A Key assumptions and additional lemmas

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

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

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad a_i^T x - b_i = 0, \\
& \quad \forall i \in [m].
\end{align*}
\]

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

\[
\begin{align*}
\text{maximize} & \quad \min_{\lambda \in \mathbb{R}^m} \{ \mathcal{L}(x, \lambda) \mid x \in \mathbb{R}^n \}, \\
\text{minimize} & \quad \max_{\lambda \in \mathbb{R}^n} \{ \mathcal{L}(x, \lambda) \mid \lambda \in \mathbb{R}^m \},
\end{align*}
\]

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

\[
\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i (a_i^T x - b_i).
\]

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

\[
f(x) = \sup_{\alpha \in A} f_\alpha(x).
\]

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

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

where $\text{conv}$ denotes the convex hull of a set.
Lemma 14. Given $U \subseteq \mathbb{R}^p$, let $f : \mathbb{R}^n \times U \rightarrow \mathbb{R}$ be the objective function of the optimization problem

$$\max_{x \in S(u)} f(x, u),$$

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

$$\Phi(u) = \arg\max_x \{f(x, u) \mid x \in S(u)\}.$$

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 $f : \mathbb{R}^n \times \mathcal{A} \rightarrow [-\infty, +\infty]$. Let $g : \mathbb{R}^n \rightarrow [-\infty, +\infty]$ be the function defined by

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

If $f(\cdot, y)$ is convex for each $y \in \mathcal{A}$, $g$ 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:

$$\min_{w \in \mathbb{R}^{m \times p}, b \in \mathbb{R}^m, \eta \in \mathbb{R}^{n \times m}} \max_{\alpha \in \mathbb{R}^{n \times m}} \sum_{i=1}^n \ell(y_i, \eta_i) + \frac{1}{2\gamma} \sum_{r=1}^m \|w_r\|^2 + \sum_{i=1}^n \sum_{r=1}^m \alpha_r(w_r^\top x_i + b_r - \eta_i).$$

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

$$\max_{\alpha \in \mathbb{R}^{n \times m}} \sum_{i=1}^n \min \left\{ \ell(y_i, \eta_i) - \alpha_i^\top \eta_i \mid \eta_i \in \mathbb{R}^m \right\}$$

$$+ \sum_{r=1}^m \min \left\{ b_r 1^\top \alpha_r \mid b_r \in \mathbb{R} \right\}$$

$$+ \sum_{r=1}^m \min \left\{ \frac{1}{2\gamma} \|w_r\|^2 + w_r^\top X^\top \alpha_r \mid w_r \in \mathbb{R}^p \right\}. \tag{17}$$

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

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

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

### A.2 Proof of Lemma 2

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

$$\partial c(z) = \text{conv} \left\{ \left( -\frac{\gamma}{2} \left\| \alpha_r^\top x_j \right\|^2 \right)_{j \in [p]} \mid 1^\top \alpha_r = 0, \forall r \in [m], \alpha_r \in \text{dom} \ \hat{\ell}(y, \cdot), \forall i \in [n], f_2(\alpha) = c(z) \right\}.$$
where $f_z$ is the objective function (10). Moreover, $f_z$ is strictly convex because of $XX^\top \succeq O$ and the strict convexity of $\ell(y, \cdot)$ from Assumption 2. That is, the map $\alpha^*$ is a monomorphism. We thus obtain the partial derivatives of $z$ as (11).

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

$$A = \{\alpha \in \mathbb{R}^{n \times m} \mid 1^T \alpha_r = 0, \forall r \in [m], \alpha_i \in \text{dom } \ell(y_i, \cdot), \forall i \in [n]\}$$

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

From the above discussion, $c$ is continuously differentiable.

### A.3 Proof of Lemma 3

For each $\alpha \in A$, the objective function (10) is linear in $z \in [0, 1]^p$ 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 $\ell_{\text{MNL}}(y, \cdot)$ is proper convex for any $y \in [m]$. From the definition, we have $\ell_{\text{MNL}}(y, \eta) = -\log \left[\exp(\eta) / \sum_{s=1}^m \exp(\eta_s)\right]$. 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, dom $\ell_{\text{MNL}}(y, \cdot) = \mathbb{R}^m \neq \emptyset$ and $\ell_{\text{MNL}}(y, \eta) > 0$. The convexity of $\ell_{\text{MNL}}(y, \cdot)$ is discussed in Section 3 in [7].

Next, we consider problem (P) of the MNL model. As discussed above, $0 < \ell_{\text{MNL}}(y, \eta) < +\infty$, $\forall y \in [m], \eta \in \mathbb{R}^m$. Consequently, (P) is bounded and feasible. Because dom $\ell_{\text{MNL}}(y, \cdot) = \mathbb{R}^m$ 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 $\ell_{\text{MNL}}(y, \cdot)$ is strictly convex for all $y \in [m]$ because its Hessian matrix is always positive definite; this is easily observed from the following equation:

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

### A.7 Proof of Proposition 7

We have $\alpha_{iy} = -1^T \alpha_{iy}^{\top}$ from constraint (9). Consequently, the feasible region $\mathcal{A}_{\text{MNL}}$ is bounded and closed.
A.8 Proof of Proposition 9

The second equality below is satisfied by the definition of $\hat{g}$:
\[
g(\eta; \mathbf{p}) = \sup \{ \eta \alpha - \hat{g}(\alpha; \mathbf{p}) : \alpha \in [0, 1]\} = \sup \{ \eta \alpha - (p_1 \alpha^2 + p_2 \alpha + p_3) : \alpha \in [0, 1]\}. \tag{18}
\]
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 $\alpha$ and then set it equal to zero as follows:
\[
\eta - 2p_3\hat{\alpha} - p_2 = 0,
\]
where $\hat{\alpha} \in \mathbb{R}$ is the stationary point. Consequently, the following holds:
\[
\hat{\alpha} = \frac{\eta - p_2}{2p_3}. \tag{19}
\]
Because the objective function is concave, the optimal value is given at $\alpha = 0$ and $\alpha = 1$ when $\hat{\alpha} < 0$ and $\hat{\alpha} > 1$, respectively.

These intervals can be transformed into the following intervals of $\eta$ by equation (19):
\[
\begin{align*}
\hat{\alpha} < 0 & \quad \Leftrightarrow \quad \eta < p_2, \\
\hat{\alpha} \in [0, 1] & \quad \Leftrightarrow \quad \eta \in [p_2, p_2 + 2p_3], \\
\hat{\alpha} > 1 & \quad \Leftrightarrow \quad \eta > p_2 + 2p_3.
\end{align*}
\]
Consequently, we have the desired result.

A.9 Proof of Proposition 10

The second equality below is satisfied by the definition of $\ell_{\text{Titsias}}$:
\[
\ell_{\text{Titsias}}(y, \boldsymbol{\alpha}) = \sup \{ \alpha^\top \eta - \ell_{\text{Titsias}}(y, \eta) : \eta \in \mathbb{R}^m \} = \sup \{ \alpha^\top \eta - \sum_{s \neq y} \log(1 + \exp(\eta_s - \eta_y)) : \eta \in \mathbb{R}^m \}. \tag{20}
\]
Because optimization problem (20) has no constraints and has a convex objective function, we obtain an optimal solution by the gradient with respect to $\eta$. Let us consider the following two cases.

Case 1: $r \neq y$ holds. First, we calculate the partial derivative with respect to $\eta_r$, and set it equal to zero as follows:
\[
\alpha_r - \frac{\exp(\eta_y^* - \eta_r^*)}{1 + \exp(\eta_y^* - \eta_r^*)} = 0,
\]
where $\eta^* \in \mathbb{R}^m$ is an optimal solution of problem (20). That is,
\[
\frac{1}{1 + \exp(\eta_y^* - \eta_r^*)} = \alpha_r. \tag{21}
\]
From this equation, the following two equations are obtained:
\[
\begin{align*}
\eta_y^* - \eta_r^* &= \log(1 - \alpha_r) - \log \alpha_r, \tag{22} \\
1 + \exp(\eta_y^* - \eta_r^*) &= (1 - \alpha_r)^{-1}. \tag{23}
\end{align*}
\]
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 $\eta_r$, and set it equal to zero as follows:
\[
\alpha_y + \sum_{s \neq y} \left( \frac{1}{1 + \exp(\eta_y^* - \eta_s^*)} \right) = 0.
\]
Consequently, the following equation is obtained from equation (21):

$$1^\top \alpha = 0.$$  \hfill (24)

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

$$\ell^{\text{Titsias}}_{\text{quad}}(y, \alpha) = \alpha^\top \eta - \sum_{s \neq y} \log[1 + \exp(\eta_s^* - \eta_y^*)]$$

$$= \sum_{s \neq y} \alpha_s (\eta_y^* - \log(1 - \alpha_s) + \log \alpha_s) + \alpha_y \eta_y^* + \sum_{s \neq y} \log[1 - \alpha_s]$$

$$= \sum_{s \neq y} (\alpha_s \log \alpha_s + (1 - \alpha_s) \log(1 - \alpha_s)) + \eta_y^* \sum_{s \in [m]} \alpha_s.$$

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

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

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

### A.10 Proof of Theorem 11

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

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

where

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

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

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

where

$$\mathcal{A}^{\text{quad}}_y = \{\alpha \in \mathbb{R}^m \mid 1^\top \alpha = 0, \ 0 \leq \alpha^\top 1 \leq 1\}.$$  

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

First, for each $y \in [m]$, $\ell^{\text{quad}}(y, \alpha)$ is the summation of quadratic functions; $\ell^{\text{quad}}(y, \alpha)$ is thus strictly convex in $\alpha$ if and only if $p_3 > 0$. Consequently Assumption 2 is satisfied when $p_3 > 0$. Second, we have $0 \leq \alpha^\top 1 \leq 1$ and $\alpha_y = -1^\top \alpha^\top 1$ from the definition of $\mathcal{A}^{\text{quad}}_y$. Consequently, $\mathcal{A}^{\text{quad}}_y$ is bounded and closed; that is, Assumption 3 is satisfied. Finally, the function $\ell^{\text{quad}}$ is bounded below because $g(\eta; p) \geq -p_1$ when $p_3 > 0$. Because the conjugate of a closed proper convex function is still closed proper convex, $\ell^{\text{quad}}_{\text{quad}}(y, \cdot)$ 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


