Supplementary Material—Appendix

A Proofs

Proposition 3.1. *The map $\boldsymbol{\theta} \mapsto \mathbf{W}_{\boldsymbol{\theta}}$ is a continuous linear map from Θ to $\mathbb{R}^{d\times k}$. Moreover, for all $\boldsymbol{\theta} \in \Theta^+$, we have*

$$\|\mathbf{W}_{\boldsymbol{\theta}}\|_{\sigma,1} \leq \sum_{i\in\mathcal{I}} \theta_i = \|\boldsymbol{\theta}\|_1$$

and for any $\mathbf{W} \in \mathbb{R}^{d\times k}$, the vector of its singular values corresponds to $\boldsymbol{\theta} \in \Theta^+$ such that $|\mathrm{supp}(\boldsymbol{\theta})| = \mathrm{rank}(\mathbf{W})$, $\mathbf{W}_{\boldsymbol{\theta}} = \mathbf{W}$ and $\|\boldsymbol{\theta}\|_1 = \|\mathbf{W}\|_{\sigma,1}$.

Proof. The linearity is clear by definition of $\mathbf{W}_{\boldsymbol{\theta}}$. The continuity comes easily as follows. Since $\mathcal{M}$ is compact, there exists a constant M such that $\|\mathbf{M}_i\|_{\sigma,1} \leq M$ for all $i \in \mathcal{I}$. So we write

$$\|\mathbf{W}_{\boldsymbol{\theta}}\|_{\sigma,1} \leq \sum_{i\in\mathcal{I}} |\theta_i| \|\mathbf{M}_i\|_{\sigma,1} \leq \sum_{i\in\mathcal{I}} |\theta_i| M = M\|\theta\|_1 \ ,$$

which proves continuity. By definition of the trace norm, the SVD establishes that any $\mathbf{W} \in \mathbb{R}^{n\times k}$ has a non-negative representation in Θ. □

Theorem 3.2. *The function $\psi_\lambda \colon \Theta \to \mathbb{R}$ is convex and differentiable. The following optimization problems are equivalent, i.e., they have the same optimal value and correspondence of optimal solutions as*

$$\hat{\boldsymbol{\theta}} \in \operatorname*{Arg\,min}_{\boldsymbol{\theta}\in\Theta^+} \psi_\lambda(\boldsymbol{\theta}) \quad \textit{iff} \quad \mathbf{W}_{\hat{\boldsymbol{\theta}}} \in \operatorname*{Arg\,min}_{\mathbf{W}\in\mathbb{R}^{d\times k}} \phi_\lambda(\mathbf{W}) \ .$$

Proof. The function ψ is the composition of $\boldsymbol{\theta} \mapsto \mathbf{W}_{\boldsymbol{\theta}}$ with ϕ. Since the first function is linear and the second convex, ψ is convex. Since the first function is continuous and linear, and the second differentiable, ψ is also differentiable. Moreover the function $\boldsymbol{\theta} \mapsto \sum_{i\in\mathcal{I}} \theta_i$ is obviously linear and continuous (so differentiable). So we can conclude that ψ_λ is convex and differentiable.

Let $\hat{\mathbf{W}}$ be a minimizer of (3) (the existence is proved at the end of Sec. 2.5). Let $\hat{\boldsymbol{\theta}}$ be the vector of Θ^+ made by the singular values of $\hat{\mathbf{W}}$, so that we have $\phi_\lambda(\hat{\mathbf{W}}) = \psi_\lambda(\hat{\boldsymbol{\theta}})$. Now write

$$\phi_\lambda(\hat{\mathbf{W}}) = \psi_\lambda(\hat{\boldsymbol{\theta}}) \geq \min \psi_\lambda(\boldsymbol{\theta}) \geq \min \phi_\lambda(\mathbf{W}) \geq \phi_\lambda(\hat{\mathbf{W}}).$$

All the above inequalities are in fact equalities, which proves that the optimal values coincide and in particular that ψ_λ has a minimizer, showing the "if" implication.

We now prove the converse implication. Let $\hat{\boldsymbol{\theta}}$ be a minimizer of (5). Since the optimal values coincide, we have

$$\psi_\lambda(\hat{\boldsymbol{\theta}}) = \min \psi_\lambda(\boldsymbol{\theta}) = \min_{\mathbf{W}} \phi_\lambda(\mathbf{W}) \leq \phi_\lambda(\mathbf{W}_{\hat{\boldsymbol{\theta}}}) \leq \psi_\lambda(\hat{\boldsymbol{\theta}}).$$

This proves the "only if" and finishes the proof. □

Theorem 3.3. *Let ε be such that $0 \leq \varepsilon \leq \lambda$. If $\boldsymbol{\theta}$ is an ε-solution of (5), then $\mathbf{W}_{\boldsymbol{\theta}}$ is an ε-solution of (3).*

Proof. Corollary of Thm. D.3. □

Proposition 3.4. *There exist $\alpha, \delta > 0$ such that for all $\varepsilon > 0$, $\boldsymbol{\theta} \in \Theta^+$ and $i \in \mathcal{I}$ such that $\frac{\partial \psi_\lambda}{\partial \theta_i}(\boldsymbol{\theta}) \leq -\varepsilon$, we have*

$$\psi_\lambda(\boldsymbol{\theta} + \delta\, \mathbf{e}_i) \leq \psi_\lambda(\boldsymbol{\theta}) - \alpha\varepsilon^2. \tag{7}$$

Proof. To simplify notation, we set $\mathbf{W} = \mathbf{W}_{\boldsymbol{\theta}}$ and $\mathbf{M} = \mathbf{M}_i$. We introduce also $M = \max_{\mathbf{M}\in\mathcal{M}} \|\mathbf{M}\|$ where $\|\cdot\|$ is the norm from assumption (C) (note that M exists and is finite by compactness of $\mathcal{M}$). We consider the function

$$f(t) = \phi(\mathbf{W} + t\,\mathbf{M}) = \psi(\boldsymbol{\theta} + t\,\mathbf{e}_i) \ .$$

We have, for all t, $f'(t) = \langle \mathbf{M}, \nabla\phi(\mathbf{W} + t\,\mathbf{M})\rangle$, and by assumption (C),

$$f'(t) - f'(0) \leq tH\|\mathbf{M}\|^2 \leq tHM^2 \ .$$

Now we write for any $\delta > 0$

$$\begin{aligned} f(\delta) - f(0) &= \int_0^\delta f'(t)\, dt \\ &= \delta f'(0) + \int_0^\delta (f'(t) - f'(0))\, dt \\ &\leq \delta f'(0) + HM^2\delta^2/2. \end{aligned}$$

Observe now that the assumption $\frac{\partial \psi_\lambda}{\partial \theta_i}(\boldsymbol{\theta}) \leq -\varepsilon$ is equivalent to

$$f'(0) = \langle \mathbf{M}, \nabla\phi(W)\rangle \leq -\varepsilon - \lambda \ .$$

The above two inequalities yield

$$\phi(\mathbf{W} + \delta\,\mathbf{M}) + \lambda\delta \leq \phi(\mathbf{W}) - \delta\varepsilon + HM^2\delta^2/2.$$

Hence, for $\delta = \varepsilon/HM^2$,

$$\begin{aligned} \psi_\lambda(\boldsymbol{\theta} + \delta\,\mathbf{e}_j) &= \phi(\mathbf{W} + \delta\,\mathbf{M}) + \lambda\delta + \lambda\sum_{i\in\mathcal{I}} \theta_i \\ &\leq \phi(\mathbf{W}) - \frac{\varepsilon^2}{2HM^2} + \lambda\sum_{i\in\mathcal{I}} \theta_i \\ &= \psi_\lambda(\boldsymbol{\theta}) - \frac{\varepsilon^2}{2HM^2} \ . \end{aligned}$$

Therefore we have a guaranteed decrease with $\alpha = 1/2HM^2$. □

Theorem 3.5. R1D *provides ε-optimal solutions $\boldsymbol{\theta}_\varepsilon$ and $\mathbf{W}_\varepsilon$ after at most $8\psi_\lambda(\boldsymbol{\theta}_0)/\alpha\varepsilon^2$ iterations.*

Proof. The theorem follows easily from Prop. 3.4, as follows. The first observation is that it is not possible to have two iterations in a row where we enter Step 4. Suppose indeed that in iteration t we enter Step 4. Then:

- either $g_{t+1} \leq -\epsilon/2$ and the algorithm will not enter Step 4 in the next iteration;
- or $g_{t+1} > -\epsilon/2$, in which case the algorithm terminates, because $\boldsymbol{\theta}_{t+1}$ satisfies the condition (b'). This comes from the fact that the iterate $\boldsymbol{\theta}_{t+1}$ satisfies the optimality conditions of the restricted problem at the previous Step 4, namely

  (a") $\forall i \in \operatorname{supp}(\boldsymbol{\theta}_t) : \frac{\partial \psi}{\partial \theta_i}(\boldsymbol{\theta}) \geq -\lambda - \varepsilon$

  (b") $\forall i \in \operatorname{supp}(\boldsymbol{\theta}_{t+1}) : \left| \frac{\partial \psi}{\partial \theta_i}(\boldsymbol{\theta}) + \lambda \right| \leq \varepsilon$

We prove the bound on the number of iterations by contradiction. Assume that there have been more than $8\,\psi_\lambda(\boldsymbol{\theta}_0)/\alpha\epsilon^2$ iterations of the algorithm. By the first observation, this yields that there are more than $4\,\psi_\lambda(\boldsymbol{\theta}_0)/\alpha\epsilon^2$ iterations when we entered Step 3. Therefore, Prop. 3.4 implies that

$$\psi_\lambda(\boldsymbol{\theta}_t) \leq \psi_\lambda(\boldsymbol{\theta}_0) - \frac{4\,\psi_\lambda(\boldsymbol{\theta}_0)}{\alpha\epsilon^2} \cdot \frac{\alpha\varepsilon^2}{4} = 0 \ .$$

This contradicts the fact that the function ψ is non-negative and completes the bound on the number of iterations.

To conclude the proof we need to argue that on termination, the algorithm returns an ε-solution. Above we have already shown that on termination the returned $\boldsymbol{\theta}_t$ satisfies the condition (b'). Condition (a') is implied by $-g_t < \varepsilon/2$ and

$$\begin{aligned} \|\nabla\phi(\mathbf{W}_t)\|_{\sigma,\infty} &\leq \mathbf{u}_t^\top \big(-\nabla\phi(\mathbf{W}_t)\big)\mathbf{v}_t + \varepsilon/2 \\ &= -g_t + \lambda + \varepsilon/2 \leq \lambda + \varepsilon \ , \end{aligned}$$

which concludes the proof. □

Theorem 3.6. *Let $\varepsilon_\ell \to 0$. Define $\mathbf{W}_{\boldsymbol{\theta}_\ell}$ as the solution generated by* **R1D** *with $\varepsilon = \varepsilon_\ell$. Then a subsequence of $(\mathbf{W}_{\boldsymbol{\theta}_\ell})_\ell$ converges to a solution of* (3).

Proof. Since our algorithm is a descent method, we have $\psi_\lambda(\boldsymbol{\theta}_\ell) \leq \psi_\lambda(\boldsymbol{\theta}_0)$ for all ℓ. By the non-negativity of ϕ and by Thm. 3.2, this yields, for all ℓ,

$$\|\mathbf{W}_{\boldsymbol{\theta}_\ell}\|_{\sigma,1} \leq \phi_\lambda(\boldsymbol{\theta}_\ell) \leq \psi_\lambda(\boldsymbol{\theta}_0).$$

Thus the sequence $(\mathbf{W}_{\boldsymbol{\theta}_\ell})_\ell$ is bounded (by $\psi_\lambda(\boldsymbol{\theta}_0)$). Let us extract a converging subsequence $(\mathbf{W}_\ell)_\ell$; let us show that its limit $\mathbf{W}^\star$ is a solution of (3).

With a slight abuse of notation, let us call again $(\varepsilon_\ell)_\ell$ the subsequence associated to $(\mathbf{W}_\ell)_\ell$. By Thm. 3.3, we know that $\mathbf{W}_\ell$ is a ε_ℓ-solution to (3), that is, from (i') and (ii')

$$\begin{aligned} &\|\nabla\phi(\mathbf{W}_\ell)\|_{\sigma,\infty} \leq \lambda + \varepsilon_\ell, \\ &|\lambda\|\mathbf{W}_\ell\|_{\sigma,1} + \langle\nabla\phi(\mathbf{W}_\ell), \mathbf{W}_\ell\rangle| \leq \varepsilon_\ell\psi_\lambda(\boldsymbol{\theta}_0). \end{aligned}$$

Taking the limit $\varepsilon_\ell \to 0$, we get by continuity of the functions

$$\begin{aligned} &\|\nabla\phi(\mathbf{W}^\star)\|_{\sigma,\infty} \leq \lambda, \\ &\lambda\|\mathbf{W}^\star\|_{\sigma,1} + \langle\nabla\phi(\mathbf{W}^\star), \mathbf{W}^\star\rangle = 0 \end{aligned}$$

which are the first order optimality conditions characterizing a solution of (3). □

B Two loss functions

In this appendix, we establish that the two loss functions considered in this paper, namely the objective functions of the learning problems (1) and (2), satisfy the conditions (A–C). For these technical results, we need another matrix norm: the ℓ_1/ℓ_2-operator norm defined as

$$\|\mathbf{D}\|_{1,2} = \sup_{\substack{\mathbf{v}\in\mathbb{R}^k: \\ \mathbf{v}\neq\mathbf{0}}} \frac{\|\mathbf{D}\mathbf{v}\|_2}{\|\mathbf{v}\|_1} \ .$$

Proposition B.1. *Let*

$$\phi(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n L(\mathbf{W}; \mathbf{x}_i, y_i)$$

be the multi-class loss of Eq. (1) *and $M = \sup_i \|\mathbf{x}_i\|_2$. Then ϕ satisfies conditions* (A–C) *for the norm $\|\mathbf{D}\| = \|\mathbf{D}\|_{1,2}$ and the Lipschitz constant $H = M^2$.*

Proposition B.2. *Let*

$$\phi(\mathbf{W}) = \frac{1}{n}\sum_{j=1}^m \sum_{i=1}^{n_j} L_j(\mathbf{W}; \mathbf{x}_{ji}, y_{ji})$$

be the multitask loss of Eq. (2) *and $M = \sup_{j,i} \|\mathbf{x}_{ji}\|_2$. Then ϕ satisfies conditions* (A–C) *for the norm $\|\mathbf{D}\| = \max_j \|\mathbf{D}_j\|_{1,2}$ and the Lipschitz constant $H = M^2$.*

In fact, conditions (A) and (B) are clear from the convexity and non-negativeness of the multinomial logistic loss function $L(\,\cdot\,; \mathbf{x}, y)$. Condition (C) is a corollary of the following lemma.

Lemma B.3. *Let $\mathbf{W}, \mathbf{D} \in \mathbb{R}^{d\times k}$. For all $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathcal{Y}$, we have $L(\mathbf{W}; \mathbf{x}, y) \geq 0$ and*

$$0 \leq \partial^2 L(\mathbf{W} + t\mathbf{D};\, \mathbf{x},\, y)/\partial t^2 \leq \|\mathbf{x}\|_2^2 \|\mathbf{D}\|_{1,2}^2 \ .$$

Proof. For a matrix $\mathbf{D} \in \mathbb{R}^{d\times k}$, let $\mathbf{d}_\ell$ denote its ℓ-th column and $\mathbf{D}_j$ denote its j-th row. The first part follows by observation that $L(\mathbf{W}; \mathbf{x}, y) \geq \log 1 = 0$. For the second part, let $f(t) = L(\mathbf{W} + t\mathbf{D}; \mathbf{x}, y)$, and let

$$p_\ell(t) = \frac{\exp\left\{(\mathbf{w}_\ell + t\mathbf{d}_\ell)^\top \mathbf{x}\right\}}{\sum_{\ell'\in\mathcal{Y}} \exp\left\{(\mathbf{w}_{\ell'} + t\mathbf{d}_{\ell'})^\top \mathbf{x}\right\}} \quad \text{for } \ell \in \mathcal{Y} \ .$$

Note that $\sum_{\ell\in\mathcal{Y}} p_\ell(t) = 1$, i.e., $p_\ell(t)$ is a probability distribution over $\ell \in \mathcal{Y}$ for any fixed t. Furthermore,

$$f'(t) = \left(\sum_{\ell\in\mathcal{Y}} p_\ell(t)\mathbf{d}_\ell^\top \mathbf{x}\right) - \mathbf{d}_y^\top \mathbf{x} \ ,$$

$$f''(t) = \sum_{\ell\in\mathcal{Y}} p_\ell(t)\left(\mathbf{d}_\ell^\top \mathbf{x} - \sum_{\ell'\in\mathcal{Y}} p_{\ell'}(t)\mathbf{d}_{\ell'}^\top \mathbf{x}\right)^2 \ .$$

From the last identity, $f''(t) \geq 0$. Moreover, using that fact that the variance of a random variable in a range $[a, b]$ is at most $(b-a)^2/4$, we obtain

$$\begin{aligned} f''(t) &\leq \frac{1}{4}\max_{\ell,\ell'\in\mathcal{Y}} \left(\mathbf{d}_\ell^\top \mathbf{x} - \mathbf{d}_{\ell'}^\top \mathbf{x}\right)^2 \\ &\leq \|\mathbf{x}\|_2^2 (\max_{\ell\in\mathcal{Y}} \|\mathbf{d}_\ell\|_2)^2 = \|\mathbf{x}\|_2^2 \|\mathbf{D}\|_{1,2} \ . \end{aligned}$$

where the last equality follows because

$$\begin{aligned} \max_{\ell\in\mathcal{Y}} \|\mathbf{d}_\ell\|_2 &= \max_{\ell\in\mathcal{Y}} \sup_{\substack{\mathbf{u}\in\mathbb{R}^d:\\ \|\mathbf{u}\|_2\leq 1}} \mathbf{u}^\top \mathbf{d}_\ell = \sup_{\substack{\mathbf{v}\in\mathbb{R}^k:\\ \|\mathbf{v}\|_1\leq 1}} \sup_{\substack{\mathbf{u}\in\mathbb{R}^d:\\ \|\mathbf{u}\|_2\leq 1}} \mathbf{u}^\top \mathbf{D}\mathbf{v} \\ &= \sup_{\substack{\mathbf{v}\in\mathbb{R}^k:\\ \|\mathbf{v}\|_1\leq 1}} \|\mathbf{D}\mathbf{v}\|_2 = \sup_{\substack{\mathbf{v}\in\mathbb{R}^k:\\ \mathbf{v}\neq\mathbf{0}}} \frac{\|\mathbf{D}\mathbf{v}\|_2}{\|\mathbf{v}\|_1} = \|\mathbf{D}\|_{1,2} \end{aligned}$$

□

C Existing algorithms for trace norm

Here we give some details about the existing trace-norm algorithms mentioned in Sec. 3.3 and used in Sec. 4 in numerical experiments.

Proximal gradient algorithms Proximal gradient methods are specifically tailored to optimize an objective function which is the sum of a smooth function and a non-differentiable regularizer, such as trace norm. They have drawn increasing attention because of their (optimal) guaranteed convergence rate and their ability to deal with large non-smooth problems.

In our context, an iteration of the basic proximal gradient algorithm for solving (3) consists of

$$\mathbf{W}_{t+1} = \mathrm{Prox}_{\sigma,1}\left(\mathbf{W}_t - \frac{1}{H}\nabla\phi(\mathbf{W}_t)\right).$$

The proximal operator $\mathrm{Prox}_{\sigma,1}$ associated with the trace norm (and parameter λ) is obtained by computing a SVD of the matrix and then replacing each singular value σ_i by $\max\{0, (1-\lambda/|\sigma_i|)\}\sigma_i$ (its "soft-thresholding", hence the name of the method of [21]). Accelerated versions of the algorithm [3] use a second variable and combine it with $\mathbf{W}_t$ at marginal extra computational cost with information of previous step.

The basic proximal algorithm has a global convergence rate in $O(1/t)$ where t is the number of iterations of the algorithm. The accelerated version has a convergence rate in $O(1/t^2)$. However, the computational burden of the SVD computed at each iteration is prohibitive for large-scale problems.

Variational formulation by iterative rescaling The trace norm has the variational formulation as a reweighted Frobenius norm (see [1])

$$\|\mathbf{W}\|_{\sigma,1} = \min_{\mathbf{D}\succ\mathbf{0}} \ \mathrm{trace}(\mathbf{W}^\top\mathbf{D}^{-1}\mathbf{W} + \mathbf{D})/2 \ .$$

The learning problem (3) can then be written as smooth optimization problem:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times k}} \min_{\mathbf{D}\succ\mathbf{0}} \ \lambda\,\mathrm{trace}(\mathbf{W}^\top\mathbf{D}^{-1}\mathbf{W}+\mathbf{D}) + \phi(\mathbf{W}) \ . \quad (8)$$

A way to deal with the (open) constraint $\mathbf{D} \succ \mathbf{0}$ is to introduce the barrier function $\mathrm{trace}(\mathbf{D}^{-1})$ controlled by a real parameter $\delta > 0$. So in practice, we consider the family of regularized smooth optimization problems parametrized by δ

$$\min_{\mathbf{W},\mathbf{D}} \ \lambda\,\mathrm{trace}(\mathbf{W}^\top\mathbf{D}^{-1}\mathbf{W} + \mathbf{D} + \delta\mathbf{D}^{-1})/2 + \phi(\mathbf{W}) \ ,$$

replacing the non-smooth learning problem (3).

The above formulation of the problem is particularly well-suited for an alternating direction approach, as follows. The minimization with respect to $\mathbf{D}$

$$\min_{\mathbf{D}\succ\mathbf{0}} \ \mathrm{trace}(\mathbf{W}^\top\mathbf{D}^{-1}\mathbf{W} + \mathbf{D} + \delta\mathbf{D}^{-1})$$

has an explicit solution

$$\mathbf{D} = (\mathbf{W}\mathbf{W}^\top + \delta\mathbf{I}_k)^{1/2}$$

which is computed by SVD (of a $k \times k$-matrix). The minimization over $\mathbf{W}$ consists of minimizing a smooth and (strongly) convex function.

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times k}} \ \lambda\,\mathrm{trace}(\mathbf{W}^\top\mathbf{D}^{-1}\mathbf{W}) + \phi(\mathbf{W}) \ .$$

A wide range of algorithms can be applied to solve this problem, among them (accelerated and stabilized) gradient methods. In practice, rather then solving the problem in $\mathbf{W}$ to optimality, we do only several iterations of such an algorithm.

While bypassing the non-smoothness of the problem, this algorithm loses (part of) the benefit of the trace-norm regularization. Numerical experiments show that this method produces worse solutions when the optimum is of low rank.

Variational formulation by factorization As observed in several works [15, 36, 31], the trace norm has a variational formulation by low-norm factorisation

$$\|\mathbf{W}\|_{\sigma,1} = \min_{\mathbf{W}=\mathbf{U}\mathbf{V}^\top} \ (\|\mathbf{U}\|^2 + \|\mathbf{V}\|^2)/2 \ .$$

The learning problem (3) can then be written as

$$\min_{\mathbf{U},\mathbf{V}} \quad \frac{\lambda}{2}(\|\mathbf{U}\|^2 + \|\mathbf{V}\|^2) + \phi(\mathbf{U}\mathbf{V}^\top) \ . \tag{9}$$

Block-coordinate descent is then an appealing approach for solving the above problems: the minimization with respect to $\mathbf{U}$, and the one with respect to $\mathbf{V}$ are mere smooth convex optimization problems. Again, accelerated and stabilized gradient methods are adapted for tackling them.

In contrast to the original problem (3) and its formulation (8), problem (9) is not jointly convex with respect to the couple $(\mathbf{U}, \mathbf{V})$. As a result, the alternating algorithm can get stuck in local minima, or saddle-points. For example, $\mathbf{U} = \mathbf{V} = \mathbf{0}$ is a critical point but it is not the global minimum. In practice, we observed that the behavior of the algorithm is highly sensitive to the starting point. Problem-dependent tunings as in [36] might be necessary to overcome this weakness.

Conditional gradient approaches Our algorithm shares some similarities with a related family of algorithms, recently applied to learning problems with a *bounded trace-norm constraint* (or low-rank constraint): conditional gradient algorithms. Conditional gradient algorithms [16, 13], a.k.a. Frank-Wolfe algorithms, allow minimize a convex objective ϕ in a simple convex set S. At iteration $(t+1)$, the conditional gradient algorithm first minimizes the linearized objective at the current iterate within the convex set

$$\tilde{\mathbf{W}}_{t+1} := \operatorname*{Arg\,min}_{\mathbf{W}\in S} \quad \langle \mathbf{W} - \mathbf{W}_t, \nabla\phi(\mathbf{W}_t)\rangle \ ,$$

then performs a line search over the line segment joining $\mathbf{W}_t$ and $\tilde{\mathbf{W}}_{t+1}$ to obtain $\mathbf{W}_{t+1}$. Conditional gradient algorithms were applied to learning problems in [10, 19]. Recent works [22, 33] devised conditional gradient algorithms to learning problems with a trace-norm or low-rank constraint, with applications to collaborative filtering. However, these algorithms worked on *constrained formulations*, whereas we consider a *penalized formulation*. Penalized and constrained formulations are equivalent when the entire regularization path is calculated. Even though for each penalty coefficient there exists a constraint that yields the same solution, we are not aware of a method to obtain the matching constraint that does not involve solving the full optimization problem. In our experience, the regularization path calculation is more stable for penalized versions than for constrained version, and penalized versions are for example the state of the art in ℓ_1-regularization literature.

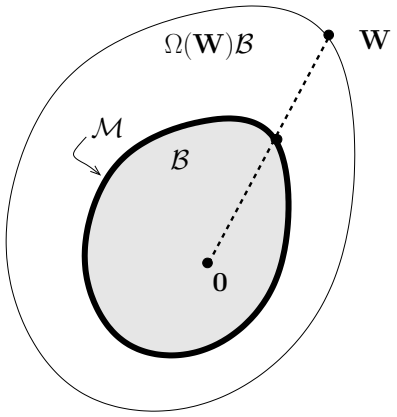

Figure 3: Illustration of the gauge function Ω. To evaluate $\Omega(\mathbf{W})$, we take a ray from the origin towards $\mathbf{W}$ and compute the ratio between the distance to $\mathbf{W}$ and the distance to the intersection of the ray (dotted) with the unit ball $\mathcal{B}$ (in bold).

D Generalization to gauge regularization

In this appendix, we discuss how the optimization algorithm given in the paper generalizes to a broader class of regularization functions. As special cases, we recover coordinate descent for lasso [18], block-coordinate descent for group lasso [28], and rank-one descent discussed in this paper for trace norm. See also [7, 37] for independent, related work.

The specific regularization examples are:

$$\Omega_{\text{lasso}}(\mathbf{W}) = \sum_{j=1}^{d}\sum_{\ell=1}^{k}|\mathbf{W}_{j\ell}| \tag{10}$$

$$\Omega_{\text{gr-lasso}}(\mathbf{W}) = \sum_{j=1}^{d}\|\mathbf{W}_j\|_2 \tag{11}$$

$$\Omega_{\text{trace}}(\mathbf{W}) = \|\mathbf{W}\|_{\sigma,1} \tag{12}$$

where $\mathbf{W}_j$ denotes the j-th row of the matrix. All of them can be naturally defined using the following construction.

Let $\mathcal{M} = \{\mathbf{M}_i \in \mathbb{R}^{d\times k} : \ i \in \mathcal{I}\}$ be a compact set of matrices, called *atoms*, and let $\mathcal{B} := \operatorname{conv}\mathcal{M}$ be its convex hull. We assume that $\mathcal{M}$ is chosen such that $\mathbf{0} \in \operatorname{int}\mathcal{B}$. We think of $\mathcal{M}$ as an "overcomplete basis" and $\mathcal{B}$ as a "unit ball". The *gauge function* Ω and *support function* Ω° associated with $\mathcal{B}$ are convex functions defined as (see illustration in Fig. 3; for further details, see [32, 20, 6])

- $\Omega(\mathbf{W}) := \inf\{t \geq 0 : \ \mathbf{W} \in t\mathcal{B}\}$
- $\Omega^\circ(\mathbf{G}) := \sup_{\mathbf{M}\in\mathcal{B}}\langle \mathbf{M}, \mathbf{G}\rangle = \sup_{\mathbf{M}\in\mathcal{M}}\langle \mathbf{M}, \mathbf{G}\rangle$.

The key property of the gauge function is *sublinearity*:

- $\Omega(t\mathbf{W}) = t\Omega(\mathbf{W})$ for all $\mathbf{W}$ and $t \geq 0$
- $\Omega(\mathbf{W}+\mathbf{W}') \leq \Omega(\mathbf{W})+\Omega(\mathbf{W}')$ for all $\mathbf{W}$ and $\mathbf{W}'$.

In addition, by assuming $\mathbf{0} \in \operatorname{int} \mathcal{B}$, we also obtain:

- $\Omega(\mathbf{W}) \geq 0$, with equality if and only if $\mathbf{W} = \mathbf{0}$
- $\{\mathbf{W} : \Omega(\mathbf{W}) \leq t\} = t\mathcal{B}$ for $t \geq 0$, i.e., level sets are compact.

Unlike norms, gauges are not required to be symmetric. The support function plays the role of the dual norm in that $\langle \mathbf{W}, \mathbf{G} \rangle \leq \Omega(\mathbf{W})\Omega^\circ(\mathbf{G})$ for all $\mathbf{W}, \mathbf{G} \in \mathbb{R}^{d\times k}$.

The three examples Eqs. (10)–(12) are obtained by:

$$\begin{aligned}
\mathcal{M}_{\text{lasso}} &= \big\{ s\mathbf{e}_j\mathbf{e}_\ell^\top : s \in \{-1,1\} \\
&\qquad\qquad j \in \{1,\dots,d\},\, \ell \in \{1,\dots,k\}\big\} \\
\mathcal{M}_{\text{gr-lasso}} &= \{\mathbf{e}_j\mathbf{v}^\top : j \in \{1,\dots,d\},\, \mathbf{v} \in \mathbb{R}^k,\, \|\mathbf{v}\|_2 = 1\} \\
\mathcal{M}_{\text{trace}} &= \{\mathbf{u}\mathbf{v}^\top : \mathbf{u} \in \mathbb{R}^d,\, \mathbf{v} \in \mathbb{R}^k,\, \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1\}
\end{aligned}$$

where $\mathbf{e}_j$ is the j-th vector of the Euclidean basis.

Positing the same assumptions on $\phi(\mathbf{W})$ as in Sec. 2.4, we consider minimization of the regularized objective

$$\operatorname*{Minimize}_{\mathbf{W}\in\mathbb{R}^{d\times k}} \quad \phi_\lambda(\mathbf{W}) \coloneqq \lambda\Omega(\mathbf{W}) + \phi(\mathbf{W}) \ . \tag{13}$$

By compactness of level sets of Ω, lower-boundedness of ϕ, and continuity, the minimum is attained. Furthermore, the subdifferential of Ω is

$$\partial\Omega(\mathbf{W}) = \big\{\mathbf{M} \in \mathbb{R}^{d\times k} : \Omega^\circ(\mathbf{M}) \leq 1,\, \langle \mathbf{M}, \mathbf{W}\rangle = \Omega(\mathbf{W})\big\}$$

hence the ε-optimality is defined as:

(i') $\Omega^\circ\big(-\nabla\phi(\mathbf{W})\big) \leq \lambda + \varepsilon$, and

(ii') $\big|\langle \nabla\phi(\mathbf{W}), \mathbf{W}\rangle + \lambda\Omega(\mathbf{W})\big| \leq \varepsilon\Omega(\mathbf{W})$.

We define Θ, Θ^+ and $\mathbf{W}_{\boldsymbol{\theta}}$ as before, and the lifted problem as

$$\operatorname*{Minimize}_{\boldsymbol{\theta}\in\Theta^+} \quad \psi_\lambda(\boldsymbol{\theta}) \coloneqq \lambda\sum_{i\in\mathcal{I}}\theta_i + \phi(\mathbf{W}_{\boldsymbol{\theta}}) \ . \tag{14}$$

The ε-optimality for Eq. (14) is defined as before:

(a') $\forall i \in \mathcal{I} : \big(-\frac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta})\big) \leq \lambda + \varepsilon$

(b') $\forall i \in \operatorname{supp}(\boldsymbol{\theta}) : \Big|\frac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta}) + \lambda\Big| \leq \varepsilon$

The following is the generalization of Prop. 3.1.

Proposition D.1. *The map $\boldsymbol{\theta} \mapsto \mathbf{W}_{\boldsymbol{\theta}}$ is a continuous linear map from Θ to $\mathbb{R}^{d\times k}$. Moreover, for all $\boldsymbol{\theta} \in \Theta^+$, we have*

$$\Omega(\mathbf{W}_{\boldsymbol{\theta}}) \leq \sum_{i\in\mathcal{I}}\theta_i = \|\boldsymbol{\theta}\|_1$$

and for any $\mathbf{W} \in \mathbb{R}^{d\times k}$ there exists $\boldsymbol{\theta} \in \Theta^+$ such that $|\operatorname{supp}(\boldsymbol{\theta})| \leq (dk+1)$, $\mathbf{W}_{\boldsymbol{\theta}} = \mathbf{W}$ and $\|\boldsymbol{\theta}\|_1 = \Omega(\mathbf{W})$.

Proof. The linearity is clear by definition of $\mathbf{W}_{\boldsymbol{\theta}}$. The continuity comes easily as follows. Consider a norm $\|\cdot\|$ in $\mathbb{R}^{d\times k}$ (all norms are equivalent). Since $\mathcal{M}$ is compact, there exists a constant M such that $\|\mathbf{M}_i\| \leq M$ for all $i \in \mathcal{I}$. So we write

$$\|\mathbf{W}_{\boldsymbol{\theta}}\| \leq \sum_{i\in\mathcal{I}}|\theta_i|\|\mathbf{M}_i\| \leq \sum_{i\in\mathcal{I}}|\theta_i|M = M\|\theta\|_1 \ ,$$

which proves continuity.

We next show that any $\mathbf{W} \in \mathbb{R}^{d\times k}$ has a non-negative representation in Θ. The statement is true for $\mathbf{W} = \mathbf{0}$. Now, take $\mathbf{W} \neq \mathbf{0}$, we have $\Omega(\mathbf{W}) \neq 0$. So, we set $\mathbf{W}' = \mathbf{W}/\Omega(\mathbf{W})$. Since $\mathbf{W}'$ lies in $\mathcal{B} = \operatorname{conv}\mathcal{M}$, it can be written as a convex combination of matrices $\mathbf{M}_i$. By Carathéodory's theorem [20], there exists $\boldsymbol{\theta}' \in \Theta^+$ such that $\sum_{i\in\mathcal{I}}\theta'_i = 1$, $\mathbf{W}' = \mathbf{W}_{\boldsymbol{\theta}'}$, and $|\operatorname{supp}(\boldsymbol{\theta}')| \leq (dk+1)$. Now, define $\boldsymbol{\theta} = \Omega(\mathbf{W})\boldsymbol{\theta}'$. Observe that $\boldsymbol{\theta} \in \Theta^+$, $|\operatorname{supp}(\boldsymbol{\theta})| \leq (dk+1)$, $\mathbf{W}_{\boldsymbol{\theta}} = \Omega(\mathbf{W})\mathbf{W}_{\boldsymbol{\theta}'} = \mathbf{W}$, and $\|\boldsymbol{\theta}\|_1 = \sum_{i\in\mathcal{I}}\theta_i = \Omega(\mathbf{W})\sum_{i\in\mathcal{I}}\theta'_i = \Omega(\mathbf{W})$.

Finally, the inequality comes from the sublinearity of Ω and non-negativity of $\boldsymbol{\theta}$ as follows:

$$\Omega(\mathbf{W}_{\boldsymbol{\theta}}) = \Omega\left(\sum_{i\in\mathcal{I}}\theta_i\mathbf{M}_i\right) \leq \sum_{i\in\mathcal{I}}\theta_i\Omega(\mathbf{M}_i) \leq \sum_{i\in\mathcal{I}}\theta_i \ .$$

□

From Prop. D.1, we obtain

$$\phi_\lambda(\mathbf{W}_{\boldsymbol{\theta}}) \leq \psi_\lambda(\boldsymbol{\theta}) \ .$$

We also obtain the equivalence similar to Thm. 3.2 and the sufficiency of lifted ε-optimality similar to Thm. 3.3.

Theorem D.2. *The function $\psi_\lambda\colon \Theta \to \mathbb{R}$ is convex and differentiable. The following optimization problems are equivalent, i.e., they have the same optimal value and correspondence of optimal solutions as*

$$\hat{\boldsymbol{\theta}} \in \operatorname*{Arg\,min}_{\boldsymbol{\theta}\in\Theta^+}\psi_\lambda(\boldsymbol{\theta}) \quad \textit{iff} \quad \mathbf{W}_{\hat{\boldsymbol{\theta}}} \in \operatorname*{Arg\,min}_{\mathbf{W}\in\mathbb{R}^{d\times k}}\phi_\lambda(\mathbf{W}) \ .$$

Proof. The proof is identical to proof of Thm. 3.2. □

Theorem D.3. *Let ε be such that $0 \leq \varepsilon \leq \lambda$. If $\boldsymbol{\theta}$ is an ε-solution of (14), then $\mathbf{W}_{\boldsymbol{\theta}}$ is an ε-solution of (13).*

Proof. Assume that $\boldsymbol{\theta}$ satisfies conditions (a') and (b'). Note that $\frac{\partial\psi}{\partial\theta_i}(\mathbf{W}) = \langle \mathbf{M}_i, \nabla\phi(\mathbf{W})\rangle$. Hence, condition (a') implies

$$\forall \mathbf{M} \in \mathcal{M} : \ \langle \mathbf{M}, -\nabla\phi(\mathbf{W}_{\boldsymbol{\theta}})\rangle \leq \lambda + \varepsilon \tag{15}$$

which yields

$$\Omega^\circ\big(-\nabla\phi(\mathbf{W}_{\boldsymbol{\theta}})\big) = \sup_{\mathbf{M}\in\mathcal{M}} \langle \mathbf{M}, -\nabla\phi(\mathbf{W}_{\boldsymbol{\theta}})\rangle \leq \lambda + \varepsilon \ ,$$

i.e., condition (i') holds.

It remains to show condition (ii'). First note that

$$\begin{aligned}\langle \nabla\phi(\mathbf{W}_{\boldsymbol{\theta}}), \mathbf{W}_{\boldsymbol{\theta}}\rangle &= \sum_{i\in\mathcal{I}} \theta_i \langle \nabla\phi(\mathbf{W}_{\boldsymbol{\theta}}), \mathbf{M}_i\rangle \\ &= \sum_{i\in\mathcal{I}} \theta_i \left(\tfrac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta})\right) \\ &\leq (-\lambda+\varepsilon)\sum_{i\in\mathcal{I}}\theta_i \leq (-\lambda+\varepsilon)\Omega(\mathbf{W}_{\boldsymbol{\theta}}) \qquad (16)\end{aligned}$$

where the last two inequalities follow by (b') and by Prop. D.1. We also know by Prop. D.1 that there exists $\boldsymbol{\theta}^* \in \Theta^+$ such that $\mathbf{W}_{\boldsymbol{\theta}^*} = \mathbf{W}_{\boldsymbol{\theta}}$, and $\Omega(\mathbf{W}_{\boldsymbol{\theta}}) = \sum_{i\in\mathcal{I}}\theta^*_i$. We can write

$$\begin{aligned}\langle \nabla\phi(\mathbf{W}_{\boldsymbol{\theta}}), \mathbf{W}_{\boldsymbol{\theta}}\rangle &= \sum_{i\in\mathcal{I}} \theta^*_i \langle \nabla\phi(\mathbf{W}_{\boldsymbol{\theta}}), \mathbf{M}_i\rangle \\ &\geq (-\lambda-\varepsilon)\sum_{i\in\mathcal{I}}\theta^*_i = (-\lambda-\varepsilon)\Omega(\mathbf{W}_{\boldsymbol{\theta}})\end{aligned}$$

where the above inequality follows by Eq. (15). Combining with Eq. (16), we obtain

$$(-\lambda-\varepsilon)\Omega(\mathbf{W}_{\boldsymbol{\theta}}) \leq \langle \nabla\phi(\mathbf{W}_{\boldsymbol{\theta}}), \mathbf{W}_{\boldsymbol{\theta}}\rangle \leq (-\lambda+\varepsilon)\Omega(\mathbf{W}_{\boldsymbol{\theta}})$$

i.e., condition (ii') holds as well. □

Using Thm. D.3, we can derive **AtomDescent** (Algorithm 3), a gauge version of **R1D** (Algorithm 1). The only difference is that the computation of top singular-vector pair is now replaced by the *extremal point* evaluation. For $\mathbf{G} \in \mathbb{R}^{d\times k}$, we define

$$\mathrm{Ext}(\mathbf{G}) = \operatorname*{Arg\,max}_{\mathbf{M}\in\mathcal{M}} \ \langle \mathbf{M}, \mathbf{G}\rangle \ .$$

Note that if $\mathbf{M}^* \in \mathrm{Ext}(\mathbf{G})$, we have $\langle \mathbf{M}^*, \mathbf{G}\rangle = \Omega^\circ(\mathbf{G})$. For our three examples, we obtain:

- *lasso:* $\mathbf{M}^* = s^* \mathbf{e}_{j^*}\mathbf{e}_{\ell^*}^\top$
  where $(j^*, \ell^*) = \mathrm{Arg\,max}_{(j,\ell)} |G_{j\ell}|$,
  $s^* = \operatorname{sign} G_{j^*\ell^*}$
- *group lasso:* $\mathbf{M}^* = \mathbf{e}_{j^*}\mathbf{v}^{*\top}$
  where $j^* = \mathrm{Arg\,max}_j \|\mathbf{G}_j\|_2$,
  $\mathbf{v}^{*\top} = \mathbf{G}_{j^*}/\|\mathbf{G}_{j^*}\|_2$
- *trace norm:* $\mathbf{M}^* = \mathbf{u}^*\mathbf{v}^{*\top}$
  where $(\mathbf{u}^*, \mathbf{v}^*)$ is the top singular-vector pair of $\mathbf{G}$

Hence, for lasso we obtain coordinate descent; for group lasso, block-coordinate descent; and for trace norm, rank-one descent. As in **R1D**, also in **AtomDescent**, we only insist on approximate extremal points. All of the convergence results for **R1D** also apply to **AtomDescent**, because analysis of **R1D** only concerned the lifted problem Eq. (5) and did not depend on the particular linear map $\boldsymbol{\theta} \mapsto \mathbf{W}_{\boldsymbol{\theta}}$ (which is the only change between **R1D** and **AtomDescent**).

Algorithm 3 AtomDescent$(\phi, \Omega, \lambda, \boldsymbol{\theta}_0, \varepsilon)$

Input: empirical risk ϕ, gauge Ω, regularization λ
initial point $\mathbf{W}_{\boldsymbol{\theta}_0}$, convergence threshold ε

Output: ε-optimal $\mathbf{W}_{\boldsymbol{\theta}}$

Notation: $\mathbf{W}_t \coloneqq \mathbf{W}_{\theta_t}$, $\mathbf{M}_t \coloneqq \mathbf{M}_{i_t}$, $\mathbf{e}_t \coloneqq \mathbf{e}_{i_t}$

Algorithm:

For $t = 0, 1, 2, \ldots$:

1. Find $i_t \in \mathcal{I}$ corresponding to the coordinate of $\boldsymbol{\theta}$ with approximately steepest descent in positive direction, i.e.,
   $$\langle \mathbf{M}_t, -\nabla\phi(\mathbf{W}_t)\rangle \geq \Omega^\circ(-\nabla\phi(\mathbf{W}_t)) - \varepsilon/2$$
2. Let $g_t \coloneqq \frac{\partial\psi_\lambda}{\partial\theta_{i_t}}(\boldsymbol{\theta}_t) = \lambda + \langle \mathbf{M}_t, \nabla\phi(\mathbf{W}_t)\rangle$
3. If $g_t \leq -\varepsilon/2$
   $\mathbf{W}_{t+1} = \mathbf{W}_t + \delta\,\mathbf{M}_t$ with δ given by Prop. 3.4
   $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \delta\,\mathbf{e}_t$
4. Else (i.e., $g_t > -\varepsilon/2$)
   If $\boldsymbol{\theta}_t$ satisfies (b'), terminate and return $\boldsymbol{\theta}_t$
   Otherwise, compute $\boldsymbol{\theta}_{t+1}$ as an ε-solution of the restricted problem $\min_{\boldsymbol{\theta}\in\mathbb{R}_+^{\mathrm{supp}(\boldsymbol{\theta}_t)}} \psi_\lambda(\boldsymbol{\theta})$

E Infinite dimensional space Θ

In this appendix, we briefly recall some additional material (especially about differentiability and optimality conditions) to demystify the infinite dimensional space Θ. We assume the gauge setting introduced in the previous section.

The completion of the normed space $(\Theta, \|\cdot\|_1)$ is the complete normed space $(\ell_1(\mathcal{I}), \|\cdot\|_1)$, the space of $(\theta_i)_{i\in\mathcal{I}}$ such that $\sum_{i\in\mathcal{I}}|\theta_i| < +\infty$. The two spaces are in duality with the space $(\ell_\infty(\mathcal{I}), \|\cdot\|_\infty)$ equipped with

$$\|\boldsymbol{\delta}\|_\infty = \max_{i\in\mathcal{I}} |\delta_i|$$

through the bracket notation

$$\langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle = \sum_{i\in\mathcal{I}} \delta_i\theta_i \ \leq \ \|\boldsymbol{\delta}\|_\infty \|\boldsymbol{\theta}\|_1 \ .$$

Let $\psi : \Theta \to \mathbb{R}$ be a differentiable function. Its differential $d\psi(\boldsymbol{\theta}) \in \ell_\infty(\mathcal{I})$ can be written with the help of partial derivatives as $d\psi(\boldsymbol{\theta}) = (\frac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta}))_{i\in\mathcal{I}}$. The

general optimality conditions in this context are the following; they are used in the proof of Prop. E.2.

Proposition E.1. *Let $\psi : \Theta \to \mathbb{R}$ be a convex differentiable function, and K a convex subset of Θ. Then $\boldsymbol{\theta}^*$ is a minimum of ψ over K if and only if*

$$\sum_{i\in\mathcal{I}} \frac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta}^*)(\theta_i - \theta_i^*) \geq 0 \qquad \textit{for all } \boldsymbol{\theta} \in K.$$

Proof. The proof is based on the following basic property of convex functions (see [29]). Let $\psi : \Theta \to \mathbb{R}$ be a convex differentiable function; then

$$\psi(\boldsymbol{\eta}) \geq \psi(\boldsymbol{\theta}) + \langle d\psi(\boldsymbol{\theta}), (\boldsymbol{\eta} - \boldsymbol{\theta})\rangle \qquad \text{for all } \boldsymbol{\eta}, \boldsymbol{\theta}. \quad (17)$$

With the help of the above inequality, the implication "if" is obvious. To prove the "only if" implication, take $t > 0$ and write for any $\boldsymbol{\theta} \in K$, by definition of differentiability,

$$\frac{\psi(\boldsymbol{\theta}^* - t(\boldsymbol{\theta} - \boldsymbol{\theta}^*))}{t} = \langle d\psi(\boldsymbol{\theta}^*), (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\rangle + \frac{o(t)}{t} \ .$$

Note that $t > 0$ so $\boldsymbol{\theta}^* - t(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \in K$ and then the left-hand-side is nonnegative. Taking the limit $t \to 0$, we obtain $\langle d\psi(\boldsymbol{\theta}^*), (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\rangle \geq 0$. □

The key assumption is the differentiability of the mapping ψ, that we get, in our context, simply by construction.

In spite of the infinite dimension, the space Θ and the new optimization problem (14) have simple-looking structures, and they share many properties with finite-dimensional analogs. In particular, the optimality conditions are as expected.

Proposition E.2. *The three following properties are equivalent*

(i) $\bar{\boldsymbol{\theta}}$ *is an optimal solution to problem (14)*

(ii) $\forall i \in \mathcal{I} : \ \frac{\partial\psi_\lambda}{\partial\theta_i}(\bar{\boldsymbol{\theta}}) \geq 0$
and $\quad \forall i \in \operatorname{supp}(\bar{\boldsymbol{\theta}}) : \ \frac{\partial\psi_\lambda}{\partial\theta_i}(\bar{\boldsymbol{\theta}}) = 0$

(iii) $\min_{i\in\mathcal{I}} \frac{\partial\psi_\lambda}{\partial\theta_i}(\bar{\boldsymbol{\theta}}) \geq 0$
and $\quad \bar{\boldsymbol{\theta}} \in \operatorname{Arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}_+^{\operatorname{supp}(\bar{\boldsymbol{\theta}})}} \psi_\lambda(\boldsymbol{\theta})$

Proof. To prove the equivalence between (i) and (ii), we apply both implications of Prop. E.1 with $\psi = \psi_\lambda$ and $K = \Theta^+$. We show first (i)⇐(ii). Let $\boldsymbol{\theta} \in K$; we have

$$\sum_{i\in I} \frac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta}^*)(\theta_i - \theta_i^*) = \sum_{i\notin\operatorname{supp}(\boldsymbol{\theta}^*)} \frac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta}^*)\theta_i \leq 0.$$

This is the optimality condition of (14), so we have (i).

We now prove the converse (i)⇒(ii). For all $i \in \mathcal{I}$, we write the optimality condition with $\boldsymbol{\eta} \in \Theta$ defined by

$$\eta_\ell = \begin{cases} \theta_\ell^* & \text{if } \ell \neq i \\ \theta_i^* + 1 & \text{otherwise} \end{cases}$$

to get $\frac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta}^*) \geq 0$. Similarly for all $i \in \mathcal{I}$ such that $\boldsymbol{\theta}_i^* > 0$, we write the optimality condition with $\boldsymbol{\eta} \in \Theta$ defined by

$$\eta_\ell = \begin{cases} \theta_\ell^* & \text{if } \ell \neq i \\ \theta_i^*/2 & \text{otherwise} \end{cases}$$

to get $\frac{\partial\psi}{\partial\theta_i}(\boldsymbol{\theta}^*) \leq 0$, and we can conclude. □