Appendix

PD Methods for Solving Eq. (7)

Algorithm 2 PD Algorithm for solving Eq. (7)

Input: Indicator vector $\widehat{\boldsymbol{\alpha}}$, hyperparameter τ, precision sequence $\{\epsilon^{(b)}\}$, and threshold ϵ_H.
Initialization: Choose any $(\boldsymbol{\beta}^{\text{feas}}, \beta_0^{\text{feas}}) \in \mathbb{S} \times \mathbb{R}$ and $\Upsilon \geq \max\{f_H(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^{\text{feas}}, \beta_0^{\text{feas}}), \min_{\beta_0, \boldsymbol{\gamma}} q(\boldsymbol{\beta}^{\text{feas}}, \beta_0, \boldsymbol{\gamma})\}$ (where f_H is defined in Eq. (2) and q in Eq. (8)). Let $\rho_0 > 0$ and $\sigma > 1$ be arbitrarily chosen. Set $b = 0$ and $\tilde{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^{\text{feas}}$.

1: **repeat** ▷ Beginning of PD
2: Solve Eq. (8) with $\rho = \rho^{(b)}$ by BCD (initialize $s = 0$):
3: **repeat**
3.1: Use IPM to solve Eq. (8) for $\beta = \tilde{\boldsymbol{\beta}}^{(s)}$, *i.e.*, $(\tilde{\beta}_0^{(s+1)}, \tilde{\boldsymbol{\gamma}}^{(s+1)}) \in \text{Argmin}_{\beta_0, \boldsymbol{\gamma}} q(\tilde{\boldsymbol{\beta}}^{(s)}, \beta_0, \boldsymbol{\gamma})$
3.2: Solve Eq. (8) with $\beta_0 = \tilde{\beta}_0^{(s+1)}$ and $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}^{(s+1)}$: $\{i_1, \ldots, i_p\} \leftarrow$ Sort indices of $\tilde{\boldsymbol{\gamma}}^{(s+1)}$ s.t. $|\tilde{\boldsymbol{\gamma}}_{i_j}^{(s+1)}| \geq |\boldsymbol{\gamma}_{i_{j+1}}^{(s+1)}|$

$$\tilde{\boldsymbol{\beta}}_i^{(s+1)} \leftarrow \begin{cases} \tilde{\boldsymbol{\gamma}}_i^{(s+1)}, \text{ if } i \in \{i_1, \ldots, i_\tau\} \\ 0, \text{ otherwise} \end{cases}$$

3.3: $s \leftarrow s + 1$
4: **until** $\max\left\{\frac{\|\tilde{\boldsymbol{\beta}}^{(s+1)} - \tilde{\boldsymbol{\beta}}^{(s)}\|_\infty}{\max\{\|\tilde{\boldsymbol{\beta}}^{(s)}\|_\infty, 1\}}, \frac{|\tilde{\beta}_0^{(s+1)} - \tilde{\beta}_0^{(s)}|}{\max\{|\tilde{\beta}_0^{(s)}|, 1\}}, \frac{\|\tilde{\boldsymbol{\gamma}}^{(s+1)} - \tilde{\boldsymbol{\gamma}}^{(s)}\|_\infty}{\max\{\|\tilde{\boldsymbol{\gamma}}^{(s)}\|_\infty, 1\}}\right\} < \epsilon^{(b)}$
5: Update $\boldsymbol{\beta}^{(b)} \leftarrow \tilde{\boldsymbol{\beta}}^{(s)}$, $\beta_0^{(b)} \leftarrow \tilde{\beta}_0^{(s)}$, $\boldsymbol{\gamma}^{(b)} \leftarrow \tilde{\boldsymbol{\gamma}}^{(s)}$
6: Update $\rho^{(b+1)} \leftarrow \sigma \rho^{(b)}$
7: Update $\tilde{\boldsymbol{\beta}}^{(0)} \leftarrow \begin{cases} \boldsymbol{\beta}^{(b)}, \text{ if } \min_{\beta_0, \boldsymbol{\gamma}} q(\boldsymbol{\beta}^{(b)}, \beta_0, \boldsymbol{\gamma}) \leq \Upsilon \\ \boldsymbol{\beta}^{\text{feas}}, \text{ otherwise} \end{cases}$
8: $b \leftarrow b + 1$
9: **until** $\|\boldsymbol{\beta}^{(b)} - \boldsymbol{\gamma}^{(b)}\|_\infty \leq \epsilon_H$ ▷ Stopping Criteria
Output: $(\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0) = (\boldsymbol{\beta}^{(b)}, \beta_0^{(b)})$

Remarks to Algorithm 2: Each PD iteration performs another BCD to approximate Eq. (8) until the stopping criterion (Step 4 in Algorithm 2) is reached. Specifically, by first fixing $\boldsymbol{\beta}$, Eq. (8) simplifies to a convex optimization problem (Step 3.1 in the Algorithm), which can be solved, for example, by the Interior Point Method (IPM) [90]. Next, $(\beta_0, \boldsymbol{\gamma})$ is fixed and then Eq. (8) becomes $\min_{\boldsymbol{\beta}} \{\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 : \|\boldsymbol{\beta}\|_0 \leq \tau\}$, which can be solved in closed-form (*i.e.*, Step 3.2 in Algorithm 2) according to [65, Proposition 3.1] (also quoted in Section S1 of the Supplement). The computational complexity of this problem is $O(p\log(p))$. Therefore, the computational cost of Algorithm 2 is

$$O(N_P N_B(C_I(n, p) + p\log(p))),$$

where $C_I(n, p)$ is the computational complexity of IPM for ℓ_2-regularized logistic regression problems and N_P (resp., N_B) is the maximum number of PD (resp., BCD) iterations. Note that $C_I(n, p)$ can be different in various implementations but it is not more than polynomial.

The following proposition derives the convergence of PD to a local minimum.

Proposition A.1. *Suppose that $(\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0)$ is an accumulation point of the sequence $\{(\boldsymbol{\beta}^{(b)}, \beta_0^{(b)})\}$ generated by PD. Then $(\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0)$ is a local minimum of Eq. (7).*

Proof. Let $x := (\boldsymbol{\beta}, \beta_0)$, $\mathcal{X} := \mathbb{R}^{p+1}$, $J := \{1, \cdots, p\}$ and $f(x) := f_H(\widehat{\boldsymbol{\alpha}}, x) = l(\widehat{\boldsymbol{\alpha}}, x) + \frac{\lambda}{2}\|x_J\|_2^2$. Then, $\mathcal{X}$ is a closed convex set and $f : \mathbb{R}^{p+1} \to \mathbb{R}$ is a continuously differentiable function. Moreover, f is convex and any level set $\mathcal{X}_\Upsilon := \{x \in \mathcal{X} : f(x) \leq \Upsilon\}$ is compact. Define $\widehat{x} := (\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0)$ and $r := \tau$ then the Robinson condition [65, Theorem 2.1] (see also Section S1 of the Supplement) holds at $\widehat{x}$. It then follows from Theorem 4.3 of [65] that $\widehat{x}$ is a local minimum of Eq. (7). □

P-BCD Methods for Solving Eq. (9)

Algorithm 3 P-BCD Algorithm for solving Eq. (9)

Input: Indicator vector $\widehat{\boldsymbol{\alpha}}$, hyperparameters λ and r, precision sequence $\{\epsilon^{(b)}\}$, and threshold ϵ_P.
Initialization: Choose $\boldsymbol{\beta}^{(0)}$, $0 < \mathcal{L}_{\min} < \mathcal{L}_{\max}$, $\nu > 1$, $c > 0$ and integer $N \geq 0$ arbitrarily. Set $b = 0$ and $f^{(0)} = \infty$.

1: **repeat** ▷ Beginning of BCD
2: Use simplex search to solve Eq. (10) with $\boldsymbol{\beta} = \boldsymbol{\beta}^{(b)}$, *i.e.*, $\beta_0^{(b+1)} \leftarrow \text{argmin}_{\beta_0} l(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^{(b)}, \beta_0)$
3: Solve Eq. (11) with $\beta_0 = \beta_0^{(b+1)}$ and $\tilde{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^{(b)}$ via NPG (initialize $s = 0$):
4: **repeat**
Choose any $\mathcal{L}^{(s)} \in [\mathcal{L}_{\min}, \mathcal{L}_{\max}]$
5: **repeat**
5.1: $g^{(s)} = \tilde{\boldsymbol{\beta}}^{(s)} - \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}^{(s)}, \beta_0^{(b+1)}) / \mathcal{L}^{(s)}$
5.2: $\{i_1, \ldots, i_p\} \leftarrow$ Sort $g^{(s)}$, s. t. $|g_{i_j}^{(s)}| \leq |g_{i_{j+1}}^{(s)}|$

5.3: $$\tilde{\boldsymbol{\beta}}_i^{(s+1)} \leftarrow \begin{cases} \text{sign}(g_i^{(s)}) \max(|g_i^{(s)}| - \lambda/\mathcal{L}^{(s)}, 0), \\ \quad \text{if } i \in \{i_1, \ldots, i_{p-r}\} \\ g_i^{(s)}, \text{ otherwise} \end{cases}$$

5.4: $\mathcal{L}^{(s)} \leftarrow \nu \mathcal{L}^{(s)}$
5.5: $f^{(s+1)} \leftarrow f_P(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}^{(s+1)}, \beta_0^{(b+1)})$ (as in Eq. (4))
6: **until** $f^{(s+1)} \leq \max_{\max(s-N, 0) \leq j \leq s} f^{(j)} - \frac{c}{2}\|\tilde{\boldsymbol{\beta}}^{(s+1)} - \tilde{\boldsymbol{\beta}}^{(s)}\|_2^2$
7: $s \leftarrow s + 1$
8: **until** $\|\nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}^{(s)}, \beta_0^{(b+1)}) - \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}^{(s-1)}, \beta_0^{(b+1)}) - \mathcal{L}^{(s-1)}(\tilde{\boldsymbol{\beta}}^{(s)} - \tilde{\boldsymbol{\beta}}^{(s-1)})\|_\infty < \epsilon^{(b)}$
9: Update $\boldsymbol{\beta}^{(b+1)} \leftarrow \tilde{\boldsymbol{\beta}}^{(s)}$
10: Update $b \leftarrow b + 1$
11: **until** $\min\{|f^{(s)} - f_P(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^{(b-1)}, \beta_0^{(b-1)})| / |f^{(s)}|, |f^{(s)}|\} \leq \epsilon_P$
Output: $(\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0) = (\boldsymbol{\beta}^{(b)}, \beta_0^{(b)})$

Remarks to Algorithm 3: P-BCD solves Eq. (9) by iterating between updating β_0 via the simplex search approach and $\boldsymbol{\beta}$ via the Nonmonotone Proximal Gradient (NPG) method. At each iteration of P-BCD, NPG (Steps 3-8) updates $\boldsymbol{\beta}$ by iteratively determining the solution to Eq. (11) until the stopping criterion (*i.e.*, the convergence with respect to $\boldsymbol{\beta}$, Step 8) is reached. At each iteration of NPG, the method estimates $\tilde{\boldsymbol{\beta}}^{(s+1)}$ by minimizing a proximal function $l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}^{(s)}, \beta_0^{(b+1)}) + \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}^{(s)}, \beta_0^{(b+1)})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}^{(s)}) + \frac{\mathcal{L}^{(s)}}{2}\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}^{(s)}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1^{(r)}$ with $\nabla_{\boldsymbol{\beta}} l(\cdot, \cdot, \cdot)$ denoting the partial derivative of $l(\cdot, \cdot, \cdot)$ with respect to $\boldsymbol{\beta}$ and $\nabla_{\boldsymbol{\beta}} l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \beta_0) = -\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i (1 + \exp(y_i(\boldsymbol{\beta}^\top \mathbf{x}_i + \beta_0)))^{-1}$. According to [48, Theorem 5.5] (quoted also in Section S1 of the Supplement), this problem has a closed-form solution (*i.e.*, Step 5.3). The computational complexity of this problem is $O(p(n + \log(p)))$. The method updates the current estimate of $\tilde{\boldsymbol{\beta}}^{(s+1)}$ (Step 5.3) until the acceptance criterion (Step 6) is reached, that is, the current objective is slightly smaller than the largest objective from the last

N iterations. Consequently, the computational cost of Algorithm 3 is

$$O(N_B p(n(N_S+n)+(n+\log(p))(\log \bar{L}-\log \mathcal{L}_{\min})/\log \nu)),$$

where $\bar{L} = \max\{\mathcal{L}_{\max}, \nu \underline{L}, \nu(1+c)\}$ for some $\underline{L} > 0$ and N_B (resp., N_S) is the maximum number of BCD iterations (resp., Nelder–Mead search steps).

The convergence of the P-BCD method to a local minimum of Eq. (9) is established in Theorem 2.2, which relies on the assumptions that ${\beta_0}^{(b+1)}$ is an optimal solution of Eq. (10) and that $\boldsymbol{\beta}^{(b+1)}$ is a local minimum of Eq. (11). Since the simplex search method converges to the optimal solution of Eq. (10) according to [66, Theorem 4.1] (see also Section S1 of the Supplement), the former assumption is satisfied trivially. We next show that P-BCD fulfills the latter assumption, namely, NPG converges to a local minimum of Eq. (11).

Proposition A.2. *Suppose that $\tilde{\boldsymbol{\beta}}$ is an accumulation point of the sequence $\{\tilde{\boldsymbol{\beta}}^{(s)}\}$ generated by NPG for Eq. (11). Then $\tilde{\boldsymbol{\beta}}$ is a local minimum of Eq. (11).*

Proof. We first show that $\tilde{\boldsymbol{\beta}}$ is a first-order stationary point (defined as in [48, Definition 4] or Section S1 of the Supplement) of Eq. (11) and then a local minimum of Eq. (11).

To show that $\tilde{\boldsymbol{\beta}}$ is a first-order stationary point, note that $l(\widehat{\boldsymbol{\alpha}}, \cdot, \beta_0^{(b+1)})$ is a continuously differentiable function on $\mathbb{R}^p$. Moreover, $f_P(\widehat{\boldsymbol{\alpha}}, \cdot, \beta_0^{(b+1)}) = l(\widehat{\boldsymbol{\alpha}}, \cdot, \beta_0^{(b+1)}) + \lambda \|\cdot\|_1^{(r)}$ is bounded below and uniformly continuous on any level set $\mathcal{S}(\breve{\boldsymbol{\beta}}) := \{\boldsymbol{\beta} \in \mathbb{R}^p : f_P(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \beta_0^{(b+1)}) \le f_P(\widehat{\boldsymbol{\alpha}}, \breve{\boldsymbol{\beta}}, \beta_0^{(b+1)})\}$. By directly applying [48, Theorem 5.2] (see also Section S1 of the Supplement) to Eq. (11) with $f(\cdot) = l(\widehat{\boldsymbol{\alpha}}, \cdot, \beta_0^{(b+1)})$, $F(\cdot) = f_P(\widehat{\boldsymbol{\alpha}}, \cdot, \beta_0^{(b+1)})$, $\Phi(\cdot) = \|\cdot\|_1^{(r)}$, $L_f = 1$, $A = 1$ and $B = F(\breve{\boldsymbol{\beta}})$, then $\tilde{\boldsymbol{\beta}}$ is a first-order stationary point of Eq. (11), *i.e.*,

$$\mathbf{0} \in \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)}) + \lambda\, \partial\Phi(\tilde{\boldsymbol{\beta}}), \tag{25}$$

where $\partial\Phi(\tilde{\boldsymbol{\beta}}) = \{\gamma : \gamma^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \le \Phi(\boldsymbol{\beta}) - \Phi(\tilde{\boldsymbol{\beta}}), \forall \boldsymbol{\beta} \in \mathbb{R}^p\}$ denotes the subdifferential of Φ at $\tilde{\boldsymbol{\beta}}$.

Now to show that $\tilde{\boldsymbol{\beta}}$ is a local minimum of Eq. (11), let $\mathcal{N}(\tilde{\boldsymbol{\beta}}, \epsilon) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_\infty < \epsilon\}$ be a neighbourhood of $\tilde{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\beta}} \in \mathcal{N}(\tilde{\boldsymbol{\beta}}, \epsilon)$ be arbitrarily chosen. From Eq. (25), we know that

$$-\frac{1}{\lambda} \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)}) \in \partial\Phi(\tilde{\boldsymbol{\beta}}),$$

which along with the definition of $\partial\Phi(\tilde{\boldsymbol{\beta}})$ yields that

$$-\frac{1}{\lambda} \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)})^T (\bar{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \le \Phi(\bar{\boldsymbol{\beta}}) - \Phi(\tilde{\boldsymbol{\beta}}).$$

Using this relation, $\lambda \ge 0$ and the convexity of $l(\widehat{\boldsymbol{\alpha}}, \cdot, \beta_0^{(b+1)})$, we further have

$$\begin{aligned}
&f_P(\widehat{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \beta_0^{(b+1)}) = l(\widehat{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \beta_0^{(b+1)}) + \lambda\Phi(\bar{\boldsymbol{\beta}}) \\
&\ge l(\widehat{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \beta_0^{(b+1)}) - \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)})^T (\bar{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + \lambda\Phi(\tilde{\boldsymbol{\beta}}) \\
&\ge l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)}) + \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)})^T (\bar{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \\
&\quad - \nabla_{\boldsymbol{\beta}} l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)})^T (\bar{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + \lambda\Phi(\tilde{\boldsymbol{\beta}}) \\
&= l(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)}) + \lambda\Phi(\tilde{\boldsymbol{\beta}}) = f_P(\widehat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \beta_0^{(b+1)}),
\end{aligned}$$

where the second inequality is due to the convexity of $l(\widehat{\boldsymbol{\alpha}}, \cdot, \beta_0^{(b+1)})$ on $\mathbb{R}^p$. Given our choice of $\bar{\boldsymbol{\beta}}$, it thus implies that $\tilde{\boldsymbol{\beta}}$ is a local minimum of Eq. (11). □