A Supplementary material

This is the supplementary material for the paper *Model Consistency for Learning with Mirror-Stratifiable Regularizers*, submitted to AiStats 2019. It contains the detailed proofs of the propositions 2 and 3 of which are, as explained in Section 4, the building blocks of our two main results (Theorems 1 and 2). The supplementary is structured in three sections: Section A.1 gathers key technical lemmas; Section A.2 presents the proof of Proposition 2; Section A.3 presents the one of Proposition 3.

We use the same notations here as introduced at the beginning of Section 4. We also introduce what can viewed as the limit of ($\mathrm{P}_{\lambda,n}$) as $n \to +\infty$:

$$w_\lambda \in \operatorname*{Argmin}_{w\in\mathbb{R}^p} \lambda R(w) + \frac{1}{2}\langle Cw,\, w\rangle - \langle u,\, w\rangle. \qquad (1)$$

For any positive semi-definite matrix A, we also note the seminorm $\|\cdot\|_A = \sqrt{\langle A\cdot,\,\cdot\rangle}$.

A.1 Useful technical lemmas

Here we present a few technical lemmas. The first gives us some control on how $\widehat{C}_n$ converges to C (resp. $\widehat{u}_n$ converges to u) when the amount of data n tends to $+\infty$, and the second provides us with some essential compactness on these sequences. The third provides us an important variational characterization of the set to which belongs η_0. Finally, the last Lemma gives a useful estimate between $\widehat{w}_{\lambda,n}$ and w_λ.

Lemma 1. *If $\lambda_n\sqrt{n/\log\log n} \to +\infty$ and $\mathbb{E}\left[|\mathbf{y}|^4\right] + \mathbb{E}\left[\|\mathbf{x}\|^4\right] < +\infty$, then the following holds almost surely:*

(i) $\max\{\|\widehat{u}_n - u\|, \|\widehat{C}_n - C\|\} = o(\lambda_n)$,

(ii) for n large enough, $\operatorname{Im}\widehat{C}_n = \operatorname{Im} C$,

(iii) $\widehat{C}_n^\dagger \to C^\dagger$ *as* $n \to +\infty$.

Proof. It can be seen (use the Young inequality) that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{xy}\|^2\right] &= \mathbb{E}\left[|\mathbf{y}|^2\|\mathbf{x}\|^2\right]\\ &\leqslant \frac{1}{2}\mathbb{E}\left[|\mathbf{y}|^4\right] + \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}\|^4\right] < +\infty\\ \text{and } \mathbb{E}\left[\|\mathbf{x}\mathbf{x}^\top\|^2\right] &= \mathbb{E}\left[\|\mathbf{x}\|^4\right] < +\infty.\end{aligned}$$

We are then in a position to invoke the law of iterated logarithm (Van der Vaart, 1998, Proposition 2.26) to obtain that, with probability 1,

$$r_n := \max\{\|\widehat{u}_n - u\|, \|\widehat{C}_n - C\|\} = O\left(n^{-1/2}\sqrt{\log\log n}\right).$$

Our assumption that $\lambda_n\sqrt{n/\log\log n} \to +\infty$ then entails item (i).

We now turn to item (ii). Consider $w \in \operatorname{Ker} C$; it verifies by definition $\mathbb{E}_\rho[\mathbf{x}\langle \mathbf{x}, w\rangle] = 0$. By taking the scalar product of this equality with w, we see that $(\forall x \sim \rho),\ \mathbb{P}(\langle x, w\rangle = 0) = 1$. Let $(w_1, ..., w_d)$ be a basis of $\operatorname{Ker} C$, where $d = \dim(\ker C)$. Then we deduce that $(\forall x \sim \rho),\ \mathbb{P}((\forall i \in \{1,\dots,d\})\ \langle x, w_i\rangle = 0) = 1$. In other words, $x \in (\operatorname{Ker} C)^\perp$ a.s., or, equivalently:

$$(\forall x \underset{i.i.d.}{\sim} \rho) \quad \mathbb{P}(x \in \operatorname{Im} C) = 1. \qquad (2)$$

Now, observe that $\operatorname{Im}\widehat{C}_n = \operatorname{Im}(\{x_i\}_{i=1}^n)$, so the following implication holds:

$$[(\forall i \in \{1,\dots,n\})\ x_i \in \operatorname{Im} C] \Rightarrow \operatorname{Im}\widehat{C}_n \subset \operatorname{Im} C. \qquad (3)$$

Since the x_i are drawn i.i.d. from ρ, and are in finite number, we can combine (10) and (11) to obtain that

$$\begin{aligned}\mathbb{P}(\operatorname{Im}\widehat{C}_n \subset \operatorname{Im} C) &\geqslant \mathbb{P}((\forall i \in \{1,\dots,n\})\ x_i \in \operatorname{Im} C)\\ &= \prod_{i=1}^n \mathbb{P}(x_i \in \operatorname{Im} C) = 1.\end{aligned}$$

We deduce then that $\operatorname{Im}\widehat{C}_n \subset \operatorname{Im} C$ a.s., from which we get that $\operatorname{rank}\widehat{C}_k \leqslant \operatorname{rank} C$ a.s. This, together with lower semi-continuity of the rank, yields that with probability 1,

$$\begin{aligned}\operatorname{rank}(C) \leqslant \liminf_{k\to+\infty} \operatorname{rank}(\widehat{C}_k) &\leqslant \limsup_{k\to+\infty} \operatorname{rank}(\widehat{C}_k)\\ &\leqslant \operatorname{rank} C\end{aligned}$$

meaning that $\operatorname{rank}(\widehat{C}_k) \to \operatorname{rank}(C)$ a.s. Because the rank takes only discrete values, this means that $\operatorname{rank}\widehat{C}_k = \operatorname{rank} C$ a.s. for all k large enough. We can then trivially deduce from the inclusion $\operatorname{Im}\widehat{C}_k \subset \operatorname{Im} C$ a.s., that the equality $\operatorname{Im}\widehat{C}_k = \operatorname{Im} C$ holds a.s. for k large enough.

Assertion (iii) follows from (ii) and (Stewart, 1977, Theorem 3.3). □

Lemma 2. *Assume that* (H_M) *holds, and*

- $\operatorname{Im}\widehat{C}_n = \operatorname{Im} C$ *for n large enough,*
- $\sup_{n\in\mathbb{N}} \widehat{C}_n^\dagger < +\infty$.

Then, the sequences $(\widehat{w}_{\lambda_n,n})_{n\in\mathbb{N}}$ and $(w_{\lambda_n})_{n\in\mathbb{N}}$ are bounded.

Proof. Introduce $f_\lambda(w) := R(w) + (1/2\lambda)\|Cw - u\|^2_{C^\dagger}$ and $f_{\lambda,n}(w) := R(w) + (1/2\lambda)\|\widehat{C}_n w - \widehat{u}_n\|^2_{\widehat{C}_n^\dagger}$ which, by definition, verify

$$w_{\lambda_n} \in \operatorname{Argmin} f_{\lambda_n} \text{ and } \widehat{w}_{\lambda_n,n} \in \operatorname{Argmin} f_{\lambda_n,n}.$$

Define $\overline{\lambda} := \sup_n \lambda_n > 0$, and use the optimality of $\widehat{w}_{\lambda_n,n}$ to derive

$$f_{\overline{\lambda},n}(\widehat{w}_{\lambda_n,n}) \leqslant f_{\lambda_n,n}(\widehat{w}_{\lambda_n,n}) \leqslant f_n(w_0)$$

By making use of Lemma 1.(iii) and Lemma 1.(i), we have the bound

$$\begin{aligned} f_n(w_0) &\leqslant R(w_0) + \frac{\|\widehat{C}_n^\dagger\|}{2\lambda_n}\|\widehat{C}_n w_0 - \widehat{u}_n\|^2, \\ &\leqslant R(w_0) + O\left(\frac{\|\widehat{C}_n - C\| + \|u - \widehat{u}_n\|}{\lambda_n}\right)^2, \\ &\leqslant R(w_0) + o(1). \end{aligned}$$

We can make a similar reasoning on the sequence $(w_{\lambda_n})_{n\in\mathbb{N}}$, and deduce that

$$f_{\overline{\lambda}}(w_{\lambda_n}) \leqslant R(w_0) + o(1), \tag{4}$$
$$\text{and } f_{\overline{\lambda},n}(\widehat{w}_{\lambda_n,n}) \leqslant R(w_0) + o(1). \tag{5}$$

To prove the boundedness of $(\widehat{w}_{\lambda_n,n})_{n\in\mathbb{N}}$ and $(w_{\lambda_n})_{n\in\mathbb{N}}$, we will use arguments relying on the notion of asymptotic or recession function; see (Bauschke and Combettes, 2011, Definition 10.32) for a definition. Define $f_0(w) := R(w) + \iota_{\{u\}}(Cw)$, where $\iota_{\{u\}}$ is the indicator function[1] of the singleton $\{u\}$. The hypothesis ($\mathrm{H_M}$) indicates that $\operatorname{argmin} f_0 = \{w_0\}$, so in particular $\operatorname{argmin} f_0$ is compact. We can then invoke (Auslender and Teboulle, 2003, Proposition 3.1.2 and 3.1.3) to deduce that $f_0^\infty(w) > 0$ for all $w \in \mathbb{R}^p \setminus \{0\}$, where f_0^∞ is the recession function of f_0. From the sum rule (Auslender and Teboulle, 2003, Proposition 2.6.1), we deduce that $f_0^\infty = R^\infty + (\iota_{\{u\}} \circ C)^\infty$. Moreover, we know from ($\mathrm{H_M}$) that $u \in \operatorname{Im} C$, so we can use (Auslender and Teboulle, 2003, Proposition 2.6.1) to get $(\iota_{\{u\}} \circ C)^\infty = \iota_{\{0\}} \circ C = \iota_{\operatorname{Ker} C}$. We deduce from all this that $R^\infty(w) > 0$ for all $w \in \operatorname{Ker} C \setminus \{0\}$, which can be equivalently reformulated as

$$\operatorname{Ker} R^\infty \cap \operatorname{Ker} C = \{0\}. \tag{6}$$

Let us start with the boundedness of $(w_{\lambda_n})_{n\in\mathbb{N}}$. Combining (Auslender and Teboulle, 2003, Proposition 2.6.1), (Auslender and Teboulle, 2003, Example 2.5.1) and the fact that $u \in \operatorname{Im} C$, the recession function of $f_{\overline{\lambda}}$ reads $f_{\overline{\lambda}}^\infty(w) = R^\infty(w)$ if $w \in \ker C$ and $+\infty$ otherwise. Thus, (14) is equivalent to $f_{\overline{\lambda}}^\infty(w) > 0$ for all $w \neq 0$. This is equivalent to saying that $f_{\overline{\lambda}}$ is level-bounded (see (Auslender and Teboulle, 2003, Proposition 3.1.3)), from which we deduce boundedness of $(w_{\lambda_n})_{n\in\mathbb{N}}$ via (12) and (13).

[1] The indicator function ι_Ω of a set $\Omega \subset \mathbb{R}^p$ is by definition equal to 0 when evaluated on Ω, and $+\infty$ elsewhere.

We now turn on $(\widehat{w}_{\lambda_n,n})_{n\in\mathbb{N}}$. We write $\widehat{u}_n = C\widehat{p}_n$ since $\widehat{u}_n \in \operatorname{Im} \widehat{C}_n \subset \operatorname{Im} C$. We first observe that (12) and (13) can be rewritten as:

$$\frac{1}{2\overline{\lambda}}\|\widehat{C}_n(\widehat{w}_{\lambda_n,n} - \widehat{p}_n)\|^2_{\widehat{C}_n^\dagger} + R(\widehat{w}_{\lambda_n,n}) \leqslant R(w_0) + o(1).$$

Let $V_n \operatorname{diag}(s_{n,i}) V_n^\top$ be a (reduced) eigendecomposition of $\widehat{C}_n$. By our assumptions, we have $\underline{s} := \inf_{n,1\leqslant i\leqslant r} s_{n,i} = \left(\sup_n \|\widehat{C}_n\|\right)^{-1} > 0$. In addition, the columns of V_n form an orthonormal basis of $\operatorname{Im} C$ for n large enough. Thus, for all such n, we have

$$\begin{aligned} &\underline{s}\|\operatorname{proj}_{\operatorname{Im} C}(\widehat{w}_{\lambda_n,n} - \widehat{p}_n)\|^2 \\ =\ &\underline{s}\|V_n^\top(\widehat{w}_{\lambda_n,n} - \widehat{p}_n)\|^2 \\ \leqslant\ &\sum_{i=1}^r s_{n,i}|\langle v_{n,i},\, \widehat{w}_{\lambda_n,n} - \widehat{p}_n\rangle|^2 \\ =\ &\langle \widehat{C}_n(\widehat{w}_{\lambda_n,n} - \widehat{p}_n),\, \widehat{w}_{\lambda_n,n} - \widehat{p}_n\rangle \\ =\ &\|\widehat{C}_n(\widehat{w}_{\lambda_n,n} - \widehat{p}_n)\|^2_{\widehat{C}_n^\dagger}. \end{aligned}$$

Altogether, we get the bound

$$\frac{\underline{s}}{2\overline{\lambda}}\|\operatorname{proj}_{\operatorname{Im} C}(\widehat{w}_{\lambda_n,n} - \widehat{p}_n)\|^2 + R(\widehat{w}_{\lambda_n,n}) \leqslant R(w_0) + o(1)$$

for n sufficiently large. Arguing as above, the recession function of $g := \frac{\underline{s}}{2\overline{\lambda}}\|\cdot - \widehat{p}_n\|^2 \circ \operatorname{proj}_{\operatorname{Im} C} + R$ is again $g^\infty(w) = R^\infty(w)$ if $w \in \ker C$ and $+\infty$ otherwise, independently of $\widehat{p}_n$[2]. Our assumption plugged into (Auslender and Teboulle, 2003, Proposition 3.1.3) entails that g is level-bounded and thus boundedness for $(\widehat{w}_{\lambda_n,n})_{n\in\mathbb{N}}$. □

Lemma 3. *Assume that* ($\mathrm{H_M}$) *holds. Then*

$$\operatorname*{Argmin}_{\eta\in\operatorname{Im} C} R^*(\eta) - \langle C^\dagger u,\, \eta\rangle = \partial R(w_0) \cap \operatorname{Im} C.$$

Proof. Using (Bauschke and Combettes, 2011, Proposition 13.23 & Theorem 15.27), one can check that problem

$$\min_{\eta\in\operatorname{Im} C} R^*(\eta) - \langle C^\dagger u,\, \eta\rangle$$

is the Fenchel dual of ($\mathrm{P_0}$). Moreover, $(w^\star, \eta^\star)$ is a primal-dual (Kuhn-Tucker) optimal pair if and only if

$$\begin{pmatrix} w^\star \\ \eta^\star \end{pmatrix} \in \begin{pmatrix} C^\dagger u + \ker C \\ \partial R(w^\star) \cap \operatorname{Im} C \end{pmatrix}.$$

As we assumed in ($\mathrm{H_M}$) that w_0 is the unique minimizer of ($\mathrm{P_0}$), the claimed identity follows. □

Lemma 4. *Let $n \in \mathbb{N}$ and assume that $\operatorname{Im} \widehat{C}_n \subset \operatorname{Im} C$. Denote $r_n := \max\{\|\widehat{u}_n - u\|, \|\widehat{C}_n - C\|\}$. Then,*

$$\|C(\widehat{w}_{\lambda,n} - w_\lambda)\| \leqslant (\|C\|\|C^\dagger\|)^{1/2}(1 + \|\widehat{w}_{\lambda,n}\|)r_n.$$

[2] This reflects the geometric fact that the recession function is unaffected by translation of the argument.

Proof. The first-order optimality conditions for both $\widehat{w}_{\lambda,n}$ and w_λ yield

$$\begin{cases} 0 & \in \lambda \partial R(\widehat{w}_{\lambda,n}) + \widehat{C}_n \widehat{w}_{\lambda,n} - \widehat{u}_n \\ 0 & \in \lambda \partial R(w_\lambda) + C w_\lambda - u. \end{cases}$$

In view of monotonicity of ∂R, we deduce that

$$0 \leqslant \langle \widehat{u}_n - u + C w_\lambda - \widehat{C}_n \widehat{w}_{\lambda,n},\, \widehat{w}_{\lambda,n} - w_\lambda \rangle.$$

Rearranging the terms, we get

$$\begin{aligned} & \langle C(\widehat{w}_{\lambda,n} - w_\lambda),\, \widehat{w}_{\lambda,n} - w_\lambda \rangle \qquad (7) \\ \leqslant\ & \langle \widehat{u}_n - u + (C - \widehat{C}_n)\widehat{w}_{\lambda,n},\, \widehat{w}_{\lambda,n} - w_\lambda \rangle. \end{aligned}$$

By virtue of standard properties of the Moore-Penrose pseudo-inverse and the fact that $\widehat{u}_n - u$ and $C - \widehat{C}_n$ both live in $\operatorname{Im} C \supset \operatorname{Im} \widehat{C}_n$, we obtain

$$\begin{aligned} & \langle C^\dagger(C\widehat{w}_{\lambda,n} - C w_\lambda),\, C\widehat{w}_{\lambda,n} - C w_\lambda \rangle \\ \leqslant\ & \langle C^\dagger(\widehat{u}_n - u + (C - \widehat{C}_n)\widehat{w}_{\lambda,n}),\, C\widehat{w}_{\lambda,n} - C w_\lambda \rangle. \end{aligned}$$

Applying the Cauchy-Schwarz and triangle inequalities, we arrive at

$$\begin{aligned} & \|C\widehat{w}_{\lambda,n} - C w_\lambda\|_{C^\dagger} \\ \leqslant\ & \|\widehat{u}_n - u\|_{C^\dagger} + \|(C - \widehat{C}_n)\widehat{w}_{\lambda,n}\|_{C^\dagger} \\ \leqslant\ & \|C^\dagger\|^{1/2} \left(\|\widehat{u}_n - u\| + \|C - \widehat{C}_n\| \|\widehat{w}_{\lambda,n}\| \right) \\ \leqslant\ & \|C^\dagger\|^{1/2} (1 + \|\widehat{w}_{\lambda,n}\|) r_n. \end{aligned}$$

On the left side of this inequality, we exploit the fact that $\|C\|^{-1}$ is the smallest nonzero eigenvalue of $C^\dagger$ on $\operatorname{Im}(C)$ to conclude

$$\begin{aligned} \|C\widehat{w}_{\lambda,n} - C w_\lambda\|_{C^\dagger} &\leqslant \|C\|^{1/2} \|C\widehat{w}_{\lambda,n} - C w_\lambda\|_{C^\dagger} \\ &\leqslant (\|C\|\|C^\dagger\|)^{1/2} (1 + \|\widehat{w}_{\lambda,n}\|) r_n. \end{aligned}$$

□

A.2 Proof of Proposition 2

Convergence of the primal variable. To lighten notations, we will write $\widehat{w}_n := \widehat{w}_{\lambda_n,n}$. From Lemma 2 we know that $(\widehat{w}_n)_{n\in\mathbb{N}}$ is bounded a.s., so it admits a cluster point, say $w^\star$. Let $\widehat{w}_n$ be a subsequence (we do not relabel for simplicity) converging a.s. to $w^\star$. Now, let $\varepsilon_n := \widehat{u}_n - \widehat{C}_n w_0$, for which we know that both ε_n and ε_n/λ_n are $o(1)$, thanks to Lemma 1(i) and the fact that $u = C w_0$. From the optimality of $\widehat{w}_n$, we obtain

$$\begin{aligned} & \lambda_n R(\widehat{w}_n) + \frac{1}{2}\langle \widehat{C}_n \widehat{w}_n, \widehat{w}_n \rangle - \langle \widehat{u}_n, \widehat{w}_n \rangle \\ \leqslant\ & \lambda_n R(w_0) + \frac{1}{2}\langle \widehat{C}_n w_0, w_0 \rangle - \langle \widehat{u}_n, w_0 \rangle, \end{aligned}$$

which can be equivalently rewritten as

$$\begin{aligned} & \frac{1}{2}\langle \widehat{C}_n(\widehat{w}_n - w_0), \widehat{w}_n - w_0 \rangle - \langle \widehat{w}_n - w_0, \varepsilon_n \rangle \qquad (8) \\ \leqslant\ & \lambda_n (R(w_0) - R(\widehat{w}_n)). \end{aligned}$$

Passing to the limit in (16) and using the fact that R is bounded from below, we obtain

$$\langle C(w^\star - w_0),\, w^\star - w_0 \rangle = 0 \text{ a.s. },$$

or equivalently, that $C w^\star = C w_0 = u$ a.s. since C is positive semi-definite. In addition, as $\widehat{C}_n$ is also positive semi-definite, so we can rewrite (16) as

$$R(\widehat{w}_n) \leqslant R(w_0) + \langle \widehat{w}_n - w_0,\, \frac{\varepsilon_n}{\lambda_n} \rangle. \qquad (9)$$

Passing to the limit in (17), using lower-semicontinuity of R and that $\varepsilon_n/\lambda_n = o(1)$ a.s., we arrive at

$$R(w^\star) \leqslant \liminf_n R(\widehat{w}_n) \leqslant \limsup_n R(\widehat{w}_n) \leqslant R(w_0) \quad a.s.$$

Clearly $R(w^\star) \leqslant R(w_0)$ and $w^\star$ obeys the constraint $C w^\star = u$, which implies that $w^\star$ is a solution of (P_0) a.s. But since this problem has a unique solution, w_0, by assumption (H_M), we conclude that $w^\star = w_0$ a.s. This being true for any a.s. cluster point means that $\widehat{w}_n \to w_0$ as $n \to +\infty$ a.s.

Convergence of the dual variable. Here we omit systematically mentioning that the bounds and convergence we obtain hold almost surely.

It can be verified, using for instance (Bauschke and Combettes, 2011, Proposition 13.23 & Theorem 15.27), that the Fenchel dual problem of ($\mathrm{P}_{\lambda,n}$) is

$$\{\widehat{\eta}_{\lambda,n}\} := \operatorname*{Argmin}_{\eta \in \operatorname{Im} \widehat{C}_n} R^*(\eta) + \frac{\lambda}{2}\langle \widehat{C}_n^\dagger \eta, \eta \rangle - \langle \widehat{C}_n^\dagger \widehat{u}_n, \eta \rangle. \qquad (10)$$

For any fixed $\lambda > 0$, we also introduce its limit problem[3], as $n \to +\infty$ (which is the dual of (9)):

$$\{\eta_\lambda\} := \operatorname*{Argmin}_{\eta \in \operatorname{Im} C} R^*(\eta) + \frac{\lambda}{2}\langle C^\dagger \eta, \eta \rangle - \langle C^\dagger u, \eta \rangle. \qquad (11)$$

Both problems are strongly convex thanks to positive semi-definiteness of $\widehat{C}_n$ and C, hence uniqueness of the corresponding dual solutions $\widehat{\eta}_{\lambda,n}$ and η_λ. Moreover, from the primal-dual extremality relationships, see (Bauschke and Combettes, 2011, Proposition 26.1.iv.b), $\widehat{\eta}_{\lambda,n}$ and η_λ can be recovered from the corresponding primal solutions as

$$\widehat{\eta}_{\lambda,n} := \frac{\widehat{u}_n - \widehat{C}_n \widehat{w}_{\lambda,n}}{\lambda} \quad \text{and} \quad \eta_\lambda := \frac{u - C w_\lambda}{\lambda}. \qquad (12)$$

In what follows, we prove that $\widehat{\eta}_n$ converges to η_0 when $n \to +\infty$. To lighten notation, we will denote

[3] By Lemma 1, we indeed have $C_n^\dagger \to C^\dagger$ a.s. under our hypotheses.

$r_n := \max\{\|\widehat{u}_n - u\|, \|\widehat{C}_n - C\|\}$, and note $\widehat{\eta}_n = \widehat{\eta}_{\lambda_n,n}$. We have

$$\|\widehat{\eta}_n - \eta_0\| \leqslant \|\widehat{\eta}_n - \eta_{\lambda_n}\| + \|\eta_{\lambda_n} - \eta_0\|. \tag{13}$$

By using (20) and the definition of r_n, we write

$$\begin{aligned}\|\widehat{\eta}_n - \eta_{\lambda_n}\| &= \left\|\frac{\widehat{u}_n - u}{\lambda_n} + \frac{Cw_{\lambda_n} - \widehat{C}_n\widehat{w}_n}{\lambda_n}\right\| \\ &\leqslant O\left(\frac{r_n}{\lambda_n}\right) + \left\|\frac{Cw_{\lambda_n} - \widehat{C}_n\widehat{w}_n}{\lambda_n}\right\|.\end{aligned}$$

The second term on the right hand side can also be bounded as

$$\begin{aligned}\left\|\frac{Cw_{\lambda_n} - \widehat{C}_n\widehat{w}_n}{\lambda_n}\right\| &= \left\|\frac{C(w_{\lambda_n} - \widehat{w}_n)}{\lambda_n} + \frac{C\widehat{w}_n - \widehat{C}_n\widehat{w}_n}{\lambda_n}\right\| \\ &\leqslant \left\|\frac{C(w_{\lambda_n} - \widehat{w}_n)}{\lambda_n}\right\| + \|\widehat{w}_n\|\frac{r_n}{\lambda_n} \\ &= O\left(\frac{r_n}{\lambda_n}\right),\end{aligned}$$

where we used Lemma 4, and Lemma 2 with Lemma 1 in the last inequality. Combining the above inequalities with the fact that $r_n = o(\lambda_n)$ by Lemma 1.(i), we obtain

$$\|\widehat{\eta}_n - \eta_{\lambda_n}\| = O\left(\frac{r_n}{\lambda_n}\right) \overset{n\to+\infty}{\longrightarrow} 0. \tag{14}$$

It remains now to prove that η_λ converges to η_0 when $\lambda \to 0$. To do so, we start by using optimality of η_λ and η_0 for problems (19) and (D_0), together with Lemma 3, to write

$$\begin{aligned}&R^*(\eta_\lambda) + \frac{\lambda}{2}\langle C^\dagger\eta_\lambda, \eta_\lambda\rangle - \langle C^\dagger u, \eta_\lambda\rangle \qquad (15)\\ \leqslant\ & R^*(\eta_0) + \frac{\lambda}{2}\langle C^\dagger\eta_0, \eta_0\rangle - \langle C^\dagger u, \eta_0\rangle \\ \leqslant\ & R^*(\eta_\lambda) + \frac{\lambda}{2}\langle C^\dagger\eta_0, \eta_0\rangle - \langle C^\dagger u, \eta_\lambda\rangle,\end{aligned}$$

from which we deduce that

$$\langle C^\dagger\eta_\lambda, \eta_\lambda\rangle \leqslant \langle C^\dagger\eta_0, \eta_0\rangle. \tag{16}$$

Since $\eta_\lambda \in \operatorname{Im} C = (\ker C^\dagger)^\perp$ (see (19)), we can infer from (24) that $(\eta_\lambda)_{\lambda>0}$ is bounded. Let $\eta^\star$ be any cluster point of this net, and let us verify that $\eta^\star$ must be equal to η_0. First, passing to the limit in (24) shows that

$$\langle C^\dagger\eta^\star, \eta^\star\rangle \leqslant \langle C^\dagger\eta_0, \eta_0\rangle. \tag{17}$$

Second, taking the limit in (23) and using lower semi-continuity of R^*, we get

$$\begin{aligned}&R^*(\eta^\star) - \langle C^\dagger u, \eta^\star\rangle \qquad (18)\\ \leqslant\ & \liminf_{\lambda\to 0} R^*(\eta_\lambda) + \frac{\lambda}{2}\langle C^\dagger\eta_\lambda, \eta_\lambda\rangle - \langle C^\dagger u, \eta_\lambda\rangle \\ \leqslant\ & \lim_{\lambda\to 0} R^*(\eta_0) + \frac{\lambda}{2}\langle C^\dagger\eta_0, \eta_0\rangle - \langle C^\dagger u, \eta_0\rangle \\ =\ & R^*(\eta_0) - \langle C^\dagger u, \eta_0\rangle.\end{aligned}$$

From $\eta_\lambda \in \operatorname{Im} C$ we know that $\eta^\star \in \operatorname{Im} C$ as well, so we can then deduce from (26) and Lemma 3 that

$$\eta^\star \in \partial R(w_0) \cap \operatorname{Im} C. \tag{19}$$

Putting together (25) and (27) shows that $\eta^\star$ is a solution of (D_0), hence $\eta^\star = \eta_0$ by uniqueness of η_0. This being true for any cluster point shows convergence of η_λ to η_0.

A.3 Proof of Proposition 3

We use here the notations h_n and $\widehat{\xi}^k$ introduced in Section 4, and we read directly from hypothesis (H_A) that $\widehat{d}^k = \nabla h_n(\widehat{w}^k) + \widehat{\xi}^k$, $\mathbb{E}\left[\widehat{\xi}^k|\mathcal{F}_k\right] = 0$, $\mathbb{E}\left[\|\widehat{\xi}^k\|^2|\mathcal{F}_k\right] \leqslant \sigma_k^2$ and $\widehat{\xi}^k$ converges a.s. to 0.

Let us start by showing that $\widehat{w}^k$ converges to $\widehat{w}_{\lambda_n,n}$. For this, let w be any solution of ($P_{\lambda,n}$). We can write, using standard identities (e.g. (Bauschke and Combettes, 2011, Corollary 2.14)), that

$$\begin{aligned}&\|\widehat{w}^{k+1} - w\|^2 \qquad (20)\\ =\ & \|(1-\alpha_k)(\widehat{w}^k - w) + \alpha_k(\widehat{z}^k - w)\|^2 \\ =\ & (1-\alpha_k)\|\widehat{w}^k - w\|^2 + \alpha_k\|\widehat{z}^k - w\|^2 \\ & -\alpha_k(1-\alpha_k)\|\widehat{z}^k - \widehat{w}^k\|^2.\end{aligned}$$

Since w is a solution of ($P_{\lambda,n}$), it is a fixed point for the operator $\operatorname{prox}_{\lambda_n\gamma_k R}\circ(\mathrm{Id} - \gamma_k\nabla h_n)$ for any $k \in \mathbb{N}$. Use then the definition of $\widehat{z}^k$ together with the nonexpansiveness of the proximal mapping to obtain

$$\begin{aligned}\|\widehat{z}^k - w\|^2 &\leqslant \|\widehat{w}^k - w + \gamma_k(\nabla h_n(w) - \nabla h_n(\widehat{w}^k) - \xi_k)\|^2 \\ &\leqslant \|\widehat{w}^k - w\|^2 + \gamma_k^2\|\nabla h_n(w) - \nabla h_n(\widehat{w}^k) - \xi_k\|^2 \\ &\quad + 2\gamma_k\langle\widehat{w}^k - w, \nabla h_n(w) - \nabla h_n(\widehat{w}^k) - \xi_k\rangle.\end{aligned}$$

Taking the conditional expectation w.r.t. $\mathcal{F}_k$ in the above inequality, and using the assumptions $\mathbb{E}(\xi_k|\mathcal{F}_k) = 0$ and $\mathbb{E}(\|\xi_k\|^2|\mathcal{F}_k) \leqslant \sigma_k^2$, leads to

$$\begin{aligned}&\mathbb{E}(\|\widehat{z}^k - w\|^2|\mathcal{F}_k) \\ \leqslant\ & \|\widehat{w}^k - w\|^2 + \gamma_k^2\|\nabla h_n(w) - \nabla h_n(\widehat{w}^k)\|^2 + \gamma_k^2\sigma_k^2 \\ & +2\gamma_k\langle\widehat{w}^k - w, \nabla h_n(w) - \nabla h_n(\widehat{w}^k)\rangle.\end{aligned}$$

Since ∇h_n is $1/L$-cocoercive, we obtain

$$\begin{aligned}&\mathbb{E}(\|\widehat{z}^k - w\|^2|\mathcal{F}_k) \\ \leqslant\ & \|\widehat{w}^k - w\|^2 + \gamma_k^2\sigma_k^2 \\ & -\gamma_k(2/L - \gamma_k)\|\nabla h_n(w) - \nabla h_n(\widehat{w}^k)\|^2\end{aligned}$$

After taking the conditional expectation in (28) and combining with the last inequality, we obtain

$$\begin{aligned}&\mathbb{E}(\|\widehat{w}^{k+1} - w\|^2|\mathcal{F}_k) \\ \leqslant\ & \|\widehat{w}^k - w\|^2 + \alpha_k\gamma_k^2\sigma_k^2 \\ & -\gamma_k(2/L - \gamma_k)\|\nabla h_n(w) - \nabla h_n(\widehat{w}^k)\|^2 \\ & -\alpha_k(1-\alpha_k)\mathbb{E}(\|\widehat{z}^k - \widehat{w}^k\|^2|\mathcal{F}_k).\end{aligned}$$

The inequality above means that $(\widehat{w}^k)_{k\in\mathbb{N}}$ is a stochastic quasi-Féjer sequence, and hypothesis ($\mathrm{H_A}$) allows us to use invoke (Combettes and Pesquet, 2015, Proposition 2.3), from which we deduce that $(\widehat{w}^k)_{k\in\mathbb{N}}$ is bounded a.s. Thus $\widehat{w}^k$ has a cluster point. Let $\bar{w}$ be a sequential cluster point of $(\widehat{w}^k)_{k\in\mathbb{N}}$, and $\widehat{w}^k$ be a subsequence (that we do not relabel for simplicity) that converges a.s. to $\bar{w}$. Recalling (8) and (7), and in view of assumption ($\mathrm{H_A}$) and continuity of the gradient, we deduce that

$$\widehat{v}^k \to -\nabla h_n(\bar{w}) \quad \text{and} \quad \widehat{z}^k \to \bar{w} \quad a.s.$$

Since $(\widehat{z}^k, \widehat{v}^k) \in \mathrm{gph}(\lambda_n \partial R)$ and $\lambda_n \partial R$ is maximally monotone, we conclude that $0 \in \nabla h_n(\bar{w}) + \lambda_n \partial R(\bar{w})$, i.e., $\bar{w}$ is minimizer of ($\mathrm{P}_{\lambda,n}$). Since this is true for any cluster point, we invoke (Combettes and Pesquet, 2015, Proposition 2.3(iv)) which yields that $\widehat{w}^k$ converges a.s. to a minimizer of ($\mathrm{P}_{\lambda,n}$). Using again (7), we see that $\widehat{z}^k$ converges a.s. to this same minimizer.