probability space, let* $P_\mathcal{X}$ *be the marginal of* P *on* $\mathcal{X}$, *and let* Y *be a separable reflexive real Banach space. Let* $\ell \in \Upsilon(\mathcal{X} \times \mathcal{Y}, \mathsf{Y})$, *and let* $R\colon \mathcal{M}(\mathcal{X}, \mathsf{Y}) \to [0, +\infty]$ *be the risk associated with* ℓ *and* P. *Let* $\mathcal{C} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$ *be nonempty and convex. Let* $p \in [1, +\infty]$ *and assume that* $\mathcal{C}$ *is* p*-admissible and that there exists* $g \in \mathcal{C} \cap L^p(\mathcal{X}, P_\mathcal{X}; \mathsf{Y})$ *such that* $R(g) < +\infty$. *Then* $\inf R(\mathcal{C}) = \inf R(\mathcal{C} \cap L^p(\mathcal{X}, P_\mathcal{X}; \mathsf{Y}))$.

Proof. Suppose that $\mathcal{C} = \big\{f \in \mathcal{M}(\mathcal{X}, \mathsf{Y}) \bigm| (\forall x \in \mathcal{X})\ f(x) \in \mathsf{C}(x)\big\}$. Let $f \in \mathcal{C}$ be such that $R(f) < +\infty$. For every $n \in \mathbb{N}$, set $A_n = \big\{x \in \mathcal{X} \bigm| |f(x)| \leqslant n\big\}$, let A_n^c be its complement, and define $f_n\colon \mathcal{X} \to \mathsf{Y}$, $f_n = \mathbf{1}_{A_n} f + \mathbf{1}_{A_n^c} g$. For every $n \in \mathbb{N}$ and $x \in \mathcal{X}$, $f_n(x) \in \mathsf{C}(x)$ and $|f_n(x)| \leqslant \max\{n, |g(x)|\}$, hence $f_n \in \mathcal{C} \cap L^p(\mathcal{X}, P_\mathcal{X}; \mathsf{Y})$. Moreover,

$$(\forall n \in \mathbb{N}) \quad |R(f_n) - R(f)| \leqslant \int_{A_n^c \times \mathcal{Y}} |\ell(x, y, g(x)) - \ell(x, y, f(x))| P(\mathrm{d}(x, y)). \tag{4.27}$$

Set $h\colon (x, y) \mapsto |\ell(x, y, g(x)) - \ell(x, y, f(x))|$. Since $R(f) < +\infty$ and $R(g) < +\infty$, we have $h \in L^1(\mathcal{X} \times \mathcal{Y}, P)$. Since $\mathbf{1}_{A_n^c \times \mathcal{Y}} h \to 0$ pointwise and $\mathbf{1}_{A_n^c \times \mathcal{Y}} h \leqslant h$, it follows from the dominated convergence theorem that the right-hand side of (4.27) tends to zero, and hence $R(f_n) \to R(f)$. This implies that $\inf R(\mathcal{C} \cap L^p(\mathcal{X}, P_\mathcal{X}; \mathsf{Y})) \leqslant R(f)$. ☐

Proposition 4.15 *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets, let $(\mathcal{X}\times\mathcal{Y},\mathfrak{A},P)$ be a probability space, let $P_{\mathcal{X}}$ be the marginal of P on $\mathcal{X}$, and let Y be a separable reflexive real Banach space. Let $\mathcal{C}\subset\mathcal{M}(\mathcal{X},\mathcal{Y})$ be nonempty and convex and let $p\in[1,+\infty]$. Suppose that $\ell\in\Upsilon_p(\mathcal{X}\times\mathcal{Y},\mathsf{Y},P)$, that $\Phi\in L^p[\mathcal{X},P_{\mathcal{X}};\mathscr{L}(\mathsf{Y}^*,\mathcal{F}^*)]$, and that $A(\operatorname{dom}G)\subset\mathcal{C}\cap\operatorname{ran}A\subset\overline{A(\operatorname{dom}G)}$, where the closure is in $L^p(\mathcal{X},P_{\mathcal{X}};\mathsf{Y})$. Let $R\colon\mathcal{M}(\mathcal{X},\mathsf{Y})\to[0,+\infty]$ be the risk associated with ℓ and P. Then the following hold:*

(i) $\inf F(\operatorname{dom}G)=\inf R(\mathcal{C}\cap\operatorname{ran}A)$.

(ii) *Suppose that $\mathcal{C}$ is p-admissible and $\operatorname{ran}A$ is p-universal relative to $\mathcal{C}$. Then* $\inf F(\operatorname{dom}G)=\inf R(\mathcal{C})$.

Proof. (i): By Remark 3.17(i), R is continuous on $L^p(\mathcal{X},P_{\mathcal{X}};\mathsf{Y})$ and hence $\inf R(A(\operatorname{dom}G))=\inf R(\overline{A(\operatorname{dom}G)})$. Therefore, since $A(\operatorname{dom}G)\subset\mathcal{C}\cap\operatorname{ran}A\subset\overline{A(\operatorname{dom}G)}$, the assertion follows.

(ii): Suppose first that $p<+\infty$. Since R is continuous on $L^p(\mathcal{X},P_{\mathcal{X}};\mathsf{Y})$ and $\mathcal{C}\cap\operatorname{ran}A$ is dense in $\mathcal{C}\cap L^p(\mathcal{X},P_{\mathcal{X}};\mathsf{Y})$, $\inf R(\mathcal{C}\cap\operatorname{ran}A)=\inf R(\mathcal{C}\cap L^p(\mathcal{X},P_{\mathcal{X}};\mathsf{Y}))$. Thus, since $\mathcal{C}$ is p-admissible, Proposition 4.14 gives $\inf R(\mathcal{C}\cap L^p(\mathcal{X},P_{\mathcal{X}};\mathsf{Y}))=\inf R(\mathcal{C})$ and hence $\inf R(\mathcal{C}\cap\operatorname{ran}A)=\inf R(\mathcal{C})$. The statement follows from (i). Now suppose that $p=+\infty$. Let $f\in\mathcal{C}\cap L^\infty(\mathcal{X},P_{\mathcal{X}};\mathsf{Y})$. By Definition 3.6(i), there exists $(f_n)_{n\in\mathbb{N}}\in(\mathcal{C}\cap\operatorname{ran}A)^{\mathbb{N}}$ and $\rho\in\mathbb{R}_{++}$ such that $\sup_{n\in\mathbb{N}}\|f_n\|_\infty\leqslant\rho$ and $f_n\to f$ $P_{\mathcal{X}}$-a.s. It follows from (3.34) that $(\exists\, g_\rho\in L^1(\mathcal{X}\times\mathcal{Y},P;\mathbb{R}))(\forall(x,y)\in\mathcal{X}\times\mathcal{Y})$ $|\ell(x,y,f_n(x))-\ell(x,y,f(x))|\leqslant 2g_\rho(x,y)$. By the dominated convergence theorem, $R(f_n)\to R(f)$. Thus, $\inf R(\mathcal{C}\cap\operatorname{ran}A)=\inf R(\mathcal{C}\cap L^\infty(\mathcal{X},P_{\mathcal{X}};\mathsf{Y}))$ and we conclude as above. ◻

Proposition 4.16 *Suppose that Assumption 4.1 holds and that Notation 4.12 is in use. Write $\varepsilon=\varepsilon_1\varepsilon_2$, where ε_1 and ε_2 are functions from $\mathbb{R}_{++}$ to $[0,1]$, let $\lambda\in\mathbb{R}_{++}$, and define $u_\lambda=\operatorname{argmin}_{\mathcal{F}}(F+\lambda G)$. Let $\tau\in\mathbb{R}_{++}$, let $n\in\mathbb{N}\smallsetminus\{0\}$, and let $\rho\in[\|u_\lambda\|,+\infty[$. Then the following hold:*

(i) $\mathsf{P}^*\left[\|u_{n,\lambda}(Z_n)-u_\lambda\|\geqslant\varepsilon_1(\lambda)+(\widehat{\psi}_\rho)^\natural\left(\alpha_{n,\lambda}(\tau,\rho)+\dfrac{\varepsilon_2(\lambda)}{\lambda}\right)\right]\leqslant e^{-\tau}$.

(ii) $\mathsf{P}^*\left(\Big[\|u_{n,\lambda}(Z_n)\|\leqslant\rho\Big]\cap\left[F(u_{n,\lambda}(Z_n))-F(u_\lambda)\geqslant\varsigma\mathrm{Lip}(\ell;\varsigma\rho)\left(\varepsilon_1(\lambda)+(\widehat{\psi}_\rho)^\natural\left(\alpha_{n,\lambda}(\tau,\rho)+\dfrac{\varepsilon_2(\lambda)}{\lambda}\right)\right)\right]\right)$ $\leqslant e^{-\tau}$.

(iii) *Suppose that $\ell\in\Upsilon_1(\mathcal{X}\times\mathcal{Y}\times\mathsf{Y},P)$ and let $c\in\mathbb{R}_+$ be as in Definition 3.16(i). Then*

$$\mathsf{P}^*\left[F(u_{n,\lambda}(Z_n))-F(u_\lambda)\geqslant\varsigma c\left(\varepsilon_1(\lambda)+(\widehat{\psi}_\rho)^\natural\left(\alpha_{n,\lambda}(\tau,\rho)+\frac{\varepsilon_2(\lambda)}{\lambda}\right)\right)\right]\leqslant e^{-\tau}.\tag{4.28}$$

Proof. (i): Let $z=(x_i,y_i)_{1\leqslant i\leqslant n}\in(\mathcal{X}\times\mathcal{Y})^n$. Since

$$u_{n,\lambda}(z)\in\operatorname{Argmin}_{\mathcal{F}}^{\varepsilon_1(\lambda)\varepsilon_2(\lambda)}(F_n(\cdot,z)+\lambda G),\tag{4.29}$$

it follows from Proposition 4.13(ii) and Ekeland's variational principle [47, Corollary 4.2.12] that there exists $v_{n,\lambda}\in\mathcal{F}$ such that $\|u_{n,\lambda}(z)-v_{n,\lambda}\|\leqslant\varepsilon_1(\lambda)$ and $\inf\|\partial(F_n(\cdot,z)+\lambda G)(v_{n,\lambda})\|\leqslant\varepsilon_2(\lambda)$. We note that $\ell\in\Upsilon_\infty(\mathcal{X}\times\mathcal{Y}\times\mathsf{Y})$ by Remark 3.17(iv). Hence, setting $\widetilde{P}=(1/n)\sum_{i=1}^n\delta_{(x_i,y_i)}$, we derive from Theorems 3.19(ii) and 3.25(ii) that there exists a measurable and P-a.s. bounded function $h_\lambda\colon\mathcal{X}\times\mathcal{Y}\to\mathsf{Y}^*$ such that $\|h_\lambda\|_\infty\leqslant\mathrm{Lip}(\ell;\varsigma\rho)$ and

$$\|v_{n,\lambda}-u_\lambda\|\leqslant(\widehat{\psi}_\rho)^\natural\left(\frac{1}{\lambda}\Big\|\mathsf{E}_P[\Phi h_\lambda]-\frac{1}{n}\sum_{i=1}^n\Phi(x_i)h_\lambda(x_i,y_i)\Big\|+\frac{\varepsilon_2(\lambda)}{\lambda}\right).\tag{4.30}$$

Thus, for every $z \in (\mathcal{X} \times \mathcal{Y})^n$

$$\|u_{n,\lambda}(z) - u_\lambda\| \leqslant \varepsilon_1(\lambda) + (\widehat{\psi}_\rho)^\natural\bigg(\frac{1}{\lambda}\Big\|\mathsf{E}_P[\Phi h_\lambda] - \frac{1}{n}\sum_{i=1}^n \Phi(x_i)h_\lambda(x_i,y_i)\Big\| + \frac{\varepsilon_2(\lambda)}{\lambda}\bigg). \tag{4.31}$$

Now consider the family of i.i.d. random vectors $(\Phi(X_i)h_\lambda(X_i,Y_i))_{1\leqslant i\leqslant n}$, from Ω to $\mathcal{F}^*$. Since $\max_{1\leqslant i\leqslant n}\|\Phi(X_i)h_\lambda(X_i,Y_i)\| \leqslant \varsigma\mathrm{Lip}(\ell;\varsigma\rho)$ P-a.s., Theorem A.17 gives

$$\mathsf{P}\bigg[\Big\|\mathsf{E}_{\mathsf{P}}[\Phi(X)h_\lambda(X,Y)] - \frac{1}{n}\sum_{i=1}^n \Phi(X_i)h_\lambda(X_i,Y_i)\Big\| \geqslant \lambda\alpha_{n,\lambda}(\tau,\rho)\bigg] \leqslant e^{-\tau}. \tag{4.32}$$

Hence, since $(\widehat{\psi}_\rho)^\natural$ is increasing by Proposition A.5(vii), a fortiori we have

$$\begin{aligned}\mathsf{P}\bigg[\varepsilon_1(\lambda) + (\widehat{\psi}_\rho)^\natural\bigg(\frac{1}{\lambda}\Big\|\mathsf{E}_P[\Phi h_\lambda] - \frac{1}{n}\sum_{i=1}^n \Phi(X_i)h_\lambda(X_i,Y_i)\Big\| + \frac{\varepsilon_2(\lambda)}{\lambda}\bigg)\\ \geqslant \varepsilon_1(\lambda) + (\widehat{\psi}_\rho)^\natural\bigg(\alpha_{n,\lambda}(\tau,\rho) + \frac{\varepsilon_2(\lambda)}{\lambda}\bigg)\bigg] \leqslant e^{-\tau}.\end{aligned} \tag{4.33}$$

Thus (i) follows from (4.31) and (4.33).

(ii): Let $\omega \in \big[\|u_{n,\lambda}(Z_n)\| \leqslant \rho\big]$. Since $\|u_\lambda\| \leqslant \rho$ and $\|u_{n,\lambda}(Z_n(\omega))\| \leqslant \rho$, we have $\|Au_\lambda\|_\infty \leqslant \varsigma\rho$ and $\|Au_{n,\lambda}(Z_n(\omega))\|_\infty \leqslant \varsigma\rho$. Hence, we derive from Assumption 4.1(ii) that $F(u_{n,\lambda}(Z_n(\omega))) - F(u_\lambda) \leqslant \mathrm{Lip}(\ell;\varsigma\rho)\|Au_{n,\lambda}(Z_n(\omega)) - Au_\lambda\|_\infty \leqslant \varsigma\mathrm{Lip}(\ell;\varsigma\rho)\|u_{n,\lambda}(Z_n(\omega)) - u_\lambda\|$. Thus, (ii) follows from (i).

(iii): It follows from Remark 3.17(v)(a) that ℓ is globally Lipschitz continuous in the third variable uniformly with respect to the first two and that $\sup_{\rho'\in\mathbb{R}_{++}}\mathrm{Lip}(\ell;\rho') \leqslant c$. Hence, we derive from (3.32) that R is Lipschitz continuous on $L^1(\mathcal{X}, P_{\mathcal{X}};\mathsf{Y})$ with Lipschitz constant c. As a result,

$$(\forall\omega\in\Omega)\quad F(u_{n,\lambda}(Z_n(\omega))) - F(u_\lambda) \leqslant c\|Au_{n,\lambda}(Z_n(\omega)) - Au_\lambda\|_\infty \leqslant \varsigma c\|u_{n,\lambda}(Z_n(\omega)) - u_\lambda\|. \tag{4.34}$$

Thus, the statement follows from (i). ☐

The following technical result will be required subsequently.

Lemma 4.17 *Let $\alpha\colon \mathbb{R}_+ \to \mathbb{R}_+$ and let $\gamma \in \mathbb{R}_{++}$ be such that, for every $\tau \in \left]1,+\infty\right[$, $\alpha(\tau) \leqslant \gamma\tau$. Let $\phi \in \mathcal{A}_0$, let $(\eta,\epsilon) \in \mathbb{R}_{++}\times\mathbb{R}_+$, and suppose that $\phi^\natural(2\gamma) < \eta$ and $\phi^\natural(2\epsilon) < \eta$. Set $\tau_0 = \phi(\eta^-)/(2\gamma)$. Then $\phi^\natural(\alpha(\tau_0)+\epsilon) < \eta$.*

Proof. Recalling Proposition A.5(vi), we derive from the inequalities $\phi^\natural(2\gamma) < \eta$ and $\phi^\natural(2\epsilon) < \eta$ that $\tau_0 > 1$ and $\phi(\eta^-) > 2\epsilon$, respectively. Therefore, since $\gamma\tau_0 = \phi(\eta^-)/2$, we have $\alpha(\tau_0) + \epsilon \leqslant \tau_0\gamma + \epsilon = \phi(\eta^-)/2 + \epsilon < \phi(\eta^-)$. Again, by Proposition A.5(vi), we obtain that $\phi^\natural(\alpha(\tau_0)+\epsilon) < \eta$. ☐

Proposition 4.18 *Suppose that Assumption 4.1 holds, that Notation 4.12 is in use, and that $\ell(\cdot,\cdot,0)$ is bounded. Write $\varepsilon = \varepsilon_1\varepsilon_2$, where ε_1 and ε_2 are functions from $\mathbb{R}_{++}$ to $[0,1]$. Let $(\forall n\in\mathbb{N})$ $\rho_n \in \big[\psi_0^\natural\big((\|\ell(\cdot,\cdot,0)\|_\infty + 1)/\lambda_n\big), +\infty\big[$. Then the following hold:*

(i) *Let $\lambda \in \mathbb{R}_{++}$, and set $u_\lambda = \mathrm{argmin}_{\mathcal{F}}(F + \lambda G)$ and let $\rho \in \big[\psi_0^\natural\big((\|\ell(\cdot,\cdot,0)\|_\infty + 1)/\lambda\big), +\infty\big[$. Let $\tau \in \mathbb{R}_{++}$ and let $n \in \mathbb{N}\smallsetminus\{0\}$. Then*

$$\begin{aligned}\mathsf{P}^*\Big[F(u_{n,\lambda}(Z_n)) - \inf F(\mathrm{dom}\,G) \geqslant \varsigma\mathrm{Lip}(\ell;\varsigma\rho)\big(\varepsilon_1(\lambda) + (\widehat{\psi}_\rho)^\natural\big(\alpha_{n,\lambda}(\tau,\rho) + \varepsilon_2(\lambda)/\lambda\big)\big)\\ + F(u_\lambda) - \inf F(\mathrm{dom}\,G)\Big] \leqslant e^{-\tau}.\end{aligned} \tag{4.35}$$

(ii) *Suppose that* (4.5) *and* (4.6) *hold. Then* $F(u_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf F(\operatorname{dom} G)$.

(iii) *Suppose that* (4.5) *and* (4.7) *hold. Then* $F(u_{n,\lambda_n}(Z_n)) \to \inf F(\operatorname{dom} G)$ P^**–a.s.*

Proof. (i): Since for every $z_n = (x_i, y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$, $F_n(0, z_n) \leqslant \|\ell(\cdot,\cdot,0)\|_\infty$ and $F(0) \leqslant \|\ell(\cdot,\cdot,0)\|_\infty$, it follows from Proposition A.16 that $\|u_{n,\lambda}(Z_n)\| \leqslant \rho$ and $\|u_\lambda\| \leqslant \rho$. Thus, Proposition 4.16(ii) yields $\mathsf{P}^*\big[F(u_{n,\lambda}(Z_n)) - F(u_\lambda) \geqslant \varsigma \mathrm{Lip}(\ell;\varsigma\rho)\big(\varepsilon_1(\lambda) + (\widehat{\psi}_\rho)^\natural\big(\alpha_{n,\lambda}(\tau,\rho) + \varepsilon_2(\lambda)/\lambda\big)\big)\big] \leqslant e^{-\tau}$, and (4.35) follows.

(ii): Because of (4.25), conditions (4.5)-(4.6) imply that

$$(\forall \tau \in [1,+\infty[) \quad \varsigma \mathrm{Lip}(\ell;\varsigma\rho_n)\big(\varepsilon_1(\lambda_n) + (\widehat{\psi}_{\rho_n})^\natural\big(\alpha_{n,\lambda_n}(\tau,\rho_n) + \varepsilon_2(\lambda_n)/\lambda_n\big)\big) \to 0. \tag{4.36}$$

Therefore, it follows from (4.35) and Proposition A.12(ii) that for every $(\eta,\tau) \in \mathbb{R}_{++} \times [1+\infty[$, there exists $\bar{n} \in \mathbb{N}$ such that, for every integer $n \geqslant \bar{n}$, $\mathsf{P}^*\big[F(u_{n,\lambda_n}(Z_n)) - \inf F(\operatorname{dom} G) \geqslant \eta\big] \leqslant e^{-\tau}$. Hence, for every $(\eta,\tau) \in \mathbb{R}_{++} \times [1,+\infty[$, $\overline{\lim}_{n\to+\infty} \mathsf{P}^*\big[F(u_{n,\lambda_n}(Z_n)) - \inf F(\operatorname{dom} G) \geqslant \eta\big] \leqslant e^{-\tau}$. The convergence in outer probability follows.

(iii): Let $\eta \in \mathbb{R}_{++}$ and let $\xi \in]1,+\infty[$. It follows from (4.5) and (4.7) that there exists an integer $\bar{n} \geqslant 3$ such that, for every integer $n \geqslant \bar{n}$, we have

$$\mathrm{Lip}(\ell;\varsigma\rho_n)(\widehat{\psi}_{\rho_n})^\natural\Big(\frac{2\varsigma\xi(4T_{q^*}+5)\mathrm{Lip}(\ell;\varsigma\rho_n)\log n}{\lambda_n n^{1/q}}\Big) < \eta \quad \text{and} \quad \mathrm{Lip}(\ell;\varsigma\rho_n)(\widehat{\psi}_{\rho_n})^\natural\Big(2\frac{\varepsilon_2(\lambda_n)}{\lambda_n}\Big) < \eta\,. \tag{4.37}$$

Let $n \in \mathbb{N}$ be such that $n \geqslant \bar{n}$ and set $\gamma = \varsigma(4T_{q^*}+5)\mathrm{Lip}(\ell;\varsigma\rho_n)/(\lambda_n n^{1/q})$. We derive from (4.25) that $(\forall \tau \in [1,+\infty)\ \alpha_{n,\lambda_n}(\tau,\rho_n) \leqslant \tau\gamma$. Then, since $1 \leqslant \xi \log n$, it follows from Lemma 4.17 that

$$\tau_0 = \widehat{\psi}_{\rho_n}\bigg(\Big(\frac{\eta}{\mathrm{Lip}(\ell;\varsigma\rho_n)}\Big)^-\bigg)\frac{\lambda_n n^{1/q}}{2\varsigma(4T_{q^*}+5)\mathrm{Lip}(\ell;\varsigma\rho_n)} \quad\Rightarrow\quad \mathrm{Lip}(\ell;\varsigma\rho_n)(\widehat{\psi}_{\rho_n})^\natural\bigg(\alpha_{n,\lambda_n}(\tau_0,\rho_n) + \frac{\varepsilon_2(\lambda_n)}{\lambda_n}\bigg) < \eta. \tag{4.38}$$

Now set

$$\Omega_{n,\eta} = \big[F(u_{n,\lambda_n}(Z_n)) - \inf F(\operatorname{dom} G) > \varsigma\mathrm{Lip}(\ell;\varsigma\rho_n)\varepsilon_1(\lambda_n) + \varsigma\eta + F(u_{\lambda_n}) - \inf F(\operatorname{dom} G)\big]. \tag{4.39}$$

Item (i) yields

$$\mathsf{P}^*\Omega_{n,\eta} \leqslant \exp\bigg(-\widehat{\psi}_{\rho_n}\bigg(\Big(\frac{\eta}{\mathrm{Lip}(\ell;\varsigma\rho_n)}\Big)^-\bigg)\frac{\lambda_n n^{1/q}}{2\varsigma(4T_{q^*}+5)\mathrm{Lip}(\ell;\varsigma\rho_n)}\bigg). \tag{4.40}$$

We remark that, by Proposition A.5(vi)-(vii), the first condition in (4.37) is equivalent to

$$\widehat{\psi}_{\rho_n}\bigg(\Big(\frac{\eta}{\mathrm{Lip}(\ell;\varsigma\rho_n)}\Big)^-\bigg)\frac{\lambda_n n^{1/q}}{2\varsigma(4T_{q^*}+5)\mathrm{Lip}(\ell;\varsigma\rho_n)} > \xi \log n\,. \tag{4.41}$$

Thus it follows from (4.40) and (4.41) that $\sum_{n=\bar{n}}^{+\infty} \mathsf{P}^*\Omega_{n,\eta} \leqslant \sum_{n=\bar{n}}^{+\infty} 1/n^\xi < +\infty$. Hence, using the Borel-Cantelli lemma (which holds for outer measures too) we conclude that $F(u_{n,\lambda_n}(Z_n)) \to \inf F(\operatorname{dom} G)$ P^*–a.s. ☐

The next proposition considers the case of a globally Lipschitz continuous loss ℓ, and does not require the boundedness of $\ell(\cdot,\cdot,0)$.

Proposition 4.19 *Suppose that Assumption 4.1 holds, that Notation 4.12 is in use, and that* $\ell \in \Upsilon_1(\mathcal{X} \times \mathcal{Y} \times \mathsf{Y}; P)$. *Let* $c \in \mathbb{R}_+$ *be as in Definition 3.16(i) and write* $\varepsilon = \varepsilon_1\varepsilon_2$, *where* ε_1 *and* ε_2 *are functions from* $\mathbb{R}_{++}$ *to* $[0,1]$. *Let* $(\forall n \in \mathbb{N})$ $\rho_n \in \left[\psi_0^\natural((R(0)+1)/\lambda_n), +\infty\right[$. *Then the following hold:*

(i) *Let* $\lambda \in \mathbb{R}_{++}$, *set* $u_\lambda = \operatorname{argmin}_{\mathcal{F}}(F + \lambda G)$ *and let* $\rho \in \left[\psi_0^\natural\big((F(0)+1)/\lambda\big), +\infty\right[$. *Let* $\tau \in \mathbb{R}_{++}$ *and let* $n \in \mathbb{N} \smallsetminus \{0\}$. *Then*

$$\mathsf{P}^*\Big[F(u_{n,\lambda}(Z_n)) - \inf F(\operatorname{dom} G) \geqslant \varsigma c\big(\varepsilon_1(\lambda) + (\widehat{\psi}_\rho)^\natural\big(\alpha_{n,\lambda}(\tau,\rho) + \varepsilon_2(\lambda)/\lambda\big)\big) + F(u_\lambda) - \inf F(\operatorname{dom} G)\Big] \leqslant e^{-\tau}. \tag{4.42}$$

(ii) *Suppose that* (4.10) *holds. Then* $F(u_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf F(\operatorname{dom} G)$.

(iii) *Suppose that* (4.10) *and* (4.11) *hold. Then* $F(u_{n,\lambda_n}(Z_n)) \to \inf F(\operatorname{dom} G)$ P^**-a.s.*

Proof. (i): By Proposition A.16, $\|u_\lambda\| \leqslant \rho$. Thus, (4.42) follows from Proposition 4.16(iii).

(ii)-(iii): Using (i), these can be established as in the proof of Proposition 4.18(ii)–(iii). ☐

Proposition 4.20 *Suppose that Assumption 4.1 holds, that Notation 4.12 is in use, and that* $S = \operatorname{Argmin}_{\operatorname{dom} G} F \neq \varnothing$. *Let* $u^\dagger = \operatorname{argmin}_{u \in S} G(u)$ *and write* $\varepsilon = \varepsilon_1\varepsilon_2$, *where* ε_1 *and* ε_2 *are functions from* $\mathbb{R}_{++}$ *to* $[0,1]$. *For every* $\lambda \in \mathbb{R}_{++}$, *set* $u_\lambda = \operatorname{argmin}_{\mathcal{F}}(F + \lambda G)$. *Let* $\rho \in \left]\sup_{\lambda \in \mathbb{R}_{++}} \|u_\lambda\|, +\infty\right[$ *and let* $\tau \in \mathbb{R}_{++}$. *Then, for every sufficiently small* $\lambda \in \mathbb{R}_{++}$ *and every* $n \in \mathbb{N} \smallsetminus \{0\}$,

$$\mathsf{P}^*\left[\|u_{n,\lambda}(Z_n) - u^\dagger\| \geqslant \varepsilon_1(\lambda) + (\widehat{\psi}_\rho)^\natural\left(\alpha_{n,\lambda}(\tau,\rho) + \frac{\varepsilon_2(\lambda)}{\lambda}\right) + \|u_\lambda - u^\dagger\|\right] \leqslant e^{-\tau}. \tag{4.43}$$

Moreover, assume that (4.12) *is satisfied. Then the following hold:*

(i) *For every sufficiently large* $n \in \mathbb{N}$,

$$\mathsf{P}^*\left[F(u_{n,\lambda_n}(Z_n)) - F(u^\dagger) \geqslant \varsigma \operatorname{Lip}(\ell; \varsigma\rho)\left(\varepsilon_1(\lambda_n) + (\widehat{\psi}_\rho)^\natural\left(\alpha_{n,\lambda_n}(\tau,\rho) + \frac{\varepsilon_2(\lambda_n)}{\lambda_n}\right)\right) + \lambda_n\right] \leqslant 2e^{-\tau}. \tag{4.44}$$

(ii) $u_{n,\lambda}(Z_n) \xrightarrow{\mathsf{P}^*} u^\dagger$ *and* $F(u_{n,\lambda_n}(Z_n)) \xrightarrow{\mathsf{P}^*} \inf F(\operatorname{dom} G)$.

(iii) *Suppose that* (4.13) *holds. Then* $F(u_{n,\lambda_n}(Z_n)) \to \inf F(\operatorname{dom} G)$ P^**-a.s. and* $u_{n,\lambda_n}(Z_n) \to u^\dagger$ P^**-a.s.*

Proof. First note that items (i) and (v) in Proposition A.14 imply that $u^\dagger$ is well defined and that $\sup_{\lambda \in \mathbb{R}_{++}} \|u_\lambda\| < +\infty$. Now, let $\lambda \in \mathbb{R}_{++}$ and let $n \in \mathbb{N}$. Since $\|u_\lambda\| \leqslant \rho$, it follows from Proposition 4.16(i) that

$$\mathsf{P}^*\Big[\|u_{n,\lambda}(Z_n) - u_\lambda\| \geqslant \varepsilon_1(\lambda) + (\widehat{\psi}_\rho)^\natural\big(\alpha_{n,\lambda}(\tau,\rho) + \varepsilon_2(\lambda_n)/\lambda_n\big)\Big] \leqslant e^{-\tau} \tag{4.45}$$

and, since $\|u_{n,\lambda}(Z_n) - u^\dagger\| \leqslant \|u_{n,\lambda}(Z_n) - u_\lambda\| + \|u_\lambda - u^\dagger\|$, (4.43) follows. Note also that Proposition A.6(viii) implies that $\widehat{\psi}_\rho \in \mathcal{A}_0$.

(i): Let $\eta \in \mathbb{R}_{++}$ be such that $\sup_{\lambda \in \mathbb{R}_{++}} \|u_\lambda\| + \eta \leqslant \rho$. It follows from (4.12), (4.25), and Proposition A.5(v), that $\varepsilon_1(\lambda_n) + (\widehat{\psi}_\rho)^\natural\big(\alpha_{n,\lambda_n}(\tau,\rho) + \varepsilon_2(\lambda_n)/\lambda_n\big) \to 0$. Hence, there exists $\bar{n} \in \mathbb{N}$ such that for every integer $n \geqslant \bar{n}$, $\varepsilon_1(\lambda_n)+(\widehat{\psi}_\rho)^\natural\big(\alpha_{n,\lambda_n}(\tau,\rho)+\varepsilon_2(\lambda_n)/\lambda_n\big) \leqslant \eta$. Now, take an integer $n \geqslant \bar{n}$ and set $\Omega_n = \big[\|u_{n,\lambda_n}(Z_n) - u_{\lambda_n}\| \leqslant \eta\big]$. Then $\Omega_n \subset \big[\|u_{n,\lambda_n}(Z_n)\| \leqslant \rho\big]$ and it follows from (4.45) that $\mathsf{P}^*(\Omega \setminus \Omega_n) \leqslant e^{-\tau}$. Hence, we deduce from Proposition 4.16(ii) that

$$\mathsf{P}^*\Big[F(u_{n,\lambda_n}(Z_n)) - F(u_{\lambda_n}) \geqslant \varsigma \mathrm{Lip}(\ell;\varsigma\rho)\Big(\varepsilon_1(\lambda_n)+(\widehat{\psi}_\rho)^\natural\big(\alpha_{n,\lambda}(\tau,\rho)+\varepsilon_2(\lambda_n)/\lambda_n\big)\Big)\Big] \leqslant 2e^{-\tau}. \quad (4.46)$$

On the other hand, Proposition A.14(iv) implies that, for n sufficiently large, $F(u_{\lambda_n}) - F(u^\dagger) \leqslant \lambda_n$, which combined with (4.46) gives (4.44).

(ii): This follows from (4.43) and (4.44), as in the proof of Proposition 4.18(ii).

(iii): Let $\eta \in \mathbb{R}_{++}$ and $\xi \in \,]1,+\infty[$. Using (4.12) and (4.13) we obtain a version of (4.37) in which $\rho_n \equiv \rho$. The proof of the fact that $F(u_{n,\lambda_n}(Z_n)) \to F(u^\dagger)$ P^*–a.s. then follows the same line as that of Proposition 4.18(iii). Next, let $n \in \mathbb{N}$ be sufficiently large so that

$$(\widehat{\psi}_\rho)^\natural\bigg(\frac{2\varsigma\xi(4T_{q^*}+5)\mathrm{Lip}(\ell;\varsigma\rho)\log n}{\lambda_n n^{1/q}}\bigg) < \eta \quad \text{and} \quad (\widehat{\psi}_\rho)^\natural\bigg(\frac{2\varepsilon_2(\lambda_n)}{\lambda_n}\bigg) < \eta\,. \quad (4.47)$$

Using Lemma 4.17, upon setting $\tau_0 = (\widehat{\psi}_\rho(\eta^-)\lambda_n n^{1/q}/(2\varsigma(4T_{q^*}+5)\mathrm{Lip}(\ell;\varsigma\rho))$, we obtain $(\widehat{\psi}_\rho)^\natural\big(\alpha_{n,\lambda_n}(\tau_0,\rho)+\varepsilon_2(\lambda_n)/\lambda_n\big) < \eta$. It then follows from (4.43) and (4.47) that, for n sufficiently large,

$$\mathsf{P}^*\Big[\|u_{n,\lambda_n}(Z_n) - u^\dagger\| > \varepsilon_1(\lambda_n)+\eta+\|u_{\lambda_n} - u^\dagger\|\Big] \leqslant \exp\bigg(-\frac{\widehat{\psi}_\rho(\eta^-)\lambda_n n^{1/q}}{2\varsigma(4T_{q^*}+5)\mathrm{Lip}(\ell;\varsigma\rho)}\bigg) < \frac{1}{n^\xi}. \quad (4.48)$$

The conclusion follows by the Borel-Cantelli lemma. ☐

Proof of Theorem 4.4. We first note that Proposition 4.15(ii) asserts that $\inf F(\operatorname{dom} G) = \inf R(\mathcal{C})$.

(i): This follows from Proposition 4.18(ii)–(iii).

(ii): Remark 3.17(v)(b) implies that, for every $\rho \in \mathbb{R}_{++}$, $\mathrm{Lip}(\ell;\rho) \leqslant (p-1)\|b\|_\infty + 3cp\max\{1,\rho^{p-1}\}$ and $\ell(\cdot,\cdot,0)$ is bounded. Hence conditions (4.8) and (4.9) imply (4.5)-(4.6) and (4.7) respectively. Therefore, the statement follows from (i).

(iii): This follows from Proposition 4.19(ii)–(iii).

(iv): This follows from Proposition 4.20(ii)–(iii). ☐

Proof of Corollary 4.6. Since $\mathcal{F}$ is uniformly convex of power type q, $\mathcal{F}^*$ is uniformly smooth with modulus of smoothness of power type q^* [46, p. 63] and hence of Rademacher type q^* (see Section 2) in conformity with Assumption 4.1(iv). Moreover, by (4.14), the modulus of total convexity ψ_ρ of G on $B(\rho)$ is greater then that of $\eta\|\cdot\|^r$. Hence, by Proposition A.9,

$$(\forall \rho \in \mathbb{R}_+)(\forall t \in \mathbb{R}_+) \qquad \psi_\rho(t) \geqslant \begin{cases} \eta\beta t^r & \text{if } r \geqslant q \\[2mm] \dfrac{\eta\beta t^q}{(\rho+t)^{q-r}} & \text{if } r < q \end{cases} \quad (4.49)$$

and, for every $\rho \in \mathbb{R}_+$ and every $s \in \mathbb{R}_+$,

$$(\widehat{\psi}_\rho)^\natural(s) \leqslant \begin{cases} \left(\dfrac{s}{\eta\beta}\right)^{1/(r-1)} & \text{if } r \geqslant q \\ 2^q\rho\max\left\{\left(\dfrac{s}{\eta\beta\rho^{r-1}}\right)^{1/(q-1)}, \left(\dfrac{s}{\eta\beta\rho^{r-1}}\right)^{1/(r-1)}\right\} & \text{if } r < q. \end{cases} \tag{4.50}$$

(i): It follows from (4.49) that

$$(\forall n \in \mathbb{N}) \qquad \psi_0^\natural\left(\frac{\|\ell(\cdot,\cdot,0)\|_\infty + 1}{\lambda_n}\right) \leqslant \left(\frac{\|\ell(\cdot,\cdot,0)\|_\infty + 1}{\eta\beta\lambda_n}\right)^{1/r} = \rho_n. \tag{4.51}$$

Now fix $\tau \in \mathbb{R}_{++}$ and assume that $\sup_{n\in\mathbb{N}} \mathrm{Lip}(\ell;\varsigma\rho_n) > 0$. Since $\mathrm{Lip}(\ell;\varsigma\rho_n)^m/(\lambda_n^{m/r}n^{1/q}) \to 0$ and $m \geqslant 2$, we have $\mathrm{Lip}(\ell;\varsigma\rho_n)/(\lambda_n^{m/r}n^{1/q}) \to 0$. Moreover, since $m/r \geqslant 1$, we have $\mathrm{Lip}(\ell;\varsigma\rho_n)/(\lambda_n n^{1/q}) \to 0$ and, therefore, since $\rho_n \to +\infty$, there exists $\bar{n} \in \mathbb{N} \smallsetminus \{0\}$ such that, for every integer $n \geqslant \bar{n}$, $\tau\mathrm{Lip}(\ell;\varsigma\rho_n)/(\lambda_n n^{1/q}) \leqslant \eta\beta\rho_n^{r-1}$. Suppose that $q > r$ and take an integer $n \geqslant \bar{n}$. Evaluating the maximum in (4.50), we obtain

$$(\widehat{\psi}_{\rho_n})^\natural\left(\frac{\tau\mathrm{Lip}(\ell;\varsigma\rho_n)}{\lambda_n n^{1/q}}\right) \leqslant 2^q\left(\frac{\tau\rho_n^{q-r}}{\eta\beta}\frac{\mathrm{Lip}(\ell;\varsigma\rho_n)}{\lambda_n n^{1/q}}\right)^{1/(q-1)}. \tag{4.52}$$

Therefore, substituting the expression of ρ_n yields

$$\mathrm{Lip}(\ell;\varsigma\rho_n)(\widehat{\psi}_{\rho_n})^\natural\left(\frac{\tau\mathrm{Lip}(\ell;\varsigma\rho_n)}{\lambda_n n^{1/q}}\right) \leqslant 2^q\tau^{\frac{1}{q-1}}\left(\frac{(\|\ell(\cdot,\cdot,0)\|_\infty + 1)^{q/r-1}}{(\eta\beta)^{q/r}}\frac{\mathrm{Lip}(\ell;\varsigma\rho_n)^q}{\lambda_n^{q/r}n^{1/q}}\right)^{\frac{1}{(q-1)}}. \tag{4.53}$$

On the other hand, if $q \leqslant r$, (4.50) yields

$$\mathrm{Lip}(\ell;\varsigma\rho_n)(\widehat{\psi}_{\rho_n})^\natural\left(\frac{\tau\mathrm{Lip}(\ell;\varsigma\rho_n)}{\lambda_n n^{1/q}}\right) \leqslant \left(\frac{\tau}{\eta\beta}\right)^{1/(r-1)}\left(\frac{\mathrm{Lip}(\ell;\varsigma\rho_n)^r}{\lambda_n n^{1/q}}\right)^{1/(r-1)}. \tag{4.54}$$

Thus, altogether (4.53) and (4.54) imply that there exists $\gamma \in \mathbb{R}_{++}$ such that, for every integer $n \geqslant \bar{n}$

$$\mathrm{Lip}(\ell;\varsigma\rho_n)(\widehat{\psi}_{\rho_n})^\natural\left(\frac{\tau\mathrm{Lip}(\ell;\varsigma\rho_n)}{\lambda_n n^{1/q}}\right) \leqslant \gamma\tau^{1/(m-1)}\left(\frac{\mathrm{Lip}(\ell;\varsigma\rho_n)^m}{\lambda_n^{m/r}n^{1/q}}\right)^{1/(m-1)}. \tag{4.55}$$

It therefore follows from (4.15) that the right-hand side of (4.55) converges to zero and hence that (4.6) is fulfilled. Likewise, (4.16) implies (4.7). Altogether the statement follows from Theorem 4.4(i).

(ii): It follows from Remark 3.17(v)(b) that $\ell(\cdot,\cdot,0)$ is bounded and that, for every $\rho \in \mathbb{R}_{++}$, $\mathrm{Lip}(\ell;\rho) \leqslant (p-1)\|b\|_\infty + 3cp\max\{1,\rho^{p-1}\}$. Set $(\forall n \in \mathbb{N})$ $\rho_n = \big((\|\ell(\cdot,\cdot,0)\|_\infty + 1)/(\eta\beta\lambda_n)\big)^{1/r}$. Then $(\exists\,\gamma \in \mathbb{R}_{++})(\forall n \in \mathbb{N})$ $\mathrm{Lip}(\ell;\rho_n) \leqslant \gamma/\lambda_n^{(p-1)/r}$. Thus, the statement follows from (i).

(iii): Fix $\tau \in \mathbb{R}_{++}$ and set $(\forall n \in \mathbb{N})$ $\rho_n = \big((R(0)+1)/(\eta\beta\lambda_n)\big)^{1/r}$. Then (4.49) yields $(\forall n \in \mathbb{N})$ $\psi_0^\natural((R(0)+1)/\lambda_n) \leqslant \rho_n$. Since $m/r \geqslant 1$, $1/(\lambda_n^{m/r}n^{1/q}) \to 0$ implies $1/(\lambda_n n^{1/q}) \to 0$. Moreover, since $\rho_n \to +\infty$, there exists $\bar{n} \in \mathbb{N} \smallsetminus \{0\}$ such that, for every integer $n \geqslant \bar{n}$, $\tau/(\lambda_n n^{1/q}) \leqslant \eta\beta\rho_n^{r-1}$. Suppose that $q > r$ and take an integer $n \geqslant \bar{n}$. Evaluating the maximum in (4.50), we obtain

$$(\widehat{\psi}_{\rho_n})^{\natural}\left(\frac{\tau}{\lambda_n n^{1/q}}\right) \leqslant 2^q \left(\frac{\tau\rho_n^{q-r}}{\eta\beta}\frac{1}{\lambda_n n^{1/q}}\right)^{\frac{1}{q-1}} = 2^q \tau^{\frac{1}{q-1}} \left(\frac{(R(0)+1)^{q/r-1}}{(\eta\beta)^{q/r}}\frac{1}{\lambda_n^{q/r} n^{1/q}}\right)^{\frac{1}{q-1}}. \quad (4.56)$$

On the other hand, if $q \leqslant r$, (4.50) yields

$$(\widehat{\psi}_{\rho_n})^{\natural}\left(\frac{\tau}{\lambda_n n^{1/q}}\right) \leqslant \left(\frac{\tau}{\eta\beta}\frac{1}{\lambda_n n^{1/q}}\right)^{1/(r-1)}, \quad (4.57)$$

Thus (4.17), together with (4.56) and (4.57) imply that (4.10) is fulfilled. Likewise, the assumption $\log n/(\lambda_n^{m/r} n^{1/q}) \to 0$ implies that (4.11) holds. Altogether, the statement follows by Theorem 4.4(iii). □

A Appendix

A.1 Lipschitz continuity of convex functions

Proposition A.1 *Let $\mathcal{B}$ be a real Banach space and let $F\colon \mathcal{B} \to [0,+\infty]$ be proper and convex. Then the following hold:*

(i) *[54, Proposition 1.11] Let $u_0 \in \mathcal{B}$, and suppose that there exist a neighborhood $\mathcal{U}$ of u_0 and $c \in \mathbb{R}_+$ such that, $(\forall\, u \in \mathcal{U})\ |F(u) - F(u_0)| \leqslant c\|u - u_0\|$. Then $\partial F(u_0) \neq \varnothing$ and $\sup \|\partial F(u_0)\| \leqslant c$.*

(ii) *[77, Corollary 2.2.12] Let $u_0 \in \mathcal{B}$, and suppose that, for some $(\rho,\delta) \in \mathbb{R}^2_{++}$, F is bounded on $u_0 + B(\rho+\delta)$. Then F is Lipschitz continuous relative to $u_0 + B(\rho)$ with constant*

$$\frac{2\rho+\delta}{\rho+\delta}\frac{1}{\delta}\sup F(u_0 + B(\rho+\delta)). \quad (A.1)$$

Proposition A.2 *Let $\mathcal{B}$ be a real normed vector space, let $p \in [1,+\infty[$, let $b \in \mathbb{R}_+$, let $c \in \mathbb{R}_{++}$, and let $F\colon \mathcal{B} \to \mathbb{R}_+$ be a convex function such that $F \leqslant c\|\cdot\|^p + b$. Then the following hold:*

(i) *Let $u \in \mathcal{B}$. Then $\partial F(u) \neq \varnothing$ and*

$$\sup \|\partial F(u)\| \leqslant \begin{cases} c & \text{if } p = 1 \\ 3cp\max\{1, \|u\|^{p-1}\} + (p-1)b & \text{if } p > 1. \end{cases} \quad (A.2)$$

(ii) *Let $\rho \in \mathbb{R}_{++}$. Then F is Lipschitz continuous relative to $B(\rho)$ with constant*

$$\begin{cases} c & \text{if } p = 1 \\ 3cp\max\{1, \rho^{p-1}\} + (p-1)b & \text{if } p > 1. \end{cases} \quad (A.3)$$

Proof. (i): Let $(\epsilon,\delta) \in \mathbb{R}^2_{++}$. Since $F \leqslant c\|\cdot\|^p + b$, then, F is bounded on $u + B(\epsilon+\delta)$ and it follows from Proposition A.1(ii) that F is Lipschitz continuous relative to $u + B(\epsilon)$ with constant $(2\epsilon+\delta)(\epsilon+\delta)^{-1}\delta^{-1}\big(c(\|u\|+\epsilon+\delta)^p + b\big)$. Then Proposition A.1(i) entails that $\partial F(u) \neq \varnothing$ and

$$\sup \|\partial F(u)\| \leqslant \frac{2\epsilon+\delta}{\epsilon+\delta}\frac{1}{\delta}\big(c(\|u\|+\epsilon+\delta)^p + b\big). \quad (A.4)$$

Letting $\epsilon \to 0^+$ in (A.4), we get

$$\sup \|\partial F(u)\| \leqslant c\Big(\frac{\|u\|}{\delta}+1\Big)(\|u\|+\delta)^{p-1}+\frac{b}{\delta}. \tag{A.5}$$

If $p = 1$, letting $\delta \to +\infty$ in (A.5) yields $\sup\|\partial F(u)\| \leqslant c$. Now, suppose that $p > 1$ and set $s = \max\{\|u\|, 1\}$. Then, since $\|u\| \leqslant s$, (A.5) implies that

$$\sup \|\partial F(u)\| \leqslant c\Big(\frac{s}{\delta}+1\Big)s^{p-1}\Big(1+\frac{\delta}{s}\Big)^{p-1}+\frac{b}{\delta} \leqslant c\Big(\frac{s}{\delta}+1\Big)s^{p-1}e^{\delta(p-1)/s}+\frac{b}{\delta}, \tag{A.6}$$

where we took into account that $(1+\delta/s)^{s/\delta} \leqslant e$. By choosing $\delta = s/(p-1)$, we get $\sup\|\partial F(u)\| \leqslant 3cps^{p-1} + (p-1)b/s$ and (A.3) follows since $1/s \leqslant 1$.

(ii): Let $(u, v) \in B(\rho)^2$. It follows from (i) that $\partial F(u) \neq \varnothing$ and $\partial F(v) \neq \varnothing$. Let $u^* \in \partial F(u)$ and $v^* \in \partial F(v)$. Then $F(v) - F(u) \geqslant \langle v - u, u^*\rangle$ and $F(u) - F(v) \geqslant \langle u - v, v^*\rangle$. Hence $|F(u) - F(v)| \leqslant \max\{\|u^*\|, \|v^*\|\}\|u - v\|$ and the statement follows by (i). ☐

Proposition A.3 *Let $\mathcal{B}$ be a real Banach space, let $\rho \in \mathbb{R}_{++}$, let $p \in \,]1, +\infty[$, let $b \in \mathbb{R}_+$, let $c \in \mathbb{R}_{++}$, and set $F = c\|\cdot\|^p + b$. Then F is Lipschitz continuous relative to $B(\rho)$ with constant $cp\rho^{p-1}$.*

Proof. Let $(u, v) \in \mathcal{B}^2$ and let $u^* \in J_{\mathcal{B},p}(u)$. Then (2.7) yields $\|u\|^p - \|v\|^p \leqslant p\langle u - v, u^*\rangle \leqslant p\|u^*\|\|v - u\| = p\|u\|^{p-1}\|u - v\|$. Swapping u and v yields

$$\big|\|u\|^p - \|v\|^p\big| \leqslant p\max\big\{\|u\|^{p-1}, \|v\|^{p-1}\big\}\|u - v\|, \tag{A.7}$$

and the claim follows. ☐

A.2 Totally convex functions

Let $\mathcal{F}$ be a reflexive real Banach space and let $G\colon \mathcal{F} \to \,]-\infty, +\infty]$ be a proper convex function. Following (2.8), we denote by $\psi\colon \operatorname{dom} G \times \mathbb{R} \to [0, +\infty]$ the modulus of total convexity of G and, following (2.9), for every $\rho \in \mathbb{R}_{++}$ such that $B(\rho) \cap \operatorname{dom} G \neq \varnothing$, we denote by $\psi_\rho\colon \mathbb{R} \to [0, +\infty]$ the modulus of total convexity of G on $B(\rho)$. G is *totally convex at* $u \in \operatorname{dom} G$ if, for every $t \in \mathbb{R}_{++}$, $\psi(u, t) > 0$. Moreover, G is *totally convex on bounded sets* if, for every $\rho \in \mathbb{R}_{++}$ such that $B(\rho) \cap \operatorname{dom} G \neq \varnothing$, G is totally convex on $B(\rho)$, meaning that $\psi_\rho > 0$ on $\mathbb{R}_{++}$. Total convexity and standard variants of convexity are related as follows:

- Suppose that G is totally convex at every point of $\operatorname{dom} G$. Then G is strictly convex.
- Total convexity is closely related to uniform convexity [70, 76]. Indeed G is uniformly convex on $\mathcal{F}$ if and only if, for every $t \in \mathbb{R}_{++}$, $\inf_{u \in \operatorname{dom} G} \psi(u, t) > 0$ [77, Theorem 3.5.10]. Alternatively, G is uniformly convex on $\mathcal{F}$ if and only if $(\forall t \in \mathbb{R}_{++})\ \inf_{\rho \in \mathbb{R}_{++}} \psi_\rho(t) > 0$.
- In reflexive spaces, total convexity on bounded sets is equivalent to uniform convexity on bounded sets [19, Proposition 4.2]. Yet, some results will require pointwise total convexity, which makes it the pertinent notion in our investigation.

Remark A.4 Let u_0 and u be in $\operatorname{dom} G$. Then (2.8) implies that

$$G(u) - G(u_0) \geqslant G'(u_0; u - u_0) + \psi(u_0, \|u - u_0\|). \tag{A.8}$$

Moreover, if $u^* \in \partial G(u_0)$, $\langle u - u_0, u^*\rangle \leqslant G'(u_0; u - u_0)$ and therefore

$$G(u) - G(u_0) \geqslant \langle u - u_0, u^*\rangle + \psi(u_0, \|u - u_0\|). \tag{A.9}$$

Thus, $\partial G(u_0) \neq \varnothing \Rightarrow \psi(u_0, \|u - u_0\|) < +\infty$.

The following proposition collects some properties of the classes $\mathcal{A}_0$ and $\mathcal{A}_1$ introduced in (2.12) and (2.13) that are used to study the modulus of total convexity.

Proposition A.5 *Let $\phi \in \mathcal{A}_0$. Then the following hold:*

(i) $\operatorname{dom}\phi$ *is an interval containing* 0.

(ii) $\operatorname{dom}\phi^\natural = [0, \sup\phi(\mathbb{R}_+)[$.

(iii) *Suppose that $\widehat{\phi}$ is increasing on $\mathbb{R}_+$. Then $\operatorname{dom}\phi^\natural = \mathbb{R}_+$ and ϕ is strictly increasing on $\operatorname{dom}\phi$.*

(iv) *Suppose that $(t_n)_{n\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ satisfies $\phi(t_n) \to 0$. Then $t_n \to 0$.*

(v) *$\phi^\natural$ is increasing on $\mathbb{R}_+$ and $\lim_{s\to 0^+}\phi^\natural(s) = 0 = \phi^\natural(0)$.*

(vi) *Let $(s,t) \in \mathbb{R}_+ \times \mathbb{R}_{++}$. Then $\phi^\natural(s) < t \Leftrightarrow s < \phi(t^-)$.*

(vii) *Suppose that $\phi \in \mathcal{A}_1$. Then $\operatorname{int}(\operatorname{dom}\phi) \neq \varnothing$, $\widehat{\phi} \in \mathcal{A}_0$, $\widehat{\phi}$ is right-continuous at 0, and $(\widehat{\phi})^\natural \in \mathcal{A}_0$.*

Proof. (i): This follows from (2.12).

(ii): For every $s \in \mathbb{R}_+$, $[\phi \leqslant s] \subset \operatorname{dom}\phi$. Therefore, if $\operatorname{dom}\phi$ is bounded, $\phi^\natural$ is real-valued. Now, suppose that $\operatorname{dom}\phi = \mathbb{R}_+$. Let $s \in \mathbb{R}_+$ with $s < \sup\phi(\mathbb{R}_+)$. Then there exists $t_1 \in \mathbb{R}_+$ such that $s < \phi(t_1)$. Moreover, since ϕ is increasing, $t \in [\phi \leqslant s] \Rightarrow \phi(t) \leqslant s < \phi(t_1) \Rightarrow t \leqslant t_1$. Hence, $\phi^\natural(s) = \sup[\phi \leqslant s] \leqslant t_1 < +\infty$. Therefore $[0, \sup\phi(\mathbb{R}_+)[\subset \operatorname{dom}\phi^\natural$. On the other hand, if $s \in [\sup\phi(\mathbb{R}_+), +\infty[$, then $[\phi \leqslant s] = \operatorname{dom}\phi$ and hence $\phi^\natural(s) = +\infty$.

(iii): For every $t \in [1, +\infty[$, $\phi(t) \geqslant t\phi(1) > 0$. Hence $\sup\phi(\mathbb{R}_+) = +\infty$ and therefore (ii) yields $\operatorname{dom}\phi^\natural = \mathbb{R}_+$. Let $t \in \operatorname{dom}\phi$ and $s \in \operatorname{dom}\phi$ with $t < s$. If $t > 0$, then $0 < \phi(t) = t\widehat{\phi}(t) \leqslant t\widehat{\phi}(s) = (t/s)\phi(s) < \phi(s)$; otherwise, (2.12) yields $\phi(t) = \phi(0) = 0 < \phi(s)$.

(iv): Suppose that there exist $\varepsilon \in \mathbb{R}_{++}$ and a subsequence $(t_{k_n})_{n\in\mathbb{N}}$ such that $(\forall n \in \mathbb{N})\ t_{k_n} \geqslant \varepsilon$. Then $\phi(t_{k_n}) \geqslant \phi(\varepsilon) > 0$ and hence $\phi(t_n) \not\to 0$.

(v): See [77, Lemma 3.3.1(i)].

(vi): Suppose that $t \leqslant \phi^\natural(s)$. Then for every $\delta \in]0, t[$ there exists $t' \in \mathbb{R}_+$ such that $\phi(t') \leqslant s$ and $t - \delta < t'$, hence $\phi(t - \delta) \leqslant \phi(t') \leqslant s$. Therefore $0 < \sup_{\delta\in]0,t[}\phi(t - \delta) = \phi(t^-) \leqslant s$. Conversely, suppose that $t > \phi^\natural(s)$. Let $t' \in \left]\phi^\natural(s), t\right[$. Then (2.11) gives $\phi(t') > s$, and hence $\phi(t^-) > s$.

(vii): By (2.12) and (2.13), $\operatorname{int}(\operatorname{dom}\phi) \neq \varnothing$, $\widehat{\phi} \in \mathcal{A}_0$, and $\widehat{\phi}$ is continuous at 0. Let $s \in \mathbb{R}_{++}$. In view of (v), to prove that $(\widehat{\phi})^\natural \in \mathcal{A}_0$, it remains to show that $(\widehat{\phi})^\natural(s) > 0$. By continuity of $\widehat{\phi}$ at 0, $\{t \in \mathbb{R}_+ \mid \widehat{\phi}(t) \leqslant s\}$ is a neighborhood of 0 and hence $(\widehat{\phi})^\natural(s) = \sup\{t \in \mathbb{R}_+ \mid \widehat{\phi}(t) \leqslant s\} > 0$. ☐

The properties of the modulus of total convexity are summarized below.

Proposition A.6 *Let $\mathcal{F}$ be a reflexive real Banach space, let $G\colon \mathcal{F} \to]-\infty, +\infty]$ be a proper convex function the domain of which is not a singleton, let ψ be the modulus of total convexity of G, and let $u_0 \in \operatorname{dom} G$. Then the following hold:*

(i) *Let* $c \in \left]1,+\infty\right[$ *and let* $t \in \mathbb{R}_+$*. Then* $\psi(u_0, ct) \geqslant c\psi(u_0,t)$*.*

(ii) $\psi(u_0,\cdot)\colon \mathbb{R} \to [0,+\infty]$ *is increasing on* $\mathbb{R}_+$*.*

(iii) *Let* $t \in \mathbb{R}_+$*. Then*

$$\psi(u_0,t) = \inf\big\{G(u) - G(u_0) - G'(u_0; u-u_0) \;\big|\; u \in \operatorname{dom} G,\ \|u-u_0\| \geqslant t\big\}. \tag{A.10}$$

(iv) *Suppose that* G *is totally convex at* u_0*. Then* $\psi(u_0,\cdot) \in \mathcal{A}_0$ *and* $\psi(u_0,\cdot)\widehat{\ } \in \mathcal{A}_0$*.*

(v) $\operatorname{dom}\psi(u_0,\cdot)$ *is an interval containing* 0*; moreover, if* $\partial G(u_0) \neq \varnothing$*, then* $\operatorname{int}\operatorname{dom}\psi(u_0,\cdot) \neq \varnothing$*.*

(vi) *Suppose that* $\partial G(u_0) \neq \varnothing$*. Then* $\lim_{t\to 0^+}\psi(u_0,\cdot)\widehat{\ }(t) = 0$*.*

(vii) *Suppose that* $\partial G(u_0) \neq \varnothing$ *and that* G *is totally convex at* u_0*. Then* $\psi(u_0,\cdot) \in \mathcal{A}_1$*.*

(viii) *Let* $\rho \in \mathbb{R}_{++}$ *and suppose that* G *is totally convex on* $B(\rho)$*. Then* $\psi_\rho \in \mathcal{A}_0$ *and* $\widehat{\psi}_\rho \in \mathcal{A}_0$*. Moreover, if* $B(\rho) \cap \operatorname{dom}\partial G \neq \varnothing$*, then* $\psi_\rho \in \mathcal{A}_1$*.*

(ix) *Suppose that* $u_0 \in \operatorname{Argmin}_{\mathcal{F}} G$ *and that* G *is totally convex at* u_0*. Then* G *is coercive.*

Proof. (i): Suppose that $u \in \operatorname{dom} G$ satisfies $\|u-u_0\| = ct$ and set $v = (1-c^{-1})u_0 + c^{-1}u = u_0 + c^{-1}(u-u_0)$. Then $v \in \operatorname{dom} G$ and $\|v-u_0\| = t$. Therefore, since G is convex and $G'(u_0;\cdot)$ is positively homogeneous [9, Proposition 17.2],

$$\begin{aligned}\psi(u_0,t) &\leqslant G(v) - G(u_0) - G'(u_0; v-u_0)\\ &\leqslant (1-c^{-1})G(u_0) + c^{-1}G(u) - G(u_0) - c^{-1}G'(u_0;u-u_0)\\ &= c^{-1}\big(G(u) - G(u_0) - G'(u_0;u-u_0)\big)\,.\end{aligned}$$

Hence $c\psi(u_0,t) \leqslant \psi(u_0,ct)$.

(ii): Let $(s,t) \in \mathbb{R}^2_{++}$ be such that $t < s$, and set $c = s/t$. Then using (i), we have $\psi(u_0,t) \leqslant c^{-1}\psi(u_0,ct) \leqslant \psi(u_0,s)$.

(iii): Suppose that $u \in \operatorname{dom} G$ satisfies $\|u-u_0\| \geqslant t$ and set $s = \|u-u_0\|$. Then by (ii) we have $\psi(u_0,t) \leqslant \psi(u_0,s) \leqslant G(u) - G(u_0) - G'(u_0;u-u_0)$.

(iv): Since $\psi(u_0,0) = 0$, (ii) yields $\psi(u_0,\cdot) \in \mathcal{A}_0$. Moreover, it follows from (i) that $\psi(u_0,\cdot)\widehat{\ }$ is increasing, hence $\psi(u_0,\cdot)\widehat{\ } \in \mathcal{A}_0$.

(v): The first claim follows from the fact that $\psi(u_0,\cdot)$ is increasing and $\psi(u_0,0) = 0$. Next, since $\operatorname{dom} G$ is not a singleton, there exists $u \in \operatorname{dom} G, u \neq u_0$. Finally, Remark A.4 asserts that $\partial G(u_0) \neq \varnothing \Rightarrow \psi(u_0,\|u-u_0\|) < +\infty$.

(vi): Since (i) asserts that $\psi(u_0,\cdot)\widehat{\ }$ is increasing, $\lim_{t\to 0^+}\psi(u_0,\cdot)\widehat{\ }(t) = \inf_{t\in\mathbb{R}_{++}}\psi(u_0,\cdot)\widehat{\ }(t)$. Suppose that $\inf_{t\in\mathbb{R}_{++}}\psi(u_0,\cdot)\widehat{\ }(t) > 0$. Then there exists $\epsilon \in \mathbb{R}_{++}$ such that, for every $t \in \mathbb{R}_{++}$, $\psi(u_0,t) \geqslant \epsilon t$. Let $u \in \operatorname{dom} G \smallsetminus \{u_0\}$. For every $t \in \left]0,1\right]$, define $u_t = u_0 + tv$, where $v = u - u_0$. Then $\epsilon t\|v\| = \epsilon\|u_t - u_0\| \leqslant \psi(u_0,\|u_t-u_0\|) \leqslant G(u_0+tv) - G(u_0) - G'(u_0;tv)$. Hence, since $G'(u_0;\cdot)$ is positively homogeneous, $\epsilon\|v\| + G'(u_0;v) \leqslant (G(u_0+tv) - G(u_0))/t$. Letting $t \to 0^+$ yields $\epsilon\|v\| + G'(u_0;v) \leqslant G'(u_0;v)$, which contradicts the facts that $G'(u_0;v) \in \mathbb{R}$ and $\epsilon\|v\| > 0$.

(vii)–(viii): The claims follow from (iv) and (vi).

(ix): Since $0 \in \partial G(u_0)$, (A.9) yields $(\forall u \in \operatorname{dom} G)\ \psi(u_0,\|u-u_0\|) \leqslant G(u) - G(u_0)$. On the other hand, since G is also totally convex at u_0, (iv)-(v) imply that there exists $s \in \mathbb{R}_{++}$ such that

$0 < \psi(u_0, s) < +\infty$ and $(\forall t \in [s, +\infty[)\ \psi(u_0, t) \geqslant t\psi(u_0, s)/s$. Therefore, for every $u \in \operatorname{dom} G$ such that $\|u - u_0\| \geqslant s$, we have $G(u) \geqslant G(u_0) + \|u - u_0\|\psi(u_0, s)/s$, which implies that G is coercive. ☐

Remark A.7 Statements (i), (ii), (iii), and (v) are proved in [20, Proposition 2.1] with the additional assumption that $\operatorname{int} \operatorname{dom} G \neq \varnothing$, and in [18, Proposition 1.2.2] with the additional assumption that u_0 is in the algebraic interior of $\operatorname{dom} G$.

Example A.8 Let $\mathcal{F}$ be a uniformly convex real Banach space and let $\phi \in \mathcal{A}_0$ be real-valued, strictly increasing, continuous, and such that $\lim_{t \to +\infty} \phi(t) = +\infty$. Define $(\forall t \in \mathbb{R})\ \varphi(t) = \int_0^{|t|} \phi(s) \mathrm{d}s$. Then [76, Theorem 4.1(ii)] and [19, Proposition 4.2] imply that $G = \varphi \circ \|\cdot\|$ is totally convex on bounded sets (see also [70, Theorem 6]).

We now provide an example of computation of the modulus of total convexity on balls.

Proposition A.9 *Let $q \in [2, +\infty[$ and let $\mathcal{F}$ be a uniformly convex real Banach space with modulus of convexity of power type q. Let $r \in]1, +\infty[$ and for every $\rho \in \mathbb{R}_+$, denote by ψ_ρ the modulus of total convexity of $\|\cdot\|^r$ on the ball $B(\rho)$. Then there exists $\beta \in \mathbb{R}_{++}$ such that*

$$(\forall \rho \in \mathbb{R}_+)(\forall t \in \mathbb{R}_+) \qquad \psi_\rho(t) \geqslant \begin{cases} \beta t^r & \text{if } r \geqslant q \\ \dfrac{\beta t^q}{(\rho + t)^{q-r}} & \text{if } r < q. \end{cases} \tag{A.11}$$

Hence $\|\cdot\|^r$ is totally convex on bounded sets and, if $r \geqslant q$, it is uniformly convex. Moreover, for every $\rho \in \mathbb{R}_+$ and every $s \in \mathbb{R}_+$,

$$(\widehat{\psi}_\rho)^\natural(s) \leqslant \begin{cases} \left(\dfrac{s}{\beta}\right)^{1/(r-1)} & \text{if } r \geqslant q \\ 2^q \rho \max\left\{ \left(\dfrac{s}{\beta \rho^{r-1}}\right)^{1/(q-1)}, \left(\dfrac{s}{\beta \rho^{r-1}}\right)^{1/(r-1)} \right\} & \text{if } r < q. \end{cases} \tag{A.12}$$

Proof. Let $(u, v) \in \mathcal{F}^2$. We derive from [74, Theorem 1] that

$$(\forall u^* \in J_{\mathcal{F},r}(u)) \quad \|u + v\|^r - \|u\|^r \geqslant r\langle v, u^* \rangle + \vartheta_r(u, v), \tag{A.13}$$

where

$$\vartheta_r(u, v) = rK_r \int_0^1 \frac{\max\{\|u + tv\|, \|u\|\}^r}{t} \delta_{\mathcal{F}}\left(\frac{t\|v\|}{2 \max\{\|u + tv\|, \|u\|\}} \right) \mathrm{d}t$$

and $K_r \in \mathbb{R}_{++}$ is the constant defined according to [74, Lemma 3, Equation (2.13)]. Since $\delta_{\mathcal{F}}(\varepsilon) \geqslant c\varepsilon^q$ for some $c \in \mathbb{R}_{++}$, then

$$\vartheta_r(u, v) \geqslant \frac{rK_r c}{2^q} \|v\|^q \int_0^1 \max\{\|u + tv\|, \|u\|\}^{r-q} t^{q-1} \mathrm{d}t. \tag{A.14}$$

Suppose first that $r \geqslant q$. Since, $\forall\, t \in [0, 1]$, $\max\{\|u + tv\|, \|u\|\} \geqslant t\|v\|/2$,

$$\vartheta_r(u, v) \geqslant \frac{rK_r c}{2^q} \|v\|^q \int_0^1 \frac{t^{r-q}}{2^{r-q}} \|v\|^{r-q} t^{q-1} \mathrm{d}t = \frac{rK_r c}{2^r} \|v\|^r \int_0^1 t^{r-1} \mathrm{d}t = \frac{K_r c}{2^r} \|v\|^r. \tag{A.15}$$

Now, suppose that $r < q$. Then since, for every $t \in [0,1]$, $\max\{\|u+tv\|, \|u\|\} \leqslant \|u\| + \|v\|$,

$$\vartheta_r(u,v) \geqslant \frac{rK_rc}{2^q}\|v\|^q \int_0^1 \frac{1}{\max\{\|u+tv\|, \|v\|\}^{q-r}} t^{q-1}\mathrm{d}t \geqslant \frac{rK_rc}{q2^q}\frac{\|v\|^q}{(\|u\|+\|v\|)^{q-r}}. \tag{A.16}$$

Let ψ be the modulus of total convexity of $\|\cdot\|^r$. Then it follows from (A.15) and (A.16) that

$$(\forall\, u \in \mathcal{F})(\forall\, t \in \mathbb{R}_+) \quad \psi(u,t) \geqslant \begin{cases} \dfrac{K_rc}{2^r}t^r & \text{if } q \leqslant r \\ \dfrac{r}{q}\dfrac{K_rc}{2^q}\dfrac{t^q}{(\|u\|+t)^{q-r}} & \text{if } q > r. \end{cases} \tag{A.17}$$

Let $\rho \in \mathbb{R}_{++}$ and set $\beta = (r/\max\{q,r\})K_rc/2^{\max\{q,r\}}$. Then we obtain (A.11) by taking the infimum over $u \in B(\rho)$ in (A.17). Thus, if $r \geqslant q$, the modulus of total convexity is independent from ρ, and hence $\|\cdot\|^r$ is uniformly convex on $\mathcal{F}$. On the other hand, if $r < q$, we deduce that $\|\cdot\|^r$ is totally convex on bounded sets. Hence,

$$(\forall\, t \in \mathbb{R}_+) \quad \widehat{\psi}_\rho(t) \geqslant \begin{cases} \beta t^{r-1} & \text{if } r \geqslant q \\ \dfrac{\beta t^{q-1}}{(\rho+t)^{q-r}} & \text{if } r < q. \end{cases} \tag{A.18}$$

A simple calculation shows that, if $r < q$,

$$(\forall\, t \in \mathbb{R}_+) \quad \widehat{\psi}_\rho(t) \geqslant \nu_\rho(t), \quad \text{where } \nu_\rho(t) = \frac{\beta\rho^{r-1}}{2^q}\min\left\{(t/\rho)^{q-1}, (t/\rho)^{r-1}\right\}. \tag{A.19}$$

The function ν_ρ is strictly increasing and continuous on $\mathbb{R}_+$, thus $\nu_\rho^\natural = \nu_\rho^{-1}$. Since for arbitrary functions $\psi_1 \colon \mathbb{R}_+ \to \mathbb{R}_+$ and $\psi_2 \colon \mathbb{R}_+ \to \mathbb{R}_+$ we have $\psi_1 \geqslant \psi_2 \Rightarrow \psi_1^\natural \leqslant \psi_2^\natural$, we obtain (A.12). □

Remark A.10

(i) An inspection of the proof of Proposition A.9 reveals that the constant β is explicitly available in terms of r and of a constant depending on the space $\mathcal{F}$. In particular, it follows from [74, Equation (2.13)] that, when $r \in \left]1,2\right]$,

$$K_r \geqslant 4(2+\sqrt{3})\min\{r(r-1)/2, (r-1)\log(3/2), 1-(2/3)^{r-1}\} > 14(1-(2/3)^{r-1}), \tag{A.20}$$

and when $r \in \left]2,+\infty\right[$

$$K_r \geqslant 4(2+\sqrt{3})\min\{1, (r-1)(2-\sqrt{3}), 1-(2/3)^{\frac{r}{2}}\}14(1-(2/3)^{\frac{r-1}{2}}). \tag{A.21}$$

As an example, for the case $\mathcal{F} = l^r(\mathbb{K})$ and $\|\cdot\|_r^r$, with $r \in \left]1,2\right]$, since $\mathcal{F}$ has modulus of convexity of power type 2 with $c = (r-1)/8$ [46], we have $\beta \geqslant (7/32)r(r-1)(1-(2/3)^{r-1})$.

(ii) In [73, Theorem 1] and [11, Lemma 2 p. 310] the case $r = q$ is considered. It is proved that $\|\cdot\|_{\mathcal{F}}^r$ is uniformly convex and that its modulus of uniform convexity, say ν, satisfies $\nu(t) \geqslant \beta t^r$, for every $t \in \mathbb{R}_{++}$.

A.3 Tikhonov-like regularization

In this section we work with the following scenario.

Assumption A.11 $\mathcal{F}$ is a reflexive real Banach space, $F\colon \mathcal{F} \to \,]-\infty, +\infty]$ is bounded from below, $G\colon \mathcal{F} \to [0, +\infty]$, $\operatorname{dom} G$ is not a singleton, and $\operatorname{dom} F \cap \operatorname{dom} G \neq \varnothing$. The function $\varepsilon\colon \mathbb{R}_{++} \to [0,1]$ satisfies $\lim_{\lambda\to 0^+} \varepsilon(\lambda) = 0$ and, for every $\lambda \in \mathbb{R}_{++}$, $u_\lambda \in \operatorname{Argmin}_{\mathcal{F}}^{\varepsilon(\lambda)}(F + \lambda G)$.

We study the behavior of the regularized problem

$$\underset{u\in\mathcal{F}}{\text{minimize}} \;\; F(u) + \lambda G(u) \tag{A.22}$$

as $\lambda \to 0^+$ in connection with the limiting problem

$$\underset{u\in\mathcal{F}}{\text{minimize}} \;\; F(u). \tag{A.23}$$

We present results similar to those of [4] under weaker assumptions and with approximate solutions of (A.22), as opposed to exact ones. In particular, Proposition A.12 does not require the family $(u_\lambda)_{\lambda\in\mathbb{R}_{++}}$ to be bounded or F to have minimizers. Indeed, although these are common requirements in the inverse problems literature, where the convergence of the minimizers $(u_\lambda)_{\lambda\in\mathbb{R}_{++}}$ is relevant, from the statistical learning point of view this assumption is not always appropriate. In that context, as discussed in the introduction, it is primarily the convergence of the values $(F(u_\lambda))_{\lambda\in\mathbb{R}_{++}}$ to $\inf F(\mathcal{F})$ which is of interest. On the other hand, when $(u_\lambda)_{\lambda\in\mathbb{R}_{++}}$ is bounded and when additional convexity properties are imposed on G, we provide bounds and strong convergence results.

Proposition A.12 *Suppose that Assumption A.11 holds. Then the following hold:*

(i) $\lim_{\lambda\to 0^+} \inf(F + \lambda G)(\mathcal{F}) = \inf F(\operatorname{dom} G)$.

(ii) $\lim_{\lambda\to 0^+} F(u_\lambda) = \inf F(\operatorname{dom} G)$.

(iii) $\lim_{\lambda\to 0^+} \lambda G(u_\lambda) = 0$.

Proof. (i): Since $\operatorname{dom} F \cap \operatorname{dom} G \neq \varnothing$, $\inf(F + \lambda G)(\mathcal{F}) < +\infty$. Let $u \in \operatorname{dom} G$. Then

$$\begin{aligned}(\forall \lambda \in \mathbb{R}_{++}) \quad \inf F(\operatorname{dom} G) &\leqslant F(u_\lambda) \leqslant F(u_\lambda) + \lambda G(u_\lambda) \leqslant \inf(F + \lambda G)(\mathcal{F}) + \varepsilon(\lambda) \\ &\leqslant F(u) + \lambda G(u) + \varepsilon(\lambda).\end{aligned} \tag{A.24}$$

Hence, $\inf F(\operatorname{dom} G) \leqslant \underline{\lim}_{\lambda\to 0^+} \big(\inf(F+\lambda G)(\mathcal{F})+\varepsilon(\lambda)\big) \leqslant \overline{\lim}_{\lambda\to 0^+} \big(\inf(F+\lambda G)(\mathcal{F})+\varepsilon(\lambda)\big) \leqslant F(u)$. Therefore, $\lim_{\lambda\to 0^+} \big(\inf(F + \lambda G)(\mathcal{F}) + \varepsilon(\lambda)\big) = \inf F(\operatorname{dom} G)$, and the statement follows.

(ii): This follows from (i) and (A.24).

(iii): By (i) and (A.24) we have $\lim_{\lambda\to 0^+} F(u_\lambda) + \lambda G(u_\lambda) = \inf F(\operatorname{dom} G)$ which, together with (ii), yields the statement. □

Remark A.13 Assume that $\inf F(\mathcal{F}) = \inf F(\operatorname{dom} G)$. Then Proposition A.12 yields $\lim_{\lambda\to 0^+} F(u_\lambda) = \inf F(\mathcal{F})$ and $\lim_{\lambda\to 0^+} \inf(F + \lambda G)(\mathcal{F}) = \inf F(\mathcal{F})$. In particular the condition $\inf F(\mathcal{F}) = \inf F(\operatorname{dom} G)$ is satisfied in each of the following cases:

(i) The lower semicontinuous envelopes of $F + \iota_{\operatorname{dom} G}$ and F coincide [4, Theorem 2.6].

(ii) $\overline{\operatorname{dom} G} \supset \operatorname{dom} F$ and F is upper semicontinuous [9, Proposition 11.1(i)].

(iii) $\operatorname{Argmin}_{\mathcal{F}} F \cap \operatorname{dom} G \neq \varnothing$.

Proposition A.14 *Suppose that Assumption A.11 holds and set $S = \operatorname{Argmin}_{\operatorname{dom} G} F$. Suppose that F and G are weakly lower semicontinuous, that G is coercive, and that $\varepsilon(\lambda)/\lambda \to 0$ as $\lambda \to 0^+$. Then*

$$S \neq \varnothing \quad \Leftrightarrow \quad (\exists\, t \in \mathbb{R})(\forall \lambda \in \mathbb{R}_{++}) \quad G(u_\lambda) \leqslant t. \tag{A.25}$$

Now suppose that $S \neq \varnothing$. Then the following hold:

(i) *$(u_\lambda)_{\lambda\in\mathbb{R}_{++}}$ is bounded and there exists a vanishing sequence $(\lambda_n)_{n\in\mathbb{N}}$ in $\mathbb{R}_{++}$ such that $(u_{\lambda_n})_{n\in\mathbb{N}}$ converges weakly.*

(ii) *Suppose that $u^\dagger \in \mathcal{F}$, that $(\lambda_n)_{n\in\mathbb{N}}$ is a vanishing sequence in $\mathbb{R}_{++}$, and that $u_{\lambda_n} \rightharpoonup u^\dagger$. Then $u^\dagger \in \operatorname{Argmin}_S G$.*

(iii) $\lim_{\lambda\to 0^+} G(u_\lambda) = \inf G(S)$.

(iv) $\lim_{\lambda\to 0^+} \big(F(u_\lambda) - \inf F(\operatorname{dom} G)\big)/\lambda = 0$.

(v) *Suppose that G is strictly quasiconvex [9, Definition 10.25]. Then there exists $u^\dagger \in \mathcal{F}$ such that $\operatorname{Argmin}_S G = \{u^\dagger\}$ and $u_\lambda \rightharpoonup u^\dagger$ as $\lambda \to 0^+$.*

(vi) *Suppose that G is totally convex on bounded sets. Then $u_\lambda \to u^\dagger = \operatorname{argmin}_S G$ as $\lambda \to 0^+$.*

Proof. Assume that $S \neq \varnothing$ and let $u \in S$. For every $\lambda \in \mathbb{R}_{++}$, $F(u_\lambda) + \lambda G(u_\lambda) \leqslant F(u) + \lambda G(u) + \varepsilon(\lambda)$, so that $u_\lambda \in \operatorname{dom} G$ and

$$G(u_\lambda) \leqslant \frac{F(u) - F(u_\lambda)}{\lambda} + \frac{\varepsilon(\lambda)}{\lambda} + G(u) \leqslant G(u) + \frac{\varepsilon(\lambda)}{\lambda}. \tag{A.26}$$

Thus, since $(\varepsilon(\lambda)/\lambda)_{\lambda\in\mathbb{R}_{++}}$ is bounded, so is $(G(u_\lambda))_{\lambda\in\mathbb{R}_{++}}$. Hence $(u_\lambda)_{\lambda\in\mathbb{R}_{++}}$ is in some sublevel set of G. Conversely, suppose that there exists $t \in \mathbb{R}_{++}$ such that $\sup_{\lambda\in\mathbb{R}_{++}} G(u_\lambda) \leqslant t$. It follows from the coercivity of G that $(u_\lambda)_{\lambda\in\mathbb{R}_{++}}$ is bounded. Therefore, since $\mathcal{F}$ is reflexive, there exist $u^\dagger \in \mathcal{F}$ and a sequence $(\lambda_n)_{n\in\mathbb{N}}$ in $\mathbb{R}_{++}$ such that $\lambda_n \to 0$ and $u_{\lambda_n} \rightharpoonup u^\dagger$. In turn, we derive from the weak lower semicontinuity of F and Proposition A.12(ii) that

$$F(u^\dagger) \leqslant \underline{\lim}\, F(u_{\lambda_n}) = \lim F(u_{\lambda_n}) = \inf F(\operatorname{dom} G). \tag{A.27}$$

Moreover, since G is weakly lower semicontinuous,

$$G(u^\dagger) \leqslant \underline{\lim}\, G(u_{\lambda_n}) \leqslant \overline{\lim}\, G(u_{\lambda_n}) \leqslant t. \tag{A.28}$$

Hence $u^\dagger \in \operatorname{dom} G$ and it follows from (A.27) that $u^\dagger \in S$.

(i): This follows from the reflexivity of $\mathcal{F}$ and the boundedness of $(u_\lambda)_{\lambda\in\mathbb{R}_{++}}$.

(ii): Arguing as above, we obtain that (A.27) holds. Moreover, for every $u \in S$, it follows from (A.26) that, since G is weakly lower semicontinuous and $\varepsilon(\lambda_n)/\lambda_n \to 0$,

$$G(u^\dagger) \leqslant \underline{\lim}\, G(u_{\lambda_n}) \leqslant \overline{\lim}\, G(u_{\lambda_n}) \leqslant G(u) < +\infty. \tag{A.29}$$

Inequalities (A.27) and (A.29) imply that $u^\dagger \in S$ and that (ii) holds.

(iii): It follows from (A.29) and (ii) that $G(u_{\lambda_n}) \to \inf G(S)$.

(iv): Let $\lambda \in \mathbb{R}_{++}$. Since u_λ is an $\varepsilon(\lambda)$-minimizer of $F + \lambda G$, for every $u \in \operatorname{dom} G$, we have

$$\frac{F(u_\lambda) - \inf F(\operatorname{dom} G)}{\lambda} + G(u_\lambda) \leqslant \frac{F(u) - \inf F(\operatorname{dom} G)}{\lambda} + G(u) + \frac{\varepsilon(\lambda)}{\lambda}. \tag{A.30}$$

In particular, taking $u = u^\dagger$ in (A.30) yields

$$\frac{F(u_\lambda) - \inf F(\operatorname{dom} G)}{\lambda} + G(u_\lambda) \leqslant G(u^\dagger) + \frac{\varepsilon(\lambda)}{\lambda}. \tag{A.31}$$

Since $\varepsilon(\lambda)/\lambda \to 0$, passing to the limit superior in (A.31) as $\lambda \to 0^+$, and using (ii) and (iii), we get

$$\varlimsup_{\lambda\to 0^+} \frac{F(u_\lambda) - \inf F(\operatorname{dom} G)}{\lambda} + G(u^\dagger) \leqslant G(u^\dagger), \tag{A.32}$$

which implies (iv), since $F(u_\lambda) - \inf F(\operatorname{dom} G) \geqslant 0$.

(v): It follows from (i) and (ii) that $\operatorname{Argmin}_S G \neq \varnothing$. Since S is convex and G is strictly quasiconvex, $\operatorname{Argmin}_S G$ reduces to a singleton $\{u^\dagger\}$ and (ii) yields $u_\lambda \rightharpoonup u^\dagger$ as $\lambda \to 0^+$.

(vi): Since $(u_\lambda)_{\lambda\in\mathbb{R}_{++}}$ is bounded, it follows from [77, Proposition 3.6.5] (see also [19]) that there exists $\phi \in \mathcal{A}_0$ such that

$$(\forall\lambda \in \mathbb{R}_{++}) \quad \phi\Big(\frac{\|u_\lambda - u^\dagger\|}{2}\Big) \leqslant \frac{G(u^\dagger) + G(u_\lambda)}{2} - G\Big(\frac{u_\lambda + u^\dagger}{2}\Big). \tag{A.33}$$

Hence, arguing as in [26, Proof of Proposition 3.1(vi)] and using (v) and the weak lower semicontinuity of G, we obtain $u_\lambda \to u^\dagger$ as $\lambda \to 0^+$. □

Remark A.15 If $\operatorname{Argmin}_{\mathcal{F}} F \cap \operatorname{dom} G \neq \varnothing$, then $S = \operatorname{Argmin}_{\operatorname{dom} G} F = \operatorname{Argmin}_{\mathcal{F}} F \cap \operatorname{dom} G$ and $\operatorname{Argmin}_S G = \operatorname{Argmin}_{\operatorname{Argmin}_{\mathcal{F}} F} G$ (see [4, Theorem 2.6] for related results).

The following proposition provides an estimate of the growth of the function $\lambda \mapsto \|u_\lambda\|$ as $\lambda \to 0^+$ when the condition $\operatorname{Argmin}_{\operatorname{dom} G} F \neq \varnothing$ is possibly not satisfied.

Proposition A.16 *Suppose that Assumption A.11 holds, that G is convex with modulus of total convexity ψ, and that there exists $u \in \mathcal{F}$ such that* $\operatorname{Argmin}_{\mathcal{F}} G \cap \operatorname{dom} F = \{u\}$. *Then*

$$(\forall\lambda \in \mathbb{R}_{++}) \quad \|u_\lambda - u\| \leqslant \psi(u,\cdot)^\natural\Big(\frac{F(u) - \inf F(\operatorname{dom} G) + \varepsilon(\lambda)}{\lambda}\Big). \tag{A.34}$$

Proof. Let $\lambda \in \mathbb{R}_{++}$. Since $F(u_\lambda) + \lambda G(u_\lambda) \leqslant F(u) + \lambda G(u) + \varepsilon(\lambda)$, we have

$$G(u_\lambda) - G(u) \leqslant \frac{F(u) - F(u_\lambda) + \varepsilon(\lambda)}{\lambda} \leqslant \frac{F(u) - \inf F(\operatorname{dom} G) + \varepsilon(\lambda)}{\lambda}. \tag{A.35}$$

Hence, recalling (A.9) and noting that $u \in \operatorname{Argmin}_{\mathcal{F}} G \Leftrightarrow 0 \in \partial G(u)$, we obtain $\psi(u, \|u_\lambda - u\|) \leqslant (F(u) - \inf F(\operatorname{dom} G) + \varepsilon(\lambda))/\lambda$ and the claim follows. □

A.4 Concentration inequalities in Banach spaces

This section provides the Banach space valued versions of the classical Hoeffding inequality. The proof is similar to those of [62, Theorem 6.14 and Corollary 6.15], which deal with the Hilbert space case (see also [75]). A closely related result is [13, Corollary 2.2].

Theorem A.17 (Hoeffding's inequality) *Let $(\Omega, \mathfrak{A}, \mathsf{P})$ be a probability space and let $\mathcal{B}$ be a separable real Banach space of Rademacher type $q \in \,]1,2]$ with Rademacher constant T_q. Let $(\beta,\sigma) \in \mathbb{R}^2_{++}$, let $n \in \mathbb{N} \smallsetminus \{0\}$, let $(U_i)_{1\leqslant i\leqslant n}$ be a family of independent random variables from Ω to $\mathcal{B}$ satisfying $\max_{1\leqslant i\leqslant n}\|U_i\| \leqslant \beta$ P-a.s., and let $\tau \in \mathbb{R}_{++}$. Then the following hold:*

$$\mathsf{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n (U_i - \mathsf{E}_\mathsf{P} U_i)\right\| \geqslant \frac{4\beta T_q}{n^{1-1/q}} + 2\beta\sqrt{\frac{2\tau}{n}} + \frac{4\tau\beta}{3n}\right] \leqslant e^{-\tau}. \tag{A.36}$$

Proof. For every $i \in \{1,\ldots,n\}$, set $V_i = U_i - \mathsf{E}_\mathsf{P} U_i$, so that $\mathsf{E}_\mathsf{P} V_i = 0$, $\|V_i\| \leqslant 2\beta$ P-a.s., and $\mathsf{E}_\mathsf{P}\|V_i\|^q \leqslant (2\beta)^q$. Set $\sigma = 2\beta$. It follows from Jensen's inequality and [45, Proposition 9.11] that

$$\left(\mathsf{E}_\mathsf{P}\left\|\sum_{i=1}^n V_i\right\|\right)^q \leqslant \mathsf{E}_\mathsf{P}\left\|\sum_{i=1}^n V_i\right\|^q \leqslant (2T_q)^q \sum_{i=1}^n \mathsf{E}_\mathsf{P}\|V_i\|^q \leqslant (2T_q)^q n\sigma^q. \tag{A.37}$$

Hence $\mathsf{E}_\mathsf{P}\|\sum_{i=1}^n V_i\| \leqslant 2T_q\sigma n^{1/q}$. Now let $t \in \mathbb{R}_+$. Then

$$\sum_{i=1}^n \mathsf{E}_\mathsf{P}\big(e^{t\|V_i\|} - 1 - t\|V_i\|\big) = \sum_{i=1}^n\sum_{m=2}^{+\infty}\frac{t^m}{m!}\mathsf{E}_\mathsf{P}\|V_i\|^{m-q}\|V_i\|^q \leqslant n\big(e^{2t\beta} - 1 - 2t\beta\big) \tag{A.38}$$

and, using [62, Theorem 6.13] (see also [75, Theorem 3.3.1]), we obtain that, for every $\varepsilon \in \mathbb{R}_{++}$,

$$\mathsf{P}\left[\left\|\sum_{i=1}^n V_i\right\| \geqslant n\varepsilon\right] \leqslant \exp\Big(-t\varepsilon n + 2t\sigma T_q n^{1/q} + n\big(e^{2t\beta} - 1 - 2t\beta\big)\Big). \tag{A.39}$$

For every $\varepsilon \in \mathbb{R}_{++}$ such that $\varepsilon n - 2T_q\sigma n^{1/q} \geqslant 0$, the right-hand side of (A.39) reaches its minimum at

$$\bar{t} = \frac{1}{2\beta}\log(1+\alpha), \quad \text{where} \quad \alpha = \big(\varepsilon n - 2T_q\sigma n^{1/q}\big)/(2n\beta). \tag{A.40}$$

Moreover, as in [62, Theorem 6.14], one gets

$$-\bar{t}\varepsilon n + \bar{t}(b_q n)^{1/q}\sigma + n\big(e^{\bar{t}\beta} - 1 - \bar{t}\beta\big) \leqslant -\frac{3n}{2}\frac{\alpha^2}{\alpha+3}. \tag{A.41}$$

Now set

$$\gamma = \frac{\tau}{3n} \quad \text{and} \quad \varepsilon = \frac{2\tau\beta}{3n}\big(\sqrt{6/\gamma+1}+1\big) + \frac{2T_q\sigma}{n^{1-1/q}}. \tag{A.42}$$

Then $\varepsilon n - 2T_q\sigma n^{1/q} > 0$ and (A.40) yield

$$\alpha = \frac{3\gamma n}{2\tau\beta}\Big(\varepsilon - \frac{2T_q\sigma}{n^{1-1/q}}\Big) = \gamma + \sqrt{\gamma^2 + 6\gamma}, \tag{A.43}$$

so that $\alpha^2 = 2\gamma(\alpha+3) = 2\tau(\alpha+3)/(3n)$. Thus, (A.39) and (A.41) yield $P\big[\big\|\sum_{i=1}^n V_i\big\|/n \geqslant \varepsilon\big] \leqslant e^{-\tau}$. From (A.42), substituting the expression of γ into that of ε, we obtain

$$\varepsilon = \sqrt{\frac{8\tau\beta^2}{n} + \frac{4\beta^2\tau^2}{9n^2}} + \frac{2\tau\beta}{3n} + \frac{2T_q\sigma}{n^{1-1/q}} \leqslant \frac{4\tau\beta}{3n} + 2\beta\sqrt{\frac{2\tau}{n}} + \frac{2T_q\sigma}{n^{1-1/q}}, \tag{A.44}$$

and the statement follows. ☐

References

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces,* 2nd ed. Elsevier, Amsterdam 2003.

[2] A. Antoniadis, D. Leporini, and J.-C. Pesquet, Wavelet thresholding for some classes of non-Gaussian noise, *Statist. Neerlandica*, vol. 56, pp. 434–453, 2002.

[3] A. Argyriou, C. A. Micchelli, and M. Pontil, When is there a representer theorem? Vector versus matrix regularization, *J. Mach. Learn. Res.,* vol. 10, pp. 2507–2529, 2009.

[4] H. Attouch, Viscosity solutions of minimization problems, *SIAM J. Optim.,* vol. 6, pp. 769–805, 1996.

[5] H. Attouch, G. Buttazzo, and G. Michaille, *Variational Analysis in Sobolev and BV Spaces,* 2nd ed. SIAM, Philadelphia, PA 2014.

[6] H. Attouch and R. J.-B. Wets, Quantitative stability of variational systems: I. The epigraphical distance, *Trans. Amer. Math. Soc.,* vol. 328, pp. 695–729, 1991.

[7] P. L. Bartlett and S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results *J. Mach. Learn. Res.,* vol. 3, pp. 463–482, 2002.

[8] P. L. Bartlett, O. Bousquet, and S. Mendelson, Local Rademacher complexities. *Ann. Statist.,* vol. 33, pp. 1497–1537, 2005.

[9] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer, New York 2011.

[10] O. Blasco and J. van Neerven, Spaces of operator-valued functions measurable with respect to the strong operator topology, in: *Vector Measures, Integration and Related Topics,* pp. 65–78. Birkhäuser, Basel 2010.

[11] B. Beauzamy, *Introduction to Banach Spaces and Their Geometry,* 2nd ed. North-Holland, Amsterdam 1985.

[12] G. Beer, *Topologies on Closed and Closed Convex Sets.* Kluwer, Dordrecht 1993.

[13] D. Bosq, *Linear Processes in Function Spaces.* Springer, New York 2000.

[14] N. Bourbaki, *Intégration, Chapitres 1 à 4*, 2nd ed, Hermann, Paris, 1965. English translation: *Integration I,* Springer, New York 2004.

[15] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data.* Springer, Berlin 2011.

[16] J. Burbea and P. Masani, *Banach and Hilbert Spaces of Vector-Valued Functions.* Pitman, Boston 1984.

[17] D. Butnariu, Y. Censor, and S. Reich, Iterative averaging of entropic projections for solving stochastic convex feasibility problems, *Comput. Optim. Appl.,* vol. 8, pp. 21–39, 1997.

[18] D. Butnariu and A. N. Iusem, *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization.* Kluwer, Dordrecht 2000.

[19] D. Butnariu, A. N. Iusem, and C. Zălinescu, On uniform convexity, total convexity and convergence of the proximal point outer Bregman projection algorithm in Banach spaces, *J. Convex Anal.,* vol. 10, pp. 35–61, 2003.

[20] D. Butnariu and E. Resmerita, Bregman distances, totally convex functions, and a method for solving operator equations in Banach spaces, *Abstr. Appl. Anal.,* art. 84919, 39 pp., 2006.

[21] C. Carmeli, E. De Vito, and A. Toigo, Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem, *Anal. Appl. (Singap.)*, vol. 4, pp. 377–408, 2006.

[22] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying, Universal multi-task kernels, *J. Mach. Learn. Res.,* vol. 9, pp. 1615–1646, 2008.

[23] C. Carmeli, E. De Vito, A. Toigo, and V. Umanità, Vector valued reproducing kernel Hilbert spaces and universality, *Anal. Appl. (Singap.)*, vol. 8, pp. 19–61, 2010.

[24] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions.* Lecture Notes in Math. 580. Springer, New York 1977.

[25] I. Cioranescu, *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems.* Kluwer, Dordrecht 1990.

[26] P. L. Combettes, Strong convergence of block-iterative outer approximation methods for convex optimization, *SIAM J. Control Optim.,* vol. 38, pp. 538–565, 2000.

[27] P. L. Combettes and J.-C. Pesquet, Proximal thresholding algorithm for minimization over orthonormal bases, *SIAM J. Optim.,* vol. 18, pp. 1351–1376, 2007.

[28] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint.* Cambridge University Press, Cambridge 2007.

[29] F. Cucker and S. Smale, On the mathematical foundation of learning. *Bull. Amer. Math. Soc.,* vol. 39, pp. 1–49, 2002.

[30] C. De Mol, E. De Vito, and L. Rosasco, Elastic-net regularization in learning theory, *J. Complexity,* vol. 25, pp. 201–230, 2009.

[31] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri, Some properties of regularized kernel methods, *J. Mach. Learn. Res.,* vol. 5, pp. 1363–1390, 2004.

[32] J. Diestel and J. J. Uhl Jr., *Vector Measures.* AMS, Providence, RI 1977.

[33] N. Dinculeanu, *Vector Measures*. Pergamon Press, Oxford 1967.

[34] N. Dinculeanu, *Vector Integration and Stochastic Integration in Banach Spaces*. Wiley-Interscience, New York 2000.

[35] T. Evgeniou, C. A. Micchelli, and M. Pontil, Learning multiple tasks with kernel methods, *J. Mach. Learn. Res.,* vol. 6, pp. 615–637, 2005.

[36] G. E. Fasshauer, F. J. Hickernell, and Q. Ye, Solving support vector machines in reproducing kernel Banach spaces with positive definite functions, *Appl. Comput. Harmon. Anal.,* vol. 38, pp. 115–139, 2015.

[37] F. Ferraty, A. Laksaci, A. Tadj, and P. Vieu, Kernel regression with functional response, *Electron. J. Statist.,* vol. 5, pp. 159–171, 2011.

[38] I. Fonseca and G. Leoni. *Modern Methods in the Calculus of Variations: L^p Spaces.* Springer, New York 2007.

[39] W. Fu, Penalized regressions: the bridge versus the lasso, *J. Comput. Graph. Stat.,* vol. 7, pp. 397–416, 1998.

[40] K. Goebel and S. Reich, *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings.* Marcel Dekker, New York 1984.

[41] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York 2002.

[42] T. Hofmann, B. Schölkopf, and A. J. Smola, Kernel methods in machine learning, *Ann. Statist.,* vol. 36, pp. 1171–1220, 2008.

[43] G. S. Kimeldorf and G. Wahba, A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *Ann. Math. Statist.,* vol. 41, pp. 495–502, 1970.

[44] V. Koltchinskii, Sparsity in penalized empirical risk minimization, *Ann. Inst. Henri Poincaré Probab. Stat.,* vol. 45, pp. 7–57, 2009.

[45] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, New York 1991.

[46] L. Lindenstrauss and L. Tzafriri, *Classical Banach Spaces II.* Springer, Berlin 1979.

[47] R. Lucchetti, *Convexity and Well-Posed Problems.* Springer, New York 2006.

[48] A. Maurer and M. Pontil, Structured sparsity and generalization, *J. Mach. Learn. Res.,* vol. 13, pp. 671–690, 2012.

[49] S. Mendelson, A few notes on statistical learning theory, in: *Lecture Notes in Comput. Sci. 2600,* pp. 1–40. Springer, New York 2003.

[50] C. A. Micchelli and M. Pontil, A function representation for learning in Banach spaces, in: *Lecture Notes in Comput. Sci. 3120,* pp. 255–269. Springer, New York 1994.

[51] C. A. Micchelli and M. Pontil, On Learning vector-valued functions, *Neural Comput.,* vol. 17, pp. 177–204, 2005.

[52] C. A. Micchelli, Y. Xu, and H. Zhang, Universal kernels, *J. Mach. Learn. Res.,* vol. 7, pp. 2651–2667, 2006.

[53] J.-P. Penot, Continuity properties of projection operators, *J. Inequal. Appl.,* vol. 5, pp. 509–521, 2005.

[54] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability,* 2nd ed. Lecture Notes in Math. 1364. Springer, New York 1993.

[55] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis,* 2nd ed. Springer, New York 2005.

[56] M. M. Rao and Z. D. Ren, *Theory of Orlicz Spaces.* Marcel Dekker, New York 1991.

[57] R. T. Rockafellar, *Conjugate Duality and Optimization.* SIAM, Philadelphia, PA 1974.

[58] B. Schölkopf, R. Herbrich, and A. Smola, A generalized representer theorem, in: *Computational Learning Theory,* Lecture Notes in Comput. Sci. 2111, pp. 416–426, 2001.

[59] G. Song and H. Zhang, Reproducing kernel Banach spaces with the ℓ^1 norm II: Error analysis for regularized least square regression, *Neural Comput.,* vol. 23, pp. 2713–2729, 2011.

[60] G. Song, H. Zhang, and F. J. Hickernell, Reproducing kernel Banach spaces with the ℓ^1 norm. *Appl. Comput. Harmon. Anal.,* vol. 34, pp. 96–116, 2013.

[61] I. Steinwart, Two oracle inequalities for regularized boosting classifiers, *Stat. Interface,* vol. 2, pp. 271–284, 2009.

[62] I. Steinwart and A. Christmann, *Support Vector Machines.* Springer, New York 2008.

[63] T. Suzuki and M. Sugiyama, Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness, *Ann. Statist.,* vol. 41, pp. 1381–1405, 2013.

[64] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 58, pp. 267–288, 1996.

[65] A. B. Tsybakov, *Introduction to Nonparametric Estimation.* Springer, New York 2009.

[66] S. Van de Geer, *Empirical Processes in M-Estimation.* Cambridge University Press, Cambridge 2000.

[67] A. W. Van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes.* Springer, New York 1996.

[68] V. N. Vapnik, *Statistical Learning Theory.* Wiley, New York 1998.

[69] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, Accelerated and inexact forward-backward algorithms, *SIAM J. Optim.*, vol. 23, pp. 1607–1633, 2013.

[70] A. A. Vladimirov, Ju. E. Nesterov, and Ju. N. Čekanov, Uniformly convex functionals, *Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet.*, vol. 3, pp. 12–23, 1978.

[71] C. Wang, D.-X. Zhou, Optimal learning rates for least squares regularized regression with unbounded sampling, *J. Complexity,* vol. 27, pp. 55–67, 2011.

[72] Q. V. Wu, Y. Ying, and D.-X. Zhou, Learning rates of least-square regularized regression, *Found. Comput. Math.*, vol. 6, pp. 171–192, 2006.

[73] H. K. Xu, Inequalities in Banach spaces with applications, *Nonlinear Anal.*, vol. 16, pp. 1127–1138, 1991.

[74] Z. B. Xu and G. F. Roach, Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces, *J. Math. Anal. Appl.*, vol. 157, pp. 189–210, 1991.

[75] V. Yurinsky, *Sums and Gaussian Vectors.* Lecture Notes in Math. 1617, Springer, New York 1995.

[76] C. Zălinescu, On uniformly convex functions, *J. Math. Anal. Appl.*, vol. 95, pp. 344–374, 1983.

[77] C. Zălinescu, *Convex Analysis in General Vector Spaces.* World Scientific, River Edge, NJ 2002.

[78] H. Zhang, Y. Xu, and J. Zhang, Reproducing kernel Banach spaces for machine learning, *J. Mach. Learn. Res.,* vol. 10, pp. 2741–2775, 2009.

[79] H. Zhang and J. Zhang, Regularized learning in Banach spaces as an optimization problem: representer theorems, *J. Global Optim.,* vol. 54, pp. 235–250, 2012.

[80] H. Zhang and J. Zhang, Vector-valued reproducing kernel Banach spaces with applications to multi-task learning, *J. Complexity,* vol. 20, pp. 195–215, 2013.

[81] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B Stat. Methodol.,* vol. 67, pp. 301–320, 2005.