Supplementary Material

0.1 Proof of Theorem 1

In order to prove Theorem 1, we compute the Hessian matrix of f. We begin by evaluating the gradient of f.

Lemma 0.1. *Let $\widehat{P}(c|\mathbf{x})$ be defined as in Equation 3 in the paper, where $\pi_c, \theta_{c,j,\mathbf{x}_i} \in (0,1)$. Then*

$$\frac{\partial \widehat{P}(c'|\mathbf{x}_i)}{\partial w_{c,k}} = \begin{cases} \widehat{P}(c'|\mathbf{x}_i)(1-\widehat{P}(c'|\mathbf{x}_i)) \log \theta_{c',k,\mathbf{x}_i} & \text{if } c = c' \\ -\widehat{P}(c'|\mathbf{x}_i)\widehat{P}(c|\mathbf{x}_i) \log \theta_{c,k,\mathbf{x}_i} & \text{if } c \neq c'. \end{cases} \tag{1}$$

Proof. We first compute the derivative of $\widehat{P}(c'|\mathbf{x}_i)$ with respect to $w_{c',k}$. We have

$$\begin{aligned} \frac{\partial \widehat{P}(c'|\mathbf{x}_i)}{\partial w_{c',k}} &= \frac{\pi_{c'} \prod_{j=1}^{m} \theta_{c',j,\mathbf{x}_i}^{w_{c',j}} \log \theta_{c',k,\mathbf{x}_i} \left(\sum_{c''=1}^{l} \pi_{c''} \prod_{j=1}^{m} \theta_{c'',j,\mathbf{x}_i}^{w_{c'',j}} \right)}{\left(\sum_{c''=1}^{l} \pi_{c''} \prod_{j=1}^{m} \theta_{c'',j,\mathbf{x}_i}^{w_{c'',j}} \right)^2} \\ &\quad - \frac{\pi_{c'} \prod_{j=1}^{m} \theta_{c',j,\mathbf{x}_i}^{w_{c',j}} \cdot \pi_{c'} \prod_{j=1}^{m} \theta_{c',j,\mathbf{x}_i}^{w_{c',j}} \log \theta_{c',k,\mathbf{x}_i}}{\left(\sum_{c''=1}^{l} \pi_{c''} \prod_{j=1}^{m} \theta_{c'',j,\mathbf{x}_i}^{w_{c'',j}} \right)^2} \\ &= \widehat{P}(c'|\mathbf{x}_i)(1-\widehat{P}(c'|\mathbf{x}_i)) \log \theta_{c',k,\mathbf{x}_i}. \end{aligned} \tag{2}$$

Next, if $c' \neq c$, we obtain

$$\begin{aligned} \frac{\partial \widehat{P}(c'|\mathbf{x}_i)}{\partial w_{c,k}} &= -\frac{\pi_c' \prod_{j=1}^{m} \theta_{c',j,\mathbf{x}_i}^{w_{c',j}} \pi_c \prod_{j=1}^{m} \theta_{c,j,\mathbf{x}_i}^{w_{c,j}} \log \theta_{c,k,\mathbf{x}_i}}{\left(\sum_{c''=1}^{l} \pi_{c''} \prod_{j=1}^{m} \theta_{c'',j,\mathbf{x}_i}^{w_{c'',j}} \right)^2} \\ &= -\widehat{P}(c'|\mathbf{x}_i)\widehat{P}(c|\mathbf{x}_i) \log \theta_{c,k,\mathbf{x}_i}. \end{aligned} \tag{3}$$

Given a vector field $F = (F_1, F_2, \ldots, F_m)^T$ where $F_k = F_k(x_1, x_2, \ldots, x_m)$, we denote by ∇F the $l \times m$ matrix with (i,j)-th entry given by $(\nabla F)_{ij} = \frac{\partial F_j}{\partial x_i}$. In particular, if $f : \mathbb{R}^n \to \mathbb{R}$ is a function, then ∇f is the usual gradient of f and $\nabla\nabla f = \nabla^2 f$ is the usual Hessian matrix of f.

Let $\mathbf{e}_{1,1}, \ldots, \mathbf{e}_{1,m}, \mathbf{e}_{2,1}, \ldots, \mathbf{e}_{2,m}, \ldots, \mathbf{e}_{l,1}, \ldots, \mathbf{e}_{l,m}$ denote the canonical basis of $\mathbb{R}^{l\times m}$. The following reformulation of Lemma 0.1 provides an expression for $\nabla \widehat{P}(c'|\mathbf{x}_i)$.

Lemma 0.2. *We have*

$$\nabla \widehat{P}(c'|\mathbf{x}_i) = \widehat{P}(c'|\mathbf{x}_i) \sum_{k=1}^{m} \log \theta_{c',k,\mathbf{x}_i} \mathbf{e}_{c',k} - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \sum_{k=1}^{m} \widehat{P}(c|\mathbf{x}_i) \log \theta_{c,k,\mathbf{x}_i} \mathbf{e}_{c,k}. \tag{4}$$

We now evaluate the Hessian of the log-likelihood function. To simplify the notation, let

$$v_{c,i} = \sum_{k=1}^{m} \log \theta_{c,k,\mathbf{x}_i} \mathbf{e}_{c,k}, \qquad v_i = \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \qquad (c = 1, \ldots, l). \tag{5}$$

Lemma 0.3. *With the same notation as above, we have*

$$\nabla^2 \widehat{P}(c'|\mathbf{x}_i) = \frac{1}{\widehat{P}(c'|\mathbf{x}_i)} \nabla \widehat{P}(c'|\mathbf{x}_i) \nabla \widehat{P}(c'|\mathbf{x}_i)^T + \widehat{P}(c'|\mathbf{x}_i) H_i, \tag{6}$$

where

$$H_i = v_i v_i^T - \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} v_{c,i}^T \tag{7}$$

is a matrix that depends only on i.

Proof. Using Lemma 0.2 and the notation defined in Equation (5), we obtain

$$\begin{aligned}
\nabla^2 \widehat{P}(c'|\mathbf{x}_i) &= \nabla \left(\widehat{P}(c'|\mathbf{x}_i) v_{c',i} - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right) \\
&= \nabla \widehat{P}(c'|\mathbf{x}_i) v_{c',i}^T - \nabla \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i}^T - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \nabla \widehat{P}(c|\mathbf{x}_i) v_{c,i}^T \\
&= \left(\widehat{P}(c'|\mathbf{x}_i) v_{c',i} - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right) v_{c',i}^T - \\
&\quad \left(\widehat{P}(c'|\mathbf{x}_i) v_{c',i} - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i}^T \\
&\quad - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \left(\widehat{P}(c|\mathbf{x}_i) v_{c,i} - \widehat{P}(c|\mathbf{x}_i) \sum_{c''=1}^{l} \widehat{P}(c''|\mathbf{x}_i) v_{c'',i} \right) v_{c,i}^T \\
&= \widehat{P}(c'|\mathbf{x}_i) v_{c',i} v_{c',i}^T - \widehat{P}(c'|\mathbf{x}_i) \left(\sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right) v_{c',i}^T - \widehat{P}(c'|\mathbf{x}_i) v_{c',i} \left(\sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right)^T \\
&\quad + \widehat{P}(c'|\mathbf{x}_i) \left(\sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right) \left(\sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right)^T \\
&\quad - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} v_{c,i}^T + \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) \left(\sum_{c''=1}^{l} \widehat{P}(c''|\mathbf{x}_i) v_{c'',i} \right) v_{c,i}^T \\
&= \widehat{P}(c'|\mathbf{x}_i) \left(v_{c',i} - \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right) \left(v_{c',i} - \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right)^T \\
&\quad - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} v_{c,i}^T + \widehat{P}(c'|\mathbf{x}_i) \left(\sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right) \left(\sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right)^T \\
&= \frac{1}{\widehat{P}(c'|\mathbf{x}_i)} \nabla \widehat{P}(c'|\mathbf{x}_i) \nabla \widehat{P}(c'|\mathbf{x}_i)^T + \widehat{P}(c'|\mathbf{x}_i) \left(\sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right) \left(\sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} \right)^T \\
&\quad - \widehat{P}(c'|\mathbf{x}_i) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} v_{c,i}^T \\
&= \frac{1}{\widehat{P}(c'|\mathbf{x}_i)} \nabla \widehat{P}(c'|\mathbf{x}_i) \nabla \widehat{P}(c'|\mathbf{x}_i)^T + \widehat{P}(c'|\mathbf{x}_i) H_i.
\end{aligned}$$

We claim that the matrices H_i are negative semidefinite. More generally, we have the following.

Lemma 0.4. *Let $n, d \in \mathbb{N}$. For $i = 1, \dots, n$, let $z_i \in \mathbb{R}^d$ and let $0 \le p_i \le 1$ with $\sum_{i=1}^n p_i = 1$. Then*

$$\left(\sum_{i=1}^n p_i z_i\right)\left(\sum_{i=1}^n p_i z_i\right)^T \le \sum_{i=1}^n p_i z_i z_i^T, \tag{8}$$

where the inequality "$\le$" is with respect to the semidefinite ordering.

Proof. Let $\langle \cdot, \cdot \rangle$ denote the standard inner product on $\mathbb{R}^d$ and let $x \in \mathbb{R}^d$. Using Jensen's inequality, we obtain

$$\begin{aligned} x^T \left(\sum_{i=1}^n p_i z_i\right)\left(\sum_{i=1}^n p_i z_i\right)^T x = \left\langle \sum_{i=1}^n p_i z_i, x \right\rangle^2 &= \left(\sum_{i=1}^n p_i \langle z_i, x\rangle\right)^2 \\ \le \sum_{i=1}^n p_i \langle z_i, x\rangle^2 &= x^T \left(\sum_{i=1}^n p_i z_i z_i^T\right) x. \end{aligned} \tag{9}$$

This proves $\sum_{i=1}^n p_i z_i z_i^T - \left(\sum_{i=1}^n p_i z_i\right)\left(\sum_{i=1}^n p_i z_i\right)^T$ is positive semidefinite.

Corollary 1. *The matrices H_i in Equation (7) are negative semidefinite.*

Lemma 0.5. *With the same notation as in Equations (5) and (7), we have*

$$\nabla f(\mathbf{W}) = -\sum_{i=1}^n \left(v_{c_i,i} - v_i\right), \qquad \nabla^2 f(\mathbf{W}) = -\sum_{i=1}^n H_i. \tag{10}$$

Proof. First, using Lemma 0.2, we have

$$\begin{aligned} \nabla f(\mathbf{W}) &= -\sum_{i=1}^n \frac{\nabla \widehat{P}(c_i|\mathbf{x}_i)}{\widehat{P}(c_i|\mathbf{x}_i)} = -\sum_{i=1}^n \left(\sum_{k=1}^m \log \theta_{c_i,k,\mathbf{x}_i} \mathbf{e}_{c_i,k} - \sum_{c=1}^l \sum_{k=1}^m \widehat{P}(c|\mathbf{x}_i) \log \theta_{c,k,\mathbf{x}_i} \mathbf{e}_{c,k}\right) \\ &= -\sum_{i=1}^n \left(v_{c_i,i} - v_i\right). \end{aligned}$$

Next, recall that for any function $g : \mathbb{R}^N \to (0, \infty)$, we have

$$\nabla^2 \log g = \frac{1}{g}\nabla^2 g - \frac{1}{g^2}\nabla g \nabla g^T.$$

Thus, using Lemmas 0.2 and 0.3, we obtain

$$\begin{aligned} \nabla^2 \log \widehat{P}(c_i|\mathbf{x}_i) &= \frac{1}{\widehat{P}(c_i|\mathbf{x}_i)} \nabla^2 \widehat{P}(c_i|\mathbf{x}_i) - \frac{1}{\widehat{P}(c_i|\mathbf{x}_i)^2} \nabla \widehat{P}(c_i|\mathbf{x}_i) \nabla \widehat{P}(c_i|\mathbf{x}_i)^T \\ &= \frac{1}{\widehat{P}(c_i|\mathbf{x}_i)^2} \nabla \widehat{P}(c_i|\mathbf{x}_i) \nabla \widehat{P}(c_i|\mathbf{x}_i)^T + H_i - \frac{1}{\widehat{P}(c_i|\mathbf{x}_i)^2} \nabla \widehat{P}(c_i|\mathbf{x}_i) \nabla \widehat{P}(c_i|\mathbf{x}_i)^T \\ &= H_i. \end{aligned}$$

It follows immediately that

$$\nabla^2 f(\mathbf{W}) = -\sum_{i=1}^n \nabla^2 \log \widehat{P}(c_i|\mathbf{x}_i) = -\sum_{i=1}^n H_i.$$

Combining the above lemmas, we immediately obtain Theorem 1.

Proof (Proof of Theorem 1). By Lemma 0.5, we have $\nabla^2 f(\mathbf{W}) = -\sum_{i=1}^n H_i$. Since each H_i is negative semidefinite (Corollary 1), it follows that $\nabla^2 f$ is positive semidefinite and so f is convex.

0.2 Proof of Lemma 1

Proof. By Theorem 1,

$$\|\nabla^2 f(\mathbf{W})\|_2 = \left\| - \sum_{i=1}^{n} H_i \right\|_2 \leq \sum_{i=1}^{n} \|H_i\|_2.$$

Next, using the fact that H_i is negative semidefinite (Theorem 1), we have

$$\|H_i\|_2 = \left\| v_i v_i^T - \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} v_{c,i}^T \right\|_2 \leq \left\| \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) v_{c,i} v_{c,i}^T \right\|_2 \tag{11}$$

$$\leq \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) \|v_{c,i}\|_2^2 \leq \left(\max_{c=1,\ldots,l} \|v_{c,i}\|_2^2 \right) \sum_{c=1}^{l} \widehat{P}(c|\mathbf{x}_i) \tag{12}$$

$$= \max_{c=1,\ldots,l} \|v_{c,i}\|_2^2. \tag{13}$$

The inequality for $\nabla^2 f(\mathbf{W})$ follows by combining the above inequalities. Adding the ℓ_2 penalty, we obtain $\|\nabla g(\mathbf{W})\|_2 \leq L + 2\rho_2$ and the result follows (see e.g.[1, Theorem 5.12]).

0.3 Proof of Lemma 2

Proof. Using Lemma 0.5, we have $\nabla^2 g(\mathbf{W}) = -\sum_{i=1}^{n} H_i + 2\rho_2 I$. Hence $\lambda_{\min}(\nabla^2 g(\mathbf{W}) \geq \lambda_{\min}(-\sum_{i=1}^{n} H_i)) + 2\rho_2 = \lambda + 2\rho_2$. For the upper bound, using Lemma 1, section 0.2, we have $\lambda_{\max}(\nabla^2 g(\mathbf{W})) = \|\nabla^2 g(\mathbf{W})\|_2 \leq L + 2\rho_2$.

0.4 Semilog Convergence Results

Here we provide plots for all datasets used for experimentation.

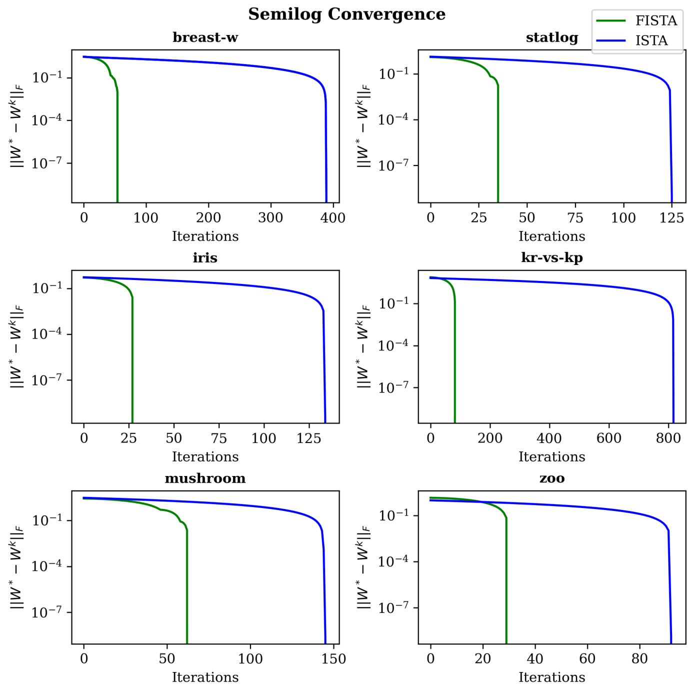

Fig. 1. ISTA v.s. FISTA to solve `BARISTA` for the breast-w, statlog, iris, kr-vs-kp, mushroom, zoo datasets.

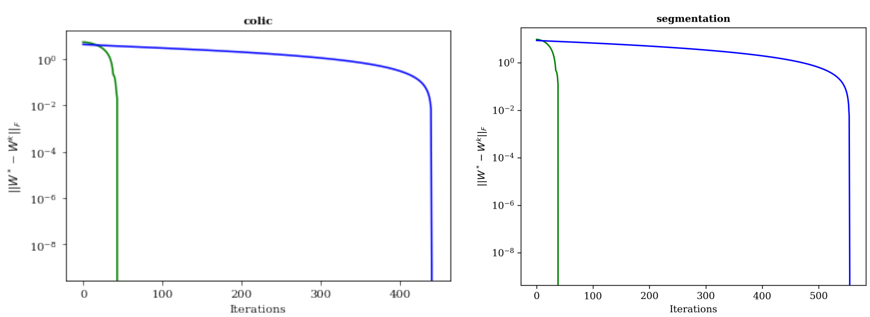

Fig. 2. ISTA v.s. FISTA to solve `BARISTA` for the colic and segmentation datasets.

Once again, FISTA is clearly outperforming ISTA for every dataset, verifying the results given by Beck and Teboulle in [2].