A Additional experiments

A.1 Gaps as a function of time

In this section, we include the counterparts of Figures 7 to 9, but display the duality gap instead of the suboptimality. Indeed, since x^* in not available in practice, the suboptimality cannot be used as a stopping criterion. To create a dual feasible point, we use the classical technique of residual rescaling (Mairal, 2010).

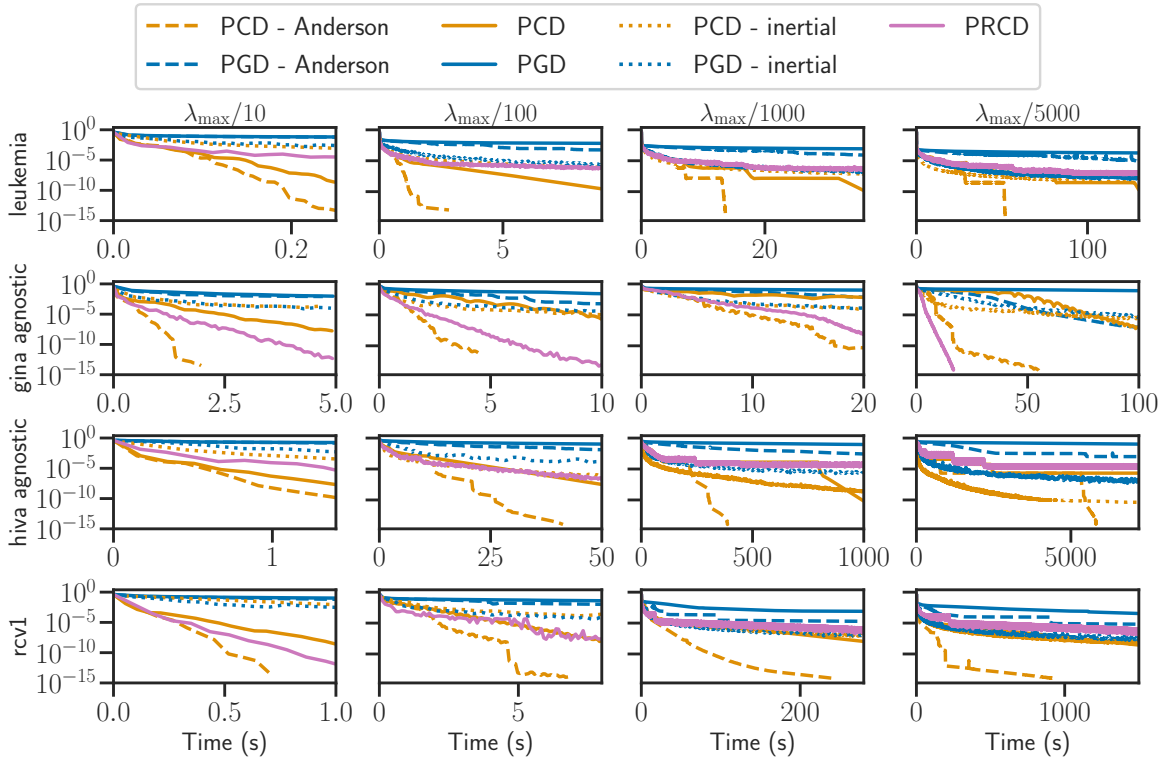

Figure 10: **Lasso, duality gap.** Duality gap along time for the Lasso on various datasets and values of λ.

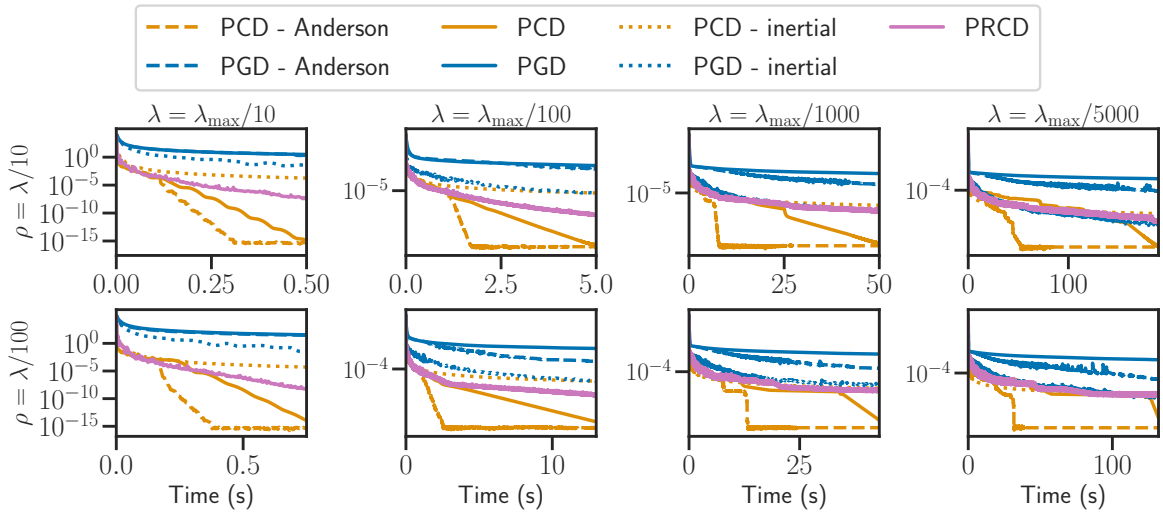

Figure 11: **Elastic net, duality gap.** Duality gap as a function of time for the elastic net on Leukemia dataset, for multiple values of λ and ρ.

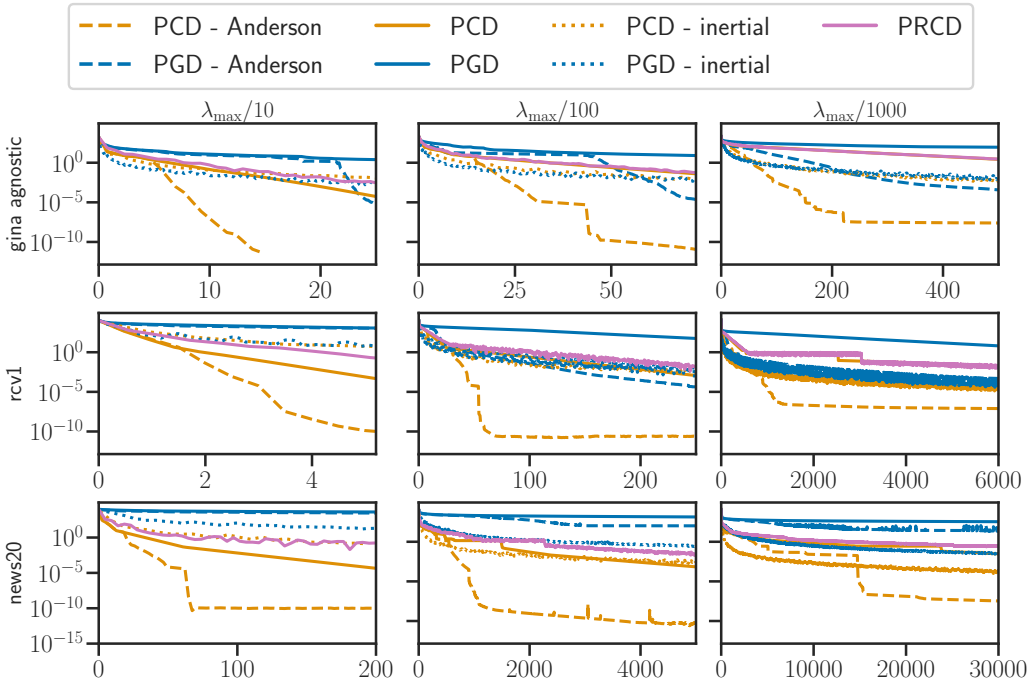

Figure 12: **ℓ_1-regularised logistic regression, duality gap.** Duality gap as a function of time for ℓ_1-regularized logistic regression on multiple datasets and values of λ.

A.2 Group Lasso

In this section we consider the group Lasso, with a design matrix $A \in \mathbb{R}^{n \times p}$, a target $y \in \mathbb{R}^n$, and a partition $\mathcal{G}$ of $[p]$ (elements of the partition being the disjoints groups):

$$\arg\min_{x \in \mathbb{R}^p} \frac{1}{2} \|y - Ax\|^2 + \lambda \sum_{g \in \mathcal{G}} \|x_g\| \quad , \tag{21}$$

where for $g \in \mathcal{G}$, $x_g \in \mathbb{R}^{|g|}$ is the subvector of x composed of coordinates in g. the group Lasso can be solved via proximal gradient descent and by block coordinate descent (BCD), the latter being amenable to Anderson acceleration. As Figure 13 shows, the superiority of Anderson accelerated block coordinate descent is on par with the one observed on the problems studied above.

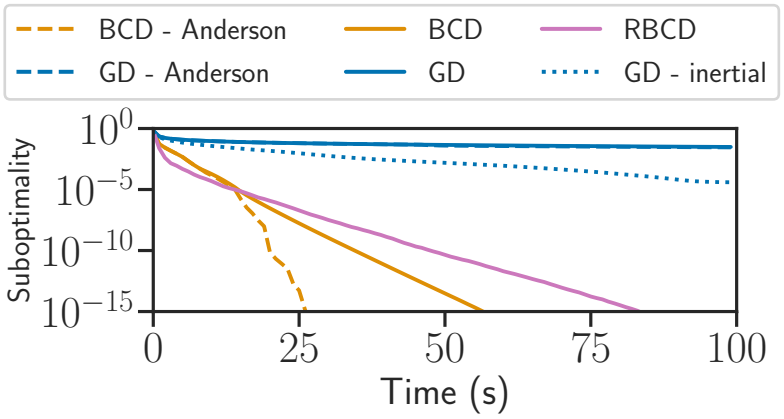

Figure 13: **Group Lasso, suboptimality.** Suboptimality as a function of time for the group Lasso on the *Leukemia* dataset, $\lambda = \lambda_{max}/100$. Groups are artificially taken as consecutive blocks of 5 features.

B Proofs of Propositions 3 and 4

B.1 Proofs of Proposition 3

Lemma 5. First we link the quantity computed in Equation (2) to the extrapolated quantity $\sum_{i=1}^k c_i x^{(i-1)}$. For all $c \in \mathbb{R}^k$ such that $\sum_{i=1}^k c_i = 1$:

$$\sum_{i=1}^k c_i (x^{(i)} - x^{(i-1)}) = (T - \mathrm{Id}) \left(\sum_{i=1}^k c_i x^{(i-1)} - x^* \right) \ . \tag{22}$$

Proof. Since $x^{(i)} = Tx^{(i-1)} + (x^* - Tx^*)$,

$$\begin{aligned} c_i (x^{(i)} - x^{(i-1)}) &= c_i (Tx^{(i-1)} + x^* - Tx^* - x^{(i-1)}) \\ &= (T - \mathrm{Id}) c_i (x^{(i-1)} - x^*) \ . \end{aligned} \tag{23}$$

Hence, since $\sum_1^k c_i = 1$,

$$\sum_{i=1}^k c_i (x^{(i)} - x^{(i-1)}) = (T - \mathrm{Id}) \left(\sum_{i=1}^k c_i x^{(i-1)} - x^* \right) \ . \tag{24}$$

□

Lemma 6. For all $c \in \mathbb{R}^k$ such that $\sum_{i=1}^k c_i = 1$,

$$\|(T - \mathrm{Id})(x^{(k)}_{\text{e-off}} - x^*)\| \leq \sqrt{\kappa(H)} \left\| \sum_{i=0}^{k-1} c_i S^i \right\| \|(T - \mathrm{Id})(x^{(0)} - x^*)\| \ . \tag{25}$$

Proof. In this proof, we denote by c^* the solution of (2). We use the fact that for all $c \in \mathbb{R}^k$ such that $\sum_{i=1}^k c_i = 1$,

$$\| \sum_{i=1}^k c_i^* (x^{(i)} - x^{(i-1)}) \| = \min_{\substack{c \in \mathbb{R}^k \\ \sum_i c_i = 1}} \| \sum_{i=1}^k c_i (x^{(i)} - x^{(i-1)}) \| \leq \| \sum_{i=1}^k c_i (x^{(i)} - x^{(i-1)}) \| \ . \tag{26}$$

Then we use twice Lemma 5 for the left-hand and right-hand side of Equation (26). Using Lemma 5 with the c_i^*

minimizing Equation (2) we have for all $c_i \in \mathbb{R}$ such that $\sum_{i=1}^k c_i = 1$:

$$\begin{aligned}
\|(T - \mathrm{Id})(x_\mathrm{e} - x^*)\| &= \|\sum_{i=1}^k c_i^*(x^{(i)} - x^{(i-1)})\| \\
&\leq \|\sum_{i=1}^k c_i(x^{(i)} - x^{(i-1)})\| \\
&= \|(T - \mathrm{Id})\sum_{i=1}^k c_i(x^{(i-1)} - x^{(*)})\| \\
&= \|(T - \mathrm{Id})\sum_{i=1}^k c_i T^{i-1}(x^{(0)} - x^*)\| \\
&= \|\sum_{i=1}^k c_i T^{i-1}(T - \mathrm{Id})(x^{(0)} - x^*)\| \\
&= \|\sum_{i=1}^k c_i T^{i-1}\| \times \|(T - \mathrm{Id})(x^{(0)} - x^*)\| \\
&\leq \|H^{-1/2}\sum_{i=1}^k c_i S^{i-1} H^{1/2}\| \times \|(T - \mathrm{Id})(x^{(0)} - x^*)\| \\
&\leq \sqrt{\kappa(H)}\|\sum_{i=1}^k c_i S^{i-1}\| \times \|(T - \mathrm{Id})(x^{(0)} - x^*)\| \ . \qquad (27)
\end{aligned}$$

□

Proof. We apply Lemma 6 by choosing c_i equal to the Chebyshev weights c_i^{Cb}. Using the proof of Barré et al. (2020, Prop. B. 2), we have, with $\zeta = \frac{1-\sqrt{1-\rho(T)}}{1+\sqrt{1-\rho(T)}}$:

$$\|\sum_{i=1}^k c_i^{\mathrm{Cb}} S^{i-1}\| \leq \frac{2\zeta^{k-1}}{1+\zeta^{2(k-1)}} \ . \qquad (28)$$

Combined with Lemma 6 this concludes the proof:

$$\begin{aligned}
\|(T - \ \mathrm{Id})(x_\mathrm{e} - x^*)\| &\leq \sqrt{\kappa(H)}\|\sum_{i=1}^k c_i S^{i-1}\|\|(T - \ \mathrm{Id})(x^{(0)} - x^*)\| \qquad (29) \\
&\leq \sqrt{\kappa(H)}\frac{2\zeta^{k-1}}{1+\zeta^{2(k-1)}}\|(T - \ \mathrm{Id})(x^{(0)} - x^*)\| \ . \qquad (30)
\end{aligned}$$

□

B.2 Proof of Proposition 4

Since g_j are $\mathcal{C}^2$ then prox_{g_j} are $\mathcal{C}^1$, see Gribonval and Nikolova (2020, Cor. 1.b). Moreover, f is $\mathcal{C}^2$ and following Massias et al. (2020); Klopfenstein et al. (2020) we have that:

$$\begin{aligned}
&\psi_j : \mathbb{R}^p \to \mathbb{R}^p \\
&x \mapsto \mathrm{prox}_{g_j}\begin{pmatrix} x_1 \\ \vdots \\ x_{j-1} \\ \mathrm{prox}_{\lambda g_j / L_j}\left(x_j - \frac{1}{L_j}\nabla_j f(x)\right) \\ x_{j+1} \\ \vdots \\ x_p \end{pmatrix} \ , \qquad (31)
\end{aligned}$$

is differentiable. Thus we have that the fixed point operator of coordinate descent: $\psi = \psi_p \circ \cdots \circ \psi_1$ is differentiable. Proposition 4 follows from the Taylor expansion of ψ in x^*.