Labs
Machine Learning Course
Fall 2023

EPFL
School of Computer and Communication Sciences
Nicolas Flammarion & Martin Jaggi
www.epfl.ch/labs/mlo/machine-learning-cs-433

Problem Set 6, Oct 26, 2023 (Theory Questions Part)

2. Support Vector Machines using Coordinate Descent

1. The dual objective function that we have to optimise is the following :

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}}{\text{maximize}} & f(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2\lambda}\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{\alpha} \\ \text{subject to} & \boldsymbol{\alpha} \in [0,1]^N \end{array}$$

where $\boldsymbol{Q} := \frac{1}{N}\text{diag}(\boldsymbol{y})\boldsymbol{X}\boldsymbol{X}^\top\text{diag}(\boldsymbol{y})$. Note that we have omitted the $\frac{1}{N}$ constant in the dual objective function for brevity, as it does not change the maximization problem. For computing coordinate update for one coordinate n, consider the following one variable sub-problem:

$$\begin{array}{ll} \underset{\gamma\in\mathbb{R}}{\text{maximize}} & f(\boldsymbol{\alpha} + \gamma \boldsymbol{e}_n) \\ \text{subject to} & 0 \le \alpha_n + \gamma \le 1 \end{array}$$

where $\boldsymbol{e}_n = [0, \cdots, 1, \cdots, 0]^\top$ (all zero vector except at the n^{th} position). We can explicitly develop $f(\boldsymbol{\alpha} + \gamma \boldsymbol{e}_n)$ as a polynomial in γ, indeed:

$$\begin{aligned} f(\boldsymbol{\alpha} + \gamma \boldsymbol{e}_n) &= (\boldsymbol{\alpha} + \gamma \boldsymbol{e}_n)^\top \mathbf{1} - \tfrac{1}{2\lambda}(\boldsymbol{\alpha} + \gamma \boldsymbol{e}_n)^\top \boldsymbol{Q}(\boldsymbol{\alpha} + \gamma \boldsymbol{e}_n) \\ &= \boldsymbol{\alpha}^\top \mathbf{1} - \tfrac{1}{2\lambda}\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{\alpha} + \gamma - \tfrac{1}{2\lambda}(\gamma^2 \boldsymbol{e}_n^\top \boldsymbol{Q}\boldsymbol{e}_n + \gamma\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{e}_n + \gamma\boldsymbol{e}_n^\top \boldsymbol{Q}\boldsymbol{\alpha}) \\ &= f(\boldsymbol{\alpha}) - \tfrac{\gamma^2}{2\lambda}\boldsymbol{Q}_{nn} + \gamma(1 - \tfrac{1}{\lambda}\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{e}_n) \end{aligned}$$

Hence we recognize a concave second degree polynomial, which has a unique maximum over $\mathbb{R}$. Differentiating with respect to γ and setting to 0 to obtain its maximum over $\mathbb{R}$, we get :

$$\begin{aligned} -\tfrac{\gamma^\star}{\lambda}\boldsymbol{Q}_{nn} + (1 - \tfrac{1}{\lambda}\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{e}_n) &= 0 \\ \gamma^\star &= \frac{\lambda}{\boldsymbol{Q}_{nn}}(1 - \tfrac{1}{\lambda}\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{e}_n) \end{aligned}$$

Note that $\boldsymbol{Q}_{nn} = \frac{1}{N}\boldsymbol{x}_n^\top \boldsymbol{x}_n y_n^2 = \frac{1}{N}\boldsymbol{x}_n^\top \boldsymbol{x}_n$ and $\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{e}_n = \sum_{i=1}^N \alpha_i \boldsymbol{Q}_{i,n} = \frac{1}{N}\sum_{i=1}^N \alpha_i y_i \boldsymbol{x}_i^\top \boldsymbol{x}_n y_n$. Using $\boldsymbol{w}(\boldsymbol{\alpha}) = \frac{1}{\lambda N}\sum_{i=1}^N \alpha_i y_i \boldsymbol{x}_i$, we get $\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{e}_n = \lambda y_n \boldsymbol{w}^\top \boldsymbol{x}_n$ and thus

$$\gamma^\star = \frac{\lambda N}{\boldsymbol{x}_n^\top \boldsymbol{x}_n}(1 - y_n \boldsymbol{w}^\top \boldsymbol{x}_n)$$

We conclude

$$\begin{aligned} \alpha_n^{\text{new}} &= \alpha_n + \gamma^\star \\ &= \alpha_n + \frac{\lambda N}{\boldsymbol{x}_n^\top \boldsymbol{x}_n}(1 - y_n \boldsymbol{w}^\top \boldsymbol{x}_n) \end{aligned}$$

In the previous equation it *seems* as if α_n^{new} depends on α_n, which should not be the case, you can check that it is indeed not the case. Since we have a constraint $\boldsymbol{\alpha} \in [0,1]^N$ and we know that function f is quadratic with respect to α_n, the optimal α_n is the projection of α_n^{new} onto the set $[0,1]^N$:

$$\alpha_n^{\text{new}} := \min\Big\{ \max\Big\{\alpha_n + \frac{\lambda N}{\boldsymbol{x}_n^\top \boldsymbol{x}_n}(1 - y_n \boldsymbol{w}^\top \boldsymbol{x}_n), 0\Big\}, 1\Big\}$$

2. In lecture we seen the relation between primal and dual variables that is

$$\boldsymbol{w}(\boldsymbol{\alpha}) = \frac{1}{\lambda N}\sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n$$

   Assume that for current dual variable $\boldsymbol{\alpha}$ the corresponding $\boldsymbol{w}(\boldsymbol{\alpha})$ is known. After a coordinate update over a coordinate n_1 the dual variable changes as $\boldsymbol{\alpha}^{\text{new}} = \boldsymbol{\alpha} + \gamma^* \boldsymbol{e}_{n_1}$, and then the primal variable changes as

$$\boldsymbol{w}^{\text{new}}(\boldsymbol{\alpha}^{\text{new}}) = \frac{1}{\lambda N}\sum_{n=1}^{N} \alpha_n^{\text{new}} y_n \boldsymbol{x}_n = \frac{1}{\lambda N}\sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n + \frac{1}{\lambda N}\gamma^* y_{n_1}\boldsymbol{x}_{n_1} = \boldsymbol{w} + \frac{1}{\lambda N}\gamma^* y_{n_1}\boldsymbol{x}_{n_1}$$

4. We see that for the dataset used in this exercise, optimizing with coordinate descent is faster than using SGD. We can also see that the duality gap goes to zero.

   In general practical performance of SGD for SVM with well tuned stepsize is identical to dual Coordinate Descent, however, the advantage of CD is that no stepsize tuning is needed.

Additional Theory Exercises

Convexity

1. We need to check that

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$$

   for all $x, y \in \mathbb{R}$ and $\theta \in [0,1]$. Since the function is linear, we get an equality

$$a(\theta x + (1-\theta)y) + b = \theta\,(ax+b) + (1-\theta)\,(ay+b)$$

2. For any elements x, y in the common fixed domain we have that

$$\begin{aligned} g(\theta x + (1-\theta)y)) &= \sum_i f_i(\theta x + (1-\theta)y) \\ &\le \sum_i [\theta f_i(x) + (1-\theta) f_i(y)] \\ &= \theta \sum_i f_i(x) + (1-\theta)\sum_i f_i(y) \\ &= \theta g(x) + (1-\theta) g(y). \end{aligned}$$

3. Using convexity of f, we know that

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)\,.$$

   Further since g is increasing, we can apply g on both sides of the above equation to get

$$g(f(\theta x + (1-\theta)y)) \le g(\theta f(x) + (1-\theta) f(y))\,.$$

   Finally, using the convexity of g we get

$$\begin{aligned} g(f(\theta x + (1-\theta)y)) &\le g(\theta f(x) + (1-\theta) f(y)) \\ &\le \theta g(f(x)) + (1-\theta) g(f(y))\,. \end{aligned}$$

4. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two elements in the domain. Let $x = \boldsymbol{w}^\top \boldsymbol{x} + b$ and $y = \boldsymbol{w}^\top \boldsymbol{y} + b$. Let $\theta \in [0,1]$. We need to show that

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y),$$

   which follows since by assumption f was convex.

5. Assume that it has two global minima at $x^\star$ and $y^\star$. Let $z^\star = (x^\star + y^\star)/2$. Then, since f is strictly convex, we have $f(z^\star) < \frac{1}{2}(f(x^\star) + f(y^\star)) = f(x^\star) = f(y^\star)$, which means neither points $x^\star$ and $y^\star$ are global minima. This contradicts the initial assumption and proves that a strictly convex function has a unique global minimizer.

Extension of Logistic Regression to Multi-Class Classification

1. We will use $\mathbf{W} = \mathbf{w}_1, ..., \mathbf{w}_K$ to avoid heavy notation. We have that

$$\log \mathbb{P}[\hat{\mathbf{y}} = \mathbf{y}|\mathbf{X}, \mathbf{W}] = \log \prod_{n=1}^{N} \mathbb{P}[\hat{y}_n = y_n|\mathbf{x}_n, \mathbf{W}]$$

Where $\hat{\mathbf{y}}$ are our predictions and $\mathbf{y}$ represent the ground truth for our samples. We can rewrite the equation as follow, dividing the samples in groups based on their class.

$$\log \mathbb{P}[\hat{\mathbf{y}} = \mathbf{y}|\mathbf{X}, \mathbf{W}] = \log \prod_{n:y_n=1} \mathbb{P}[\hat{y}_n = 1|\mathbf{x}_n, \mathbf{W}]... \prod_{n:y_n=K} \mathbb{P}[\hat{y}_n = K|\mathbf{x}_n, \mathbf{W}]$$

We introduce the following notation to simplify the expression. Let $1_{y_n=k}$ be the indicator function for $y_n = k$, i.e., it is equal to one if $y_n = k$ and 0 otherwise. Notice that we can write that

$$\mathbb{P}[\hat{y}_n = k|\mathbf{x}_n, \mathbf{W}] = \prod_{j=1}^{K} \mathbb{P}[\hat{y}_n = j|\mathbf{x}_n, \mathbf{W}]^{1_{y_n=j}},$$

as $\mathbb{P}[\hat{y}_n = j|\mathbf{x}_n, \mathbf{W}]^{1_{y_n=j}}$ is 1 when $j \neq k$ (elevating to 0), whereas $\mathbb{P}[\hat{y}_n = k|\mathbf{x}_n, \mathbf{W}]$ is left unchanged.

$$\begin{aligned}
\log \mathbb{P}[\hat{\mathbf{y}} = \mathbf{y}|\mathbf{X}, \mathbf{W}] &= \log \prod_{k=1}^{K} \prod_{n=1}^{N} \mathbb{P}[\hat{y}_n = k|\mathbf{x}_n, \mathbf{W}]^{1_{y_n=k}} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \log \mathbb{P}[\hat{y}_n = k|\mathbf{x}_n, \mathbf{W}] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \left[\mathbf{w}_k^\top \mathbf{x}_n - \log \sum_{j=1}^{K} \exp(\mathbf{w}_j^\top \mathbf{x}_n) \right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \mathbf{w}_k^\top \mathbf{x}_n - \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \log \sum_{j=1}^{K} \exp(\mathbf{w}_j^\top \mathbf{x}_n) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \mathbf{w}_k^\top \mathbf{x}_n - \sum_{n=1}^{N} \log \sum_{k=1}^{k} \exp(\mathbf{w}_k^\top \mathbf{x}_n).
\end{aligned}$$

The last step is obtained by $\sum_{k=1}^{K} 1_{y_n=k} = 1$.

2. We get

$$\frac{\partial \log \mathbb{P}[\mathbf{y}|\mathbf{X}, \mathbf{W}]}{\partial \mathbf{w}_k} = \sum_{n=1}^{N} 1_{y_n=k} \mathbf{x}_n - \sum_{n=1}^{N} \text{softmax}(n, k) \mathbf{x}_n.$$

Where softmax$(n, k) = \frac{\exp(\eta_{nk})}{\sum_{j=1}^{K} \exp(\eta_{nj})}$.

3. The negative of the log-likelihood is

$$-\sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \mathbf{w}_k \mathbf{x}_n + \sum_{n=1}^{N} \log \sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \mathbf{x}_n).$$

We have already shown that a sum of convex functions is convex, so we only need to show that the following is convex.

$$-\sum_{k=1}^{K} 1_{y_n=k} \mathbf{w}_k \mathbf{x}_n + \log \sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \mathbf{x}_n).$$

The first part is a linear function, which is convex. We only need to prove that the following is convex.

$$\log \sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \mathbf{x}_n)$$

This form is know as a log-sum-exp, and you may know that it is convex. It would be perfectly fine to use this as a fact, but we will prove it using the definition of convexity for the sake of completeness.

To prove: We want to show that for all sets of weights $\mathbf{A} = \mathbf{a}_1, ..., \mathbf{a}_K, \mathbf{B} = \mathbf{b}_1, ..., \mathbf{b}_K$, we have that

$$\lambda \log \left(\sum_k e^{\mathbf{a}_k^\top \mathbf{x}} \right) + (1-\lambda) \log \left(\sum_k e^{\mathbf{b}_k^\top \mathbf{x}} \right) \geq \log \left(\sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}} \right).$$

Simplifying the expression: First, we define $\mathbf{u}_k = e^{\mathbf{a}_k^\top \mathbf{x}}$ and $\mathbf{v}_k = e^{\mathbf{b}_k^\top \mathbf{x}}$, where $\mathbf{u}_k > 0$ and $\mathbf{v}_k > 0$. Thus,

$$\log \left(\sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}} \right) = \log \left(\sum_k \left(e^{\mathbf{a}_k^\top \mathbf{x}} \right)^\lambda \left(e^{\mathbf{b}_k^\top \mathbf{x}} \right)^{1-\lambda} \right) = \log \left(\sum_k (\mathbf{u}_k)^\lambda (\mathbf{v}_k)^{1-\lambda} \right),$$

and we would like to prove

$$\lambda \log \left(\sum_k \mathbf{u}_k \right) + (1-\lambda) \log \left(\sum_k \mathbf{v}_k \right) \geq \log \left(\sum_k \mathbf{u}_k^\lambda \mathbf{v}_k^{1-\lambda} \right).$$

From Hölder's inequality:

$$\sum_k |x_k y_k| \leq \left(\sum_k |x_k|^p \right)^{\frac{1}{p}} \left(\sum_k |y_k|^q \right)^{\frac{1}{q}},$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

We can apply this inequality with $\frac{1}{p} = \lambda$ and $\frac{1}{q} = 1 - \lambda$ to $\log \left(\sum_k \mathbf{u}_k^\lambda \mathbf{v}_k^{1-\lambda} \right)$,i.e.,

$$\log \left(\sum_k \mathbf{u}_k^\lambda \mathbf{v}_k^{1-\lambda} \right) = \log \left(\sum_k |\mathbf{u}_k^\lambda| |\mathbf{v}_k^{1-\lambda}| \right) \leq \log \left(\left(\sum_k |\mathbf{u}_k^\lambda|^{\frac{1}{\lambda}} \right)^\lambda \left(\sum_k |\mathbf{v}_k^{1-\lambda}|^{\frac{1}{1-\lambda}} \right)^{1-\lambda} \right),$$

where the right formula can be reduced to:

$$\log \left(\left(\sum_k \mathbf{u}_k \right)^\lambda \left(\sum_k \mathbf{v}_k \right)^{1-\lambda} \right) = \lambda \log \left(\sum_k \mathbf{u}_k \right) + (1-\lambda) \log \left(\sum_k \mathbf{v}_k \right).$$

As a result,

$$\log \left(\sum_k \mathbf{u}_k^\lambda \mathbf{v}_k^{1-\lambda} \right) \leq \lambda \log \left(\sum_k \mathbf{u}_k \right) + (1-\lambda) \log \left(\sum_k \mathbf{v}_k \right),$$

which concludes the proof.

Mixture of Linear Regression

1. Likelihood: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{r}_n) = \prod_{k=1}^{K}[\mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)]^{r_{nk}}$.

2. Joint likelihood: $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{r}) = \prod_{n=1}^{N}\prod_{k=1}^{K}[\mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)]^{r_{nk}}$.

3. Write the joint, then the conditional, and plug in.

$$\begin{aligned} p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi}) &= \sum_{k=1}^{K} p(y_n, r_n = k|\boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi}) = \sum_{k=1}^{K} p(y_n|r_n = k, \boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi})p(r_n = k|\boldsymbol{\pi}) \\ &= \sum_{k=1}^{K} p(y_n|r_n = k, \boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi})\pi_k = \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)\pi_k \end{aligned}$$

4.

$$\begin{aligned} -\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\pi}) &= -\log \prod_{n=1}^{N}\sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)\pi_k \\ &= -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)\pi_k \end{aligned}$$

5. (a) The model is not *convex* in general. E.g., consider the case when $N = 1$, $K = 2$. Then negative log-likelihood is equal to

$$\frac{1}{2}\log 2\pi\sigma^2 - \log\left[\exp\left(-\frac{(y - \boldsymbol{w}_1^\top \boldsymbol{x})^2}{2\sigma^2}\right)\pi_1 + \exp\left(-\frac{(y - \boldsymbol{w}_2^\top \boldsymbol{x})^2}{2\sigma^2}\right)(1 - \pi_1)\right]$$

The first term is a constant, we will look only at the second term and prove that it is not convex. Define

$$f(\boldsymbol{w}_1, \boldsymbol{w}_2, \pi_1) := -\log\left[\exp\left(-\frac{(y - \boldsymbol{w}_1^\top \boldsymbol{x})^2}{2\sigma^2}\right)\pi_1 + \exp\left(-\frac{(y - \boldsymbol{w}_2^\top \boldsymbol{x})^2}{2\sigma^2}\right)(1 - \pi_1)\right]$$

In order to prove that $f(\boldsymbol{w}_1, \boldsymbol{w}_2, \pi_1)$ is not convex we will construct two points $p^1 = (\boldsymbol{w}_1^1, \boldsymbol{w}_2^1, \pi_1^1)$ and $p^2 = (\boldsymbol{w}_1^2, \boldsymbol{w}_2^2, \pi_1^2)$ such that $f(\frac{1}{2}p^1 + \frac{1}{2}p^2) > \frac{1}{2}f(p^1) + \frac{1}{2}f(p^2)$.

Let

$$p^1 = \left(\frac{y}{\|\boldsymbol{x}\|_2^2}\boldsymbol{x}, \frac{y+2}{\|\boldsymbol{x}\|_2^2}\boldsymbol{x}, 1\right) \qquad p^2 = \left(\frac{y+2}{\|\boldsymbol{x}\|_2^2}\boldsymbol{x}, \frac{y}{\|\boldsymbol{x}\|_2^2}\boldsymbol{x}, 0\right),$$

note that $\boldsymbol{x} \neq \boldsymbol{0}$ since its first coordinate is equal to 1 as stated in the exercise. Then

$$\begin{aligned} f(p^1) &= -\log\left[\exp\left(-\frac{0}{2\sigma^2}\right)\right] = 0 \\ f(p^2) &= -\log\left[\exp\left(-\frac{0}{2\sigma^2}\right)\right] = 0 \\ f\left(\frac{1}{2}p^1 + \frac{1}{2}p^2\right) &= -\log\left[\exp\left(-\frac{1}{2\sigma^2}\right)\right] = \frac{1}{2\sigma^2} > 0 \end{aligned}$$

This proves that negative log-likelihood is not convex in general.

(b) The given model is not identifiable by permutation of indexes of mixture components.

Assume that the model is identifiable and true solution is $\boldsymbol{w}^\star, \boldsymbol{\pi}^\star$ is found my MLE when the data size grows to infinity, i.e.

$$\boldsymbol{w}^\star, \boldsymbol{\pi}^\star = \arg\min_{\boldsymbol{w}, \boldsymbol{\pi}}\left[L(\boldsymbol{w}, \boldsymbol{\pi}) := -\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\pi})\right]$$

Then we will construct the second point $\hat{\boldsymbol{w}}, \hat{\boldsymbol{\pi}} \neq \boldsymbol{w}^\star, \boldsymbol{\pi}^\star$ such that $L(\hat{\boldsymbol{w}}, \hat{\boldsymbol{\pi}}) = L(\boldsymbol{w}^\star, \boldsymbol{\pi}^\star)$. This would mean that $\hat{\boldsymbol{w}}, \hat{\boldsymbol{\pi}}$ is also a solution of MLE and there is no way to distinguish between the true solution $\boldsymbol{w}^\star, \boldsymbol{\pi}^\star$ and a point $\hat{\boldsymbol{w}}, \hat{\boldsymbol{\pi}}$, so MLE doesn't always give a true solution.

We define $\hat{\boldsymbol{w}}, \hat{\boldsymbol{\pi}}$ as follows

$$\begin{aligned} \hat{\boldsymbol{w}}_1 &= \boldsymbol{w}_2^\star & \hat{\boldsymbol{\pi}}_1 &= \boldsymbol{\pi}_2^\star \\ \hat{\boldsymbol{w}}_2 &= \boldsymbol{w}_1^\star & \hat{\boldsymbol{\pi}}_2 &= \boldsymbol{\pi}_1^\star \\ \hat{\boldsymbol{w}}_i &= \boldsymbol{w}_i^\star, i \geq 3 & \hat{\boldsymbol{\pi}}_i &= \boldsymbol{\pi}_i^\star, i \geq 3, \end{aligned}$$

i.e. vectors corresponding to the first two mixture components are permuted. (We assume that $\boldsymbol{w}_1^\star \neq \boldsymbol{w}_2^\star$ as they represent two different components).

Then indeed the losses at these two points are equal,

$$L(\hat{\boldsymbol{w}}, \hat{\boldsymbol{\pi}}) = -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \mathcal{N}(y_n | \hat{\boldsymbol{w}}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2) \hat{\pi}_k = -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \mathcal{N}(y_n | {\boldsymbol{w}^\star}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2) {\pi^\star}_k = L(\boldsymbol{w}^\star, \boldsymbol{\pi}^\star).$$

This ends the proof.