Supplementary material for feature subset selection for the multinomial logit model via mixed-integer optimization

Shunsuke Kamiya Tokyo University of Agriculture and Technology
Ryuhei Miyashiro Tokyo University of Agriculture and Technology
Yuichi Takano University of Tsukuba

A Key assumptions and additional lemmas

In the main body of the paper, the following conditions are assumed:

1. the loss function $\ell(y, \bullet)$ is proper convex for all $y \in [m]$, and (P) is bounded and has an interior feasible solution;
2. the Fenchel conjugate $\hat{\ell}(y, \bullet)$ is continuous on its effective domain and strictly convex for all $y \in [m]$;
3. dom $\hat{\ell}(y, \bullet)$ is nonempty, bounded and closed for all $y \in [m]$.

Before proving our theorems, we must introduce the following lemmas.

Lemma 12 (Strong duality; Theorem 9.6 in [6]). *Let $\boldsymbol{a}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for all $i \in [m]$. Let $f : \mathbb{R}^n \to [-\infty, +\infty]$ be the objective function of the following problem:*

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{a}_i^\top \boldsymbol{x} - b_i = 0, \qquad \forall i \in [m]. \end{aligned}$$

We assume that f is proper convex and that there is a feasible solution on ri dom f*, which denotes the relative interior of* dom f*. Moreover, if this optimization problem is bounded, the optimal values of the following two optimization problems are the same.*

$$\underset{\boldsymbol{\lambda}\in\mathbb{R}^m}{\text{maximize}} \quad \min\left\{\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \mid \boldsymbol{x} \in \mathbb{R}^n\right\},$$

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad \max\left\{\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \mid \boldsymbol{\lambda} \in \mathbb{R}^m\right\},$$

where $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is the Lagrange function

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i(\boldsymbol{a}_i^\top \boldsymbol{x} - b_i).$$

Lemma 13 (Section 3.5 in [4]). *Given $\mathcal{A} \subseteq \mathbb{R}^m$, let $f_{\boldsymbol{\alpha}} : \mathbb{R}^n \to \mathbb{R}$ be subdifferentiable functions for all $\boldsymbol{\alpha} \in \mathcal{A}$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function defined by*

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{\alpha}\in\mathcal{A}} f_{\boldsymbol{\alpha}}(\boldsymbol{x}).$$

Then f is subdifferentiable if $\mathcal{A}$ is compact and the function $\boldsymbol{\alpha} \mapsto f_{\boldsymbol{\alpha}}(\boldsymbol{x})$ is upper semi-continuous for each $\boldsymbol{x}$. The subderivative of f is given by

$$\partial f(\boldsymbol{x}) = \operatorname{conv}\left[\bigcup \{\partial f_{\boldsymbol{\alpha}}(\boldsymbol{x}) \mid \boldsymbol{\alpha} \in \mathcal{A},\, f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = f(\boldsymbol{x})\}\right],$$

where conv *denotes the convex hull of a set.*

Lemma 14. *Given $U \subseteq \mathbb{R}^p$, let $f : \mathbb{R}^n \times U \to \mathbb{R}$ be the objective function of the optimization problem*

$$\underset{\boldsymbol{x} \in S(\boldsymbol{u})}{\text{maximize}} \quad f(\boldsymbol{x}, \boldsymbol{u}), \tag{14}$$

where $S : U \to \mathcal{P}(\mathbb{R}^n)$ is a constraint map. Assume that S is continuous at a point $\bar{\boldsymbol{u}} \in U$ and that the objective function f is continuous on $S(\bar{\boldsymbol{u}}) \times \{\bar{\boldsymbol{u}}\}$. If problem (14) has the unique optimal solution $\bar{\boldsymbol{x}}$ for $\boldsymbol{u} = \bar{\boldsymbol{u}}$, the following map $\Phi : U \to \mathcal{P}(\mathbb{R}^n)$ is continuous at $\bar{\boldsymbol{u}}$:

$$\Phi(\boldsymbol{u}) = \underset{\boldsymbol{x}}{\text{argmax}}\{f(\boldsymbol{x}, \boldsymbol{u}) \mid \boldsymbol{x} \in S(\boldsymbol{u})\}.$$

Proof. The same result is proved for a minimization problem in Theorem 3.30 in [5]. We can use this result directly. □

Lemma 15 (Section 3.2.3 in [3])**.** *Let $\mathcal{A} \subseteq \mathbb{R}^m$ and $f : \mathbb{R}^n \times \mathcal{A} \to [-\infty, +\infty]$. Let $g : \mathbb{R}^n \to [-\infty, +\infty]$ be the function defined by*

$$g(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in \mathcal{A}} f(\boldsymbol{x}, \boldsymbol{y}).$$

If $f(\bullet, \boldsymbol{y})$ is convex for each $\boldsymbol{y} \in \mathcal{A}$, g is convex.

A.1 Proof of Theorem 1

The proof proceeds along the lines of the proof of Theorem 1 in [1]. First, we derive the Lagrangian relaxation of problem (P) as follows:

$$\underset{\boldsymbol{W} \in \mathbb{R}^{m \times p},\, \boldsymbol{b} \in \mathbb{R}^m,\, \boldsymbol{\eta} \in \mathbb{R}^{n \times m}}{\text{minimize}} \ \underset{\boldsymbol{\alpha} \in \mathbb{R}^{n \times m}}{\text{maximize}} \ \sum_{i=1}^n \ell(y_i, \boldsymbol{\eta}_{i\bullet}) + \frac{1}{2\gamma} \sum_{r=1}^m \|\boldsymbol{w}_{r\bullet}\|_2^2 + \sum_{i=1}^n \sum_{r=1}^m \alpha_{ir}(\boldsymbol{w}_{r\bullet}^\top \boldsymbol{x}_{i\bullet} + b_r - \eta_{ir}).$$

From Assumption 1, the loss function $\ell(y, \bullet)$ is proper convex for all $y \in [m]$ and problem (P) is bounded and has a interior feasible solution. Consequently, the strong duality holds by Lemma 12. We now transform this problem into

$$\begin{aligned}
\underset{\boldsymbol{\alpha} \in \mathbb{R}^{n \times m}}{\text{maximize}} \quad & \sum_{i=1}^n \min\left\{\ell(y_i, \boldsymbol{\eta}_{i\bullet}) - \boldsymbol{\alpha}_{i\bullet}^\top \boldsymbol{\eta}_{i\bullet} \mid \boldsymbol{\eta}_{i\bullet} \in \mathbb{R}^m\right\} && (15)\\
& + \sum_{r=1}^m \min\left\{b_r \boldsymbol{1}^\top \boldsymbol{\alpha}_{\bullet r} \mid b_r \in \mathbb{R}\right\} && (16)\\
& + \sum_{r=1}^m \min\left\{\frac{1}{2\gamma}\|\boldsymbol{w}_{r\bullet}\|_2^2 + \boldsymbol{w}_{r\bullet}^\top \boldsymbol{X}^\top \boldsymbol{\alpha}_{\bullet r} \mid \boldsymbol{w}_{r\bullet} \in \mathbb{R}^p\right\}. && (17)
\end{aligned}$$

Because their decision variables are independent, these minimization problems (15)–(17) can be solved separately as follows:

- The optimal value of (15) is $-\hat{\ell}(y_i, \boldsymbol{\alpha}_{i\bullet})$ from the definition of the conjugate function.
- An equality constraint $\boldsymbol{1}^\top \boldsymbol{\alpha}_{\bullet r} = 0$ is obtained because problem (16) must be bounded.
- Problem (17) can be solved analytically; we have the optimal solution $\boldsymbol{w}_{r\bullet}^\star = -\gamma \boldsymbol{X}^\top \boldsymbol{\alpha}_{\bullet r}$.

From these results, we have the dual problem as desired.

A.2 Proof of Lemma 2

First, for any $y \in [m]$ and $\boldsymbol{\alpha} \in \text{dom } \hat{\ell}(y, \bullet)$, the expression $\sum_{r=1}^m \sum_{j=1}^p z_j (\boldsymbol{x}_{\bullet j}^\top \boldsymbol{\alpha}_{\bullet r})^2$ is differentiable with respect to $\boldsymbol{z} \in [0, 1]^p$. In addition, dom $\hat{\ell}(y, \bullet)$ is bounded and closed from Assumption 3. The subderivative of c is therefore obtained by Lemma 13 as

$$\partial c(\boldsymbol{z}) = \text{conv}\left\{\left(-\frac{\gamma}{2}\left\|\boldsymbol{\alpha}^\top \boldsymbol{x}_{\bullet j}\right\|_2^2\right)_{j \in [p]} \,\middle|\, \boldsymbol{1}^\top \boldsymbol{\alpha}_{\bullet r} = 0,\, \forall r \in [m],\, \boldsymbol{\alpha}_{i\bullet} \in \text{dom } \hat{\ell}(y_i, \bullet),\, \forall i \in [n],\, f_{\boldsymbol{z}}(\boldsymbol{\alpha}) = c(\boldsymbol{z})\right\},$$

where $f_{\boldsymbol{z}}$ is the objective function (10). Moreover, $f_{\boldsymbol{z}}$ is strictly convex because of $\boldsymbol{X}\boldsymbol{X}^\top \succeq \boldsymbol{O}$ and the strict convexity of $\hat{\ell}(y, \bullet)$ from Assumption 2. That is, the map $\boldsymbol{\alpha}^\star$ is a monomorphism. We thus obtain the partial derivatives of $\boldsymbol{z}$ as (11).

Next we show the continuity of $\nabla c(\boldsymbol{z})$. The expression (11) is continuous at each $\boldsymbol{z} \in [0, 1]^p$ if the function $\boldsymbol{\alpha}^\star$ is continuous; we therefore show the continuity of $\boldsymbol{\alpha}^\star$ instead. The feasible region

$$\mathcal{A} = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^{n \times m} \;\middle|\; \mathbf{1}^\top \boldsymbol{\alpha}_{\bullet r} = 0,\, \forall r \in [m],\, \boldsymbol{\alpha}_i \in \operatorname{dom} \hat{\ell}(y_i, \bullet),\, \forall i \in [n] \right\}$$

does not depend on $\boldsymbol{z}$, and thus the constraint map of $(\mathrm{D}_{\boldsymbol{z}})$ is trivially continuous at each $\boldsymbol{z} \in [0, 1]^p$. For any $\boldsymbol{z} \in [0, 1]^p$, the objective function (10) is also continuous in $\boldsymbol{\alpha}$. From these facts and the uniqueness of $\boldsymbol{\alpha}^\star(\boldsymbol{z})$, $\boldsymbol{\alpha}^\star$ satisfies the assumptions of Lemma 14. Consequently, $\boldsymbol{\alpha}^\star$ is continuous at each $\boldsymbol{z} \in [0,1]^p$.

From the above discussion, c is continuously differentiable.

A.3 Proof of Lemma 3

For each $\boldsymbol{\alpha} \in \mathcal{A}$, the objective function (10) is linear in $\boldsymbol{z} \in [0,1]^p$ and thus convex. Consequently, c is convex from Lemma 15.

A.4 Proof of Theorem 4

Algorithm 1 converges to an optimal solution if the following conditions are satisfied [2]:

- The optimization problem (7) is feasible and bounded.
- The objective function c is continuously differentiable and convex.

The former condition is clearly satisfied by Assumption 1, and the latter condition is also satisfied because Lemmas 2 and 3 hold under Assumptions 1–3. Because the loss function satisfies Assumptions 1–3, Algorithm 1 converges to the optimal solution in a finite number of iterations.

A.5 Proof of Proposition 5

First, we prove that the function $\ell^{\mathrm{MNL}}(y, \bullet)$ is proper convex for any $y \in [m]$. From the definition, we have $\ell^{\mathrm{MNL}}(y, \boldsymbol{\eta}) = -\log\left[\exp(\eta_y)/\sum_{s=1}^m \exp(\eta_s)\right]$. Because $0 < \exp(v) < +\infty$ for any $v \in \mathbb{R}$, the following inequality holds:

$$0 < \frac{\exp(\eta_y)}{\sum_{s=1}^m \exp(\eta_s)} < 1.$$

Consequently, $\operatorname{dom} \ell^{\mathrm{MNL}}(y, \bullet) = \mathbb{R}^m \neq \emptyset$ and $\ell^{\mathrm{MNL}}(y, \boldsymbol{\eta}) > 0$. The convexity of $\ell^{\mathrm{MNL}}(y, \bullet)$ is discussed in Section 3 in [7].

Next, we consider problem (P) of the MNL model. As discussed above, $0 < \ell^{\mathrm{MNL}}(y, \boldsymbol{\eta}) < +\infty, \forall y \in [m], \boldsymbol{\eta} \in \mathbb{R}^m$. Consequently, (P) is bounded and feasible. Because $\operatorname{dom} \ell^{\mathrm{MNL}}(y, \bullet) = \mathbb{R}^m$ for all $y \in [m]$, (P) has an interior feasible solution.

A.6 Proof of Proposition 6

The continuity is trivial because $v \log v$ is continuous at each $v \in [0, 1]$. We also find that $\hat{\ell}^{\mathrm{MNL}}(y, \bullet)$ is strictly convex for all $y \in [m]$ because its Hessian matrix is always positive definite; this is easily observed from the following equation:

$$\frac{\partial^2 \hat{\ell}^{\mathrm{MNL}}}{\partial \alpha_s \partial \alpha_r}(y, \boldsymbol{\alpha}) = \begin{cases} \alpha_r^{-1} & \text{if } r \neq y \text{ and } s = r, \\ (1 + \alpha_r)^{-1} & \text{if } r = y \text{ and } s = r, \\ 0 & \text{otherwise}, \end{cases} \qquad \forall \boldsymbol{\alpha} \in \mathcal{A}_y^{\mathrm{MNL}}.$$

A.7 Proof of Proposition 7

We have $\alpha_{iy_i} = -\mathbf{1}^\top \boldsymbol{\alpha}_{i\bullet}^{\setminus y_i}$ from constraint (9). Consequently, the feasible region $\mathcal{A}^{\mathrm{MNL}}$ is bounded and closed.

A.8 Proof of Proposition 9

The second equality below is satisfied by the definition of $\hat{g}$:

$$\begin{aligned} g(\eta;\, \boldsymbol{p}) &= \sup\{\eta\alpha - \hat{g}(\alpha;\, \boldsymbol{p}) \mid \alpha \in [0,\, 1]\} \\ &= \sup\{\eta\alpha - (p_3\alpha^2 + p_2\alpha + p_1) \mid \alpha \in [0,\, 1]\}. \end{aligned} \tag{18}$$

We note that the objective function of problem (18) is concave from the assumption $p_3 > 0$. We differentiate the objective function with respect to α and then set it equal to zero as follows:

$$\eta - 2p_3\bar{\alpha} - p_2 = 0,$$

where $\bar{\alpha} \in \mathbb{R}$ is the stationary point. Consequently, the following holds:

$$\bar{\alpha} = \frac{\eta - p_2}{2p_3}. \tag{19}$$

Because the objective function is concave, the optimal value is given at $\alpha = 0$ and $\alpha = 1$ when $\bar{\alpha} < 0$ and $\bar{\alpha} > 1$, respectively.

These intervals can be transformed into the following intervals of η by equation (19):

$$\begin{aligned} \bar{\alpha} < 0 \quad &\Leftrightarrow \quad \eta < p_2, \\ \bar{\alpha} \in [0,\, 1] \quad &\Leftrightarrow \quad \eta \in [p_2,\, p_2 + 2p_3], \\ \bar{\alpha} > 1 \quad &\Leftrightarrow \quad \eta > p_2 + 2p_3. \end{aligned}$$

Consequently, we have the desired result.

A.9 Proof of Proposition 10

The second equality below is satisfied by the definition of ℓ^{Titsias}:

$$\begin{aligned} \hat{\ell}^{\text{Titsias}}(y, \boldsymbol{\alpha}) &= \sup\{\boldsymbol{\alpha}^\top \boldsymbol{\eta} - \ell^{\text{Titsias}}(y,\, \boldsymbol{\eta}) \mid \boldsymbol{\eta} \in \mathbb{R}^m\} \\ &= \sup\{\boldsymbol{\alpha}^\top \boldsymbol{\eta} - \sum_{s \neq y} \log[1 + \exp(\eta_s - \eta_y)] \mid \boldsymbol{\eta} \in \mathbb{R}^m\}. \end{aligned} \tag{20}$$

Because optimization problem (20) has no constraints and has a convex objective function, we obtain an optimal solution by the gradient with respect to $\boldsymbol{\eta}$. Let us consider the following two cases.

Case 1: $r \neq y$ holds. First, we calculate the partial derivative with respect to η_r, and set it equal to zero as follows:

$$\alpha_r - \frac{\exp(\eta_r^\star - \eta_y^\star)}{1 + \exp(\eta_r^\star - \eta_y^\star)} = 0,$$

where $\boldsymbol{\eta}^\star \in \mathbb{R}^m$ is an optimal solution of problem (20). That is,

$$\frac{1}{1 + \exp(\eta_y^\star - \eta_r^\star)} = \alpha_r. \tag{21}$$

From this equation, the following two equations are obtained:

$$\eta_y^\star - \eta_r^\star = \log(1 - \alpha_r) - \log \alpha_r, \tag{22}$$

$$1 + \exp(\eta_r^\star - \eta_y^\star) = (1 - \alpha_r)^{-1}. \tag{23}$$

Note that we assume $0 < \alpha_r < 1$ to derive these equations.

Case 2: $r = y$ holds. Similarly, we calculate the partial derivative with respect to η_r, and set it equal to zero as follows:

$$\alpha_y + \sum_{s \neq y} \left(\frac{1}{1 + \exp(\eta_y^\star - \eta_s^\star)} \right) = 0.$$

Consequently, the following equation is obtained from equation (21):

$$\mathbf{1}^\top \boldsymbol{\alpha} = 0. \tag{24}$$

We obtain the following derivation from equations (22) and (23):

$$\begin{aligned}
\hat{\ell}^{\mathrm{Titsias}}(y, \boldsymbol{\alpha}) &= \boldsymbol{\alpha}^\top \boldsymbol{\eta}^\star - \sum_{s \neq y} \log[1 + \exp(\eta_s^\star - \eta_y^\star)] \\
&= \sum_{s \neq y} \alpha_s \left(\eta_y^\star - \log(1 - \alpha_s) + \log \alpha_s\right) + \alpha_y \eta_y^\star + \sum_{s \neq y} \log[1 - \alpha_s] \\
&= \sum_{s \neq y} (\alpha_s \log \alpha_s + (1 - \alpha_s) \log(1 - \alpha_s)) + \eta_y^\star \sum_{s \in [m]} \alpha_s.
\end{aligned}$$

Consequently, from equation (24), the following equation holds:

$$\hat{\ell}^{\mathrm{Titsias}}(y, \boldsymbol{\alpha}) = \sum_{s \neq y} [\alpha_s \log \alpha_s + (1 - \alpha_s) \log(1 - \alpha_s)].$$

Because $v \log v \to 0$ when $v \to 0$, we have the desired result.

A.10 Proof of Theorem 11

The loss function $\ell^{\mathrm{quad}} : [m] \times \mathbb{R}^m \to \mathbb{R}$ is defined by

$$\ell^{\mathrm{quad}}(y, \boldsymbol{\eta}) = \sum_{s \neq y} g(\eta_s - \eta_y; \boldsymbol{p}),$$

where

$$g(\eta; \boldsymbol{p}) = \begin{cases} -p_1 & \text{if } \eta < p_2, \\ (\eta - p_2)^2 / 4p_3 - p_1 & \text{if } \eta \in [p_2, p_2 + 2p_3], \\ \eta - (p_1 + p_2 + p_3) & \text{otherwise.} \end{cases}$$

Moreover, for each $y \in [m]$, let $\hat{\ell}^{\mathrm{quad}}(y, \bullet)$ be the conjugate defined as follows:

$$\hat{\ell}^{\mathrm{quad}}(y, \boldsymbol{\alpha}) = \begin{cases} \sum_{s \neq y} p_3 \alpha_s^2 + p_2 \alpha_s + p_1 & \text{if } \boldsymbol{\alpha} \in \mathcal{A}_y^{\mathrm{quad}}, \\ +\infty & \text{otherwise,} \end{cases}$$

where

$$\mathcal{A}_y^{\mathrm{quad}} = \{\boldsymbol{\alpha} \in \mathbb{R}^m \mid \mathbf{1}^\top \boldsymbol{\alpha} = 0, \mathbf{0} \leq \boldsymbol{\alpha}^{\setminus y} \leq \mathbf{1}\}.$$

We show that these two functions and problem (P) satisfy Assumptions 1–3.

First, for each $y \in [m]$, $\ell^{\mathrm{quad}}(y, \boldsymbol{\alpha})$ is the summation of quadratic functions; $\ell^{\mathrm{quad}}(y, \boldsymbol{\alpha})$ is thus strictly convex in $\boldsymbol{\alpha}$ if and only if $p_3 > 0$. Consequently Assumption 2 is satisfied when $p_3 > 0$. Second, we have $\mathbf{0} \leq \boldsymbol{\alpha}^{\setminus y} \leq \mathbf{1}$ and $\alpha_y = -\mathbf{1}^\top \boldsymbol{\alpha}^{\setminus y}$ from the definition of $\mathcal{A}_y^{\mathrm{quad}}$. Consequently, $\mathcal{A}_y^{\mathrm{quad}}$ is bounded and closed; that is, Assumption 3 is satisfied. Finally, the function ℓ^{quad} is bounded below because $g(\eta; \boldsymbol{p}) \geq -p_1$ when $p_3 > 0$. Because the conjugate of a closed proper convex function is still closed proper convex, $\ell^{\mathrm{quad}}(y, \bullet)$ is a proper convex function. As with the proof of Theorem 5, (P) is bounded and has an interior feasible solution. Consequently, Assumption 1 is satisfied.

From the above discussion, all the assumptions of Theorem 4 are satisfied, and we have the desired result.

References

[1] D. Bertsimas, J. Pauphilet, and B. V. Parys. Sparse classification and phase transitions: A discrete optimization perspective. arXiv preprint arXiv:1710.01352, 2017.

[2] P. Bonami, L. Biegler, A. Conn, G. Cornuéjols, I. Grossmann, C. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya, and A. Wächter. An algorithmic framework for convex mixed integer nonlinear programs. Discrete Optimization 5(2):186–204, 2008.

[3] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge University Press, 2004.

[4] S. Boyd and L. Vandenberghe. Subgradients. Lecture notes for EE364b, Stanford University, Winter Quarter 2006–2007, 2008.

[5] M. Fukushima. Basic nonlinear optimization (in Japanese). Asakura Publishing, 2001.

[6] T. Kanamori, T. Suzuki, I. Takeuchi, and I. Sato. Continuous optimization for machine learning (in Japanese). Kodansha Scientific, 2016.

[7] J. D. M. Rennie. Regularized logistic regression is strictly convex. Technical report, MIT, [http://qwone.com/~jason/writing/convexLR.pdf], 2005. Last accessed: 2018-09-05.