Supplemental Material for Shift Happens: Adjusting Classifiers

Theodore James Thibault Heiser, Mari-Liis Allikivi, and Meelis Kull

Institute of Computer Science, University of Tartu, Tartu, Estonia
{mari-liis.allikivi,meelis.kull}@ut.ee*

1 Introduction

This supplemental material of the paper *Shift Happens: Adjusting Classifiers* first lists all the definitions of the paper, followed by theorems and proofs and also the lemmas needed to complete these proofs. The source code and the list of IDs of OpenML tasks and runs that enter our experiments have been provided as file `source_code.zip` together with the current document.

2 Definitions

Definition 1 (Proper Scoring Rule (or Proper Loss)). *In a k-class classification task a loss function $f : [0,1]^k \times \{0,1\}^k \to \mathbb{R}$ is called a* proper scoring rule *(or* proper loss*), if for any probability vectors $p, q \in [0,1]^k$ with $\sum_{i=1} p_i = 1$ and $\sum_{i=1} q_i = 1$ the following inequality holds:*

$$\mathbb{E}_{Y\sim q}[f(q,Y)] \leq \mathbb{E}_{Y\sim q}[f(p,Y)]$$

where Y is a one-hot encoded label randomly drawn from the categorical distribution over k classes with class probabilities represented by vector q. The loss function f is called strictly proper *if the inequality is strict for all $p \neq q$.*

Definition 2 (Bregman Divergence). *Let $\phi : \Omega \to \mathbb{R}$ be a strictly convex function defined on a convex set $\Omega \subseteq \mathbb{R}^k$ such that ϕ is differentiable on the relative interior of Ω, $ri(\Omega)$. The Bregman divergence $d_\phi : ri(\Omega) \times \Omega \to [0,\infty)$ is defined as*

$$d_\phi(p,q) = \phi(q) - \phi(p) - \langle q - p, \nabla\phi(p)\rangle$$

Definition 3 (Adjusted Predictions). *Let $p \in [0,1]^{n\times k}$ be the predictions of a probabilistic k-class classifier on n instances and let $\pi \in [0,1]^k$ be the actual class distribution on these instances. We say that* predictions p are adjusted on this dataset*, if the average prediction is equal to the class proportion for every class j, that is $\frac{1}{n}\sum_{i=1}^{n} p_{ij} = \pi_j$.*

* T. Heiser can be reached at teddyheiser@google.com.

Definition 4 (Additive Adjustment). Additive adjustment *is the function* $\alpha_+ : [0,1]^{n\times k} \times [0,1]^k \to [0,1]^{n\times k}$ *which takes in the predictions of a probabilistic* k*-class classifier on* n *instances and the actual class distribution* π *on these instances, and outputs adjusted predictions* $a = \alpha_+(p,\pi)$ *defined as* $a_{i\cdot} = p_{i\cdot} + (\varepsilon_1, \ldots, \varepsilon_k)$ *where* $a_{i\cdot} = (a_{i1}, \ldots, a_{ik})$, $p_{i\cdot} = (p_{i1}, \ldots, p_{ik})$, *and* $\varepsilon_j = \pi_j - \frac{1}{n}\sum_{i=1}^{n} p_{ij}$ *for each class* $j \in \{1, \ldots, k\}$.

Definition 5 (Adjustment Procedure). Adjustment procedure *is any function* $\alpha : [0,1]^{n\times k} \times [0,1]^k \to [0,1]^{n\times k}$ *which takes as arguments the predictions* p *of a probabilistic* k*-class classifier on* n *instances and the actual class distribution* π *on these instances, such that for any* p *and* π *the output predictions* $a = \alpha(p,\pi)$ *are adjusted, that is* $\frac{1}{n}\sum_{i=1}^{n} a_{ij} = \pi_j$ *for each class* $j \in \{1, \ldots, k\}$.

Definition 6 (Bounded Adjustment Procedure). *An adjustment procedure* $\alpha : [0,1]^{n\times k} \times [0,1]^k \to [0,1]^{n\times k}$ *is* bounded, *if for any* p *and* π *the output predictions* $a = \alpha(p,\pi)$ *are in the range* $[0,1]$, *that is* $a_{ij} \in [0,1]$ *for all* i, j.

Definition 7 (Multiplicative Adjustment). Multiplicative adjustment *is the function* $\alpha_* : [0,1]^{n\times k} \times [0,1]^k \to [0,1]^{n\times k}$ *which takes in the predictions of a probabilistic* k*-class classifier on* n *instances and the actual class distribution* π *on these instances, and outputs adjusted predictions* $a = \alpha_*(p,\pi)$ *defined as* $a_{ij} = \frac{w_j p_{ij}}{z_i}$, *where* $w_1, \ldots, w_k \geq 0$ *are real-valued weights chosen based on* p *and* π *such that the predictions* $\alpha_*(p,\pi)$ *would be adjusted, and* z_i *are the renormalisation factors defined as* $z_i = \sum_{j=1}^{k} w_j p_{ij}$.

Definition 8 (Coherence of Adjustment Procedure and Bregman Divergence [1]). *Let* $\alpha : [0,1]^{n\times k} \times [0,1]^k \to [0,1]^{n\times k}$ *be an adjustment procedure and* d_ϕ *be a Bregman divergence. Then* α *is called to be coherent with* d_ϕ *if and only if for any predictions* p *and class distribution* π *the following holds for all* $i = 1, \ldots, n$ *and* $j, j' = 1, \ldots, k$*:*

$$(d_\phi(a_{i\cdot}, c_j) - d_\phi(p_{i\cdot}, c_j)) - (d_\phi(a_{i\cdot}, c_{j'}) - d_\phi(p_{i\cdot}, c_{j'})) = const_{j,j'}$$

where $const_{j,j'}$ *is a quantity not depending on* i, *and where* $a = \alpha(p,\pi)$ *and* c_j *is a one-hot encoded vector corresponding to class* j *(with* 1 *at position* j *and* 0 *everywhere else).*

Definition 9 (Unbounded General Adjuster (UGA)). *Consider a* k*-class classification task with a test dataset of* n *instances, and let* d_ϕ *be a Bregman divergence. Then the* unbounded general adjuster corresponding to d_ϕ *is the function* $\alpha^\star : \mathbb{R}^{n\times k} \times \mathbb{R}^k \to \mathbb{R}^{n\times k}$ *defined as follows:*

$$\alpha^\star(p,\pi) = \operatorname*{arg\,min}_{a \in Q_\pi} \frac{1}{n}\sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_{i\cdot})$$

Definition 10 (Bounded General Adjuster (BGA)). *Consider a* k*-class classification task with a test dataset of* n *instances, and let* d_ϕ *be a Bregman divergence. Then the* bounded general adjuster corresponding to d_ϕ *is the function*

$\alpha^{\square} : [0,1]^{n\times k} \times [0,1]^k \to [0,1]^{n\times k}$ *defined as follows:*

$$\alpha^{\square}(p,\pi) = \underset{a\in Q_\pi^{\square}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_{i\cdot})$$

3 Theorems and Lemmas with Proofs

Theorem 1 (Decomposition of Bregman Divergences [1]). *Let d_ϕ be a Bregman divergence and let $\alpha : [0,1]^{n\times k} \times [0,1]^k \to [0,1]^{n\times k}$ be an adjustment procedure coherent with d_ϕ. Then for any predictions p, one-hot encoded true labels $y \in \{0,1\}^{n\times k}$ and class distribution π (with $\pi_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$) the following decomposition holds:*

$$\frac{1}{n}\sum_{i=1}^{n} d_\phi(p_{i\cdot}, y_{i\cdot}) = \frac{1}{n}\sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_{i\cdot}) + \frac{1}{n}\sum_{i=1}^{n} d_\phi(a_{i\cdot}, y_{i\cdot}) \tag{1}$$

Proof. Proof given in cited article [1].

Lemma 1. *Let $d_\phi : ri(\Omega) \times \Omega \to \mathbb{R}$ be a Bregman divergence. Then for any $p, q \in ri(\Omega)$ the following holds:*

$$\nabla_q d_\phi(p,q) = \nabla\phi(q) - \nabla\phi(p),$$

where ∇_q notates the gradient with respect to vector q.

Proof. By the definition of Bregman divergence,

$$d_\phi(p,q) = \phi(q) - \phi(p) - \langle q-p, \nabla\phi(p)\rangle.$$

The required result follows by taking ∇_q of each side and simplifying:

$$\begin{aligned}
\nabla_q d_\phi(p,q) &= \nabla_q(\phi(q) - \phi(p) - \langle q-p, \nabla\phi(p)\rangle) \\
&= \nabla_q\phi(q) - \nabla_q\phi(p) - \nabla_q\langle q-p, \nabla\phi(p)\rangle \\
&= \nabla\phi(q) - \nabla_q\langle q-p, \nabla\phi(p)\rangle \\
&= \nabla\phi(q) - \nabla_q\langle q, \nabla\phi(p)\rangle \\
&= \nabla\phi(q) - \nabla_q(q_1\frac{\partial}{\partial p_1}\phi(p) + \ldots + q_k\frac{\partial}{\partial p_k}\phi(p)) \\
&= \nabla\phi(q) - (\frac{\partial}{\partial p_1}\phi(p), \ldots, \frac{\partial}{\partial p_k}\phi(p)) \\
&= \nabla\phi(q) - \nabla\phi(p)
\end{aligned}$$

Lemma 2. *Let $d_\phi : ri(\Omega) \times \Omega \to \mathbb{R}$ be a Bregman divergence. Then for any $p, q \in ri(\Omega)$ and $z \in \Omega$ the following holds:*

$$d_\phi(p,z) - d_\phi(q,z) = \langle z-q, \nabla_q d_\phi(p,q)\rangle + d_\phi(p,q).$$

Proof. Simplifying from the definition of Bregman divergence gives:

$$\begin{aligned} d_\phi(p,z) - d_\phi(q,z) &= (\phi(z) - \phi(p) - \langle z-p, \nabla\phi(p)\rangle) - (\phi(z) - \phi(q) - \langle z-q, \nabla\phi(q)\rangle) \\ &= \phi(q) - \phi(p) + \langle z-q, \nabla\phi(q)\rangle - \langle z-p, \nabla\phi(p)\rangle \end{aligned}$$

Using Lemma 1 to rewrite the third term yields:

$$\begin{aligned} &= \phi(q) - \phi(p) + \langle z-q, \nabla_q d_\phi(p,q) + \nabla\phi(p)\rangle - \langle z-p, \nabla\phi(p)\rangle \\ &= \phi(q) - \phi(p) + \langle z-q, \nabla_q d_\phi(p,q)\rangle + \langle z-q, \nabla\phi(p)\rangle - \langle z-p, \nabla\phi(p)\rangle \\ &= \phi(q) - \phi(p) + \langle z-q, \nabla_q d_\phi(p,q)\rangle - \langle q, \nabla\phi(p)\rangle + \langle p, \nabla\phi(p)\rangle \\ &= \phi(q) - \phi(p) + \langle z-q, \nabla_q d_\phi(p,q)\rangle - \langle q-p, \nabla\phi(p)\rangle \\ &= \langle z-q, \nabla_q d_\phi(p,q)\rangle + \phi(q) - \phi(p) - \langle q-p, \nabla\phi(p)\rangle \\ &= \langle z-q, \nabla_q d_\phi(p,q)\rangle + d_\phi(p,q) \end{aligned}$$

Lemma 3. *Let d_ϕ be a Bregman divergence, let p be a set of predictions, and π be a class distribution over k classes. Denoting $a^\star = \alpha^\star(p,\pi)$, the following holds for any $q \in Q_\pi$:*

$$\frac{1}{n}\sum_{i=1}^{n}(d_\phi(p_{i\cdot}, q_{i\cdot}) - d_\phi(a^\star_{i\cdot}, q_{i\cdot})) = \frac{1}{n}\sum_{i=1}^{n}(d_\phi(p_{i\cdot}, a^\star_{i\cdot}))$$

Proof. In the following we will use a simplified notation and write p_i, q_i, $a^\star_i$ instead of $p_{i\cdot}$, $q_{i\cdot}$, $a^\star_{i\cdot}$. Using Lemma 2 we can write

$$\frac{1}{n}\sum_{i=1}^{n}(d_\phi(p_i, q_i) - d_\phi(a^\star_i, q_i)) = \frac{1}{n}\sum_{i=1}^{n}(\langle q_i - a^\star_i, \nabla_{a^\star_i} d_\phi(p_i, a^\star_i)\rangle + d_\phi(p_i, a^\star_i))$$

If we can prove that

$$\sum_{i=1}^{n}\langle q_i - a^\star_i, \nabla_{a^\star_i} d_\phi(p_i, a^\star_i)\rangle = 0$$

then the proof will be complete. So we begin by using the method of Lagrange multipliers to define what each $\nabla_{a^\star_i} d_\phi(p_i, a^\star_i)$ is for each $i \in \{1, \ldots, n\}$. We rewrite the original argument minimization problem. Keep note our new function will have $n \times k$ variables from a, and n variables from our first constraint, and k variables from our second constraint.

$$F(a,\theta,\lambda) = \sum_{i=1}^{n} d_\phi(p_i, a_i) + \sum_{i=1}^{n}\theta_i(1 - \sum_{j=1}^{k} a_{i,j}) + \sum_{j=1}^{k}\lambda_j(\pi_j - \frac{1}{n}\sum_{i=1}^{n} a_{i,j})$$

Minimum is when

$$\nabla F(a,\theta,\lambda) = \mathbf{0}.$$

Let's expand the gradient.

$$\nabla F(a,\theta,\lambda) = (\nabla_a F(a,\theta,\lambda), \nabla_\theta F(a,\theta,\lambda), \nabla_\lambda F(a,\theta,\lambda))$$

Let's expand the first term. For simplicity's sake we will represent ∇_a as a matrix, but it is a vector in actuality.

$$\nabla_a F(a,\theta,\lambda) = \begin{bmatrix} \frac{\partial}{\partial a_{1,1}} F(a,\theta,\lambda) \ \dots \ \frac{\partial}{\partial a_{1,k}} F(a,\theta,\lambda) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \frac{\partial}{\partial a_{n,1}} F(a,\theta,\lambda) \ \dots \ \frac{\partial}{\partial a_{n,k}} F(a,\theta,\lambda) \end{bmatrix}$$

$$= \begin{bmatrix} \theta_1 + \lambda_1 + \frac{\partial}{\partial a_{1,1}} d_\phi(p_1,a_1) \ \dots \ \theta_1 + \lambda_k + \frac{\partial}{\partial a_{1,k}} d_\phi(p_1,a_1) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \theta_n + \lambda_1 + \frac{\partial}{\partial a_{n,1}} d_\phi(p_n,a_n) \ \dots \ \theta_n + \lambda_k + \frac{\partial}{\partial a_{n,k}} d_\phi(p_n,a_n) \end{bmatrix}$$

We can now see that for each entry (i,j) in $\nabla_a F(a,\theta,\lambda)$ to equal 0, then each

$$\frac{\partial}{\partial a_{i,j}} d_\phi(p_i,a_i) = -\theta_i - \lambda_j$$

Since $a^\star$ is at the minimum, this implies

$$\nabla_{a_i^\star} d_\phi(p_i,a_i^\star) = (-\theta_i - \lambda_1, ..., -\theta_i - \lambda_k)$$

We can now write out

$$\begin{aligned} \sum_{i=1}^n \langle q_i - a_i^\star, \nabla_{a_i^\star} d_\phi(p_i,a_i^\star)\rangle &= \sum_{i=1}^n \sum_{j=1}^k (q_{i,j} - a_{i,j}^\star)(-\theta_i - \lambda_j) \\ &= \sum_{i=1}^n \sum_{j=1}^k (q_{i,j} - a_{i,j}^\star)(-\theta_i) + \sum_{i=1}^n \sum_{j=1}^k (q_{i,j} - a_{i,j}^\star)(-\lambda_j) \\ &= \sum_{i=1}^n (-\theta_i) \sum_{j=1}^k (q_{i,j} - a_{i,j}^\star) + \sum_{j=1}^k (-\lambda_j) \sum_{i=1}^n (q_{i,j} - a_{i,j}^\star) \end{aligned}$$

We know from the constraints that each row and column of $q - a^\star$ sums to 0.

$$\sum_{j=1}^k (q_{i,j} - a_{i,j}^\star) = 0 \text{ and } \sum_{i=1}^n (q_{i,j} - a_{i,j}^\star) = 0$$

So it's clear that

$$\sum_{i=1}^n (-\theta_i) \sum_{j=1}^k (q_{i,j} - a_{i,j}^\star) + \sum_{j=1}^k (-\lambda_j) \sum_{i=1}^n (q_{i,j} - a_{i,j}^\star) = 0$$

Theorem 2. *Let $\alpha^\star$ be the unbounded general adjuster corresponding to the Bregman divergence d_ϕ. Then $\alpha^\star$ is coherent with d_ϕ.*

Proof. Let $a^\star = \alpha^\star(p,\pi)$. For $\alpha^\star$ to be coherent. the following equation must be satisfied following the definition of coherence (we use notation e_i instead of c_i

to emphasise that these are unit vectors, we use letters i and j instead of j and j', and letter x to stand for a row in matrices $a^\star$ and p):

$$d_\phi(a^\star_x, e_i) - d_\phi(a^\star_x, e_j) - d_\phi(p_x, e_i) + d_\phi(p_x, e_j) \stackrel{?}{=} const_{i,j}$$

We can just use the definition of divergence and properties of vectors to get the equation into a new form.

$$\begin{aligned}
const_{i,j} &\stackrel{?}{=} d_\phi(a^\star_x, e_i) - d_\phi(a^\star_x, e_j) - d_\phi(p_x, e_i) + d_\phi(p_x, e_j) \\
&= \phi(e_i) - \phi(a^\star_x) - \langle e_i - a^\star_x, \nabla\phi(a^\star_x)\rangle \\
&\quad - \phi(e_j) + \phi(a^\star_x) + \langle e_j - a^\star_x, \nabla\phi(a^\star_x)\rangle \\
&\quad - \phi(e_i) + \phi(p_x) + \langle e_i - p_x, \nabla\phi(p_x)\rangle \\
&\quad + \phi(e_j) - \phi(p_x) - \langle e_j - p_x, \nabla\phi(p_x)\rangle \\
&= \langle e_j - a^\star_x, \nabla\phi(a^\star_x)\rangle \\
&\quad - \langle e_i - a^\star_x, \nabla\phi(a^\star_x)\rangle \\
&\quad - \langle e_j - p_x, \nabla\phi(p_x)\rangle \\
&\quad + \langle e_i - p_x, \nabla\phi(p_x)\rangle \\
&= \langle e_j - e_i, \nabla\phi(a^\star_x)\rangle - \langle e_j - e_i, \nabla\phi(p_x)\rangle \\
&= \langle e_j - e_i, \nabla\phi(a^\star_x) - \nabla\phi(p_x)\rangle
\end{aligned}$$

From our earlier theorem, we know.

$$\langle e_j - e_i, \nabla\phi(a^\star_x) - \nabla\phi(p_x)\rangle = \langle e_j - e_i, \nabla_{a^\star_x} d_\phi(p_x, a^\star_x)\rangle$$

We know from the proof in Lemma 3 that $\nabla_{a^\star_x} d_\phi(p_x, a^\star_x)$ is defined by the sum of two variables that depend on i and j, θ and λ. That means $const_{i,j} = \langle e_j - e_i, \nabla^\star_{a^\star_x} d_\phi(p_x, a^\star_x)\rangle$ only depends on i and j and not on x, matching the definition of coherence.

Theorem 3. *Let d_ϕ be a Bregman divergence, let p be a set of predictions, and π be a class distribution over k classes. Suppose $a \in Q_\pi$ is such that for any $y \in Q_\pi$ the decomposition of Eq.(1) holds. Then $a = \alpha^\star(p, \pi)$.*

Proof. We prove by contradiction and assume that $a \neq \alpha^\star(p, \pi)$. Take the case where $q = \alpha^\star(p, \pi)$. We can rewrite the theorem's equality to

$$\sum_{i=1}^{n} d_\phi(p_i, q_i) = \sum_{i=1}^{n} (d_\phi(p_i, a_i) + d_\phi(a_i, q_i)).$$

By the definition of $\alpha^\star$ we have $\sum_{i=1}^n d_\phi(p_i, q_i) < \sum_{i=1}^n d_\phi(p_i, a_i))$ and by the definition of Bregman divergence $\sum_{i=1}^n d_\phi(a_i, q_i) > 0$. Therefore,

$$\sum_{i=1}^{n} d_\phi(p_i, q_i) < \sum_{i=1}^{n} (d_\phi(p_i, a_i) + d_\phi(a_i, q_i))$$

. We have a contradiction, so the assumption was false.

Lemma 4. *Let d_ϕ be a Bregman divergence, let p be a set of predictions, and π be a class distribution over k classes. Denoting $a^\square = \alpha^\square(p, \pi)$, the following holds for any $q \in Q_\pi^\square$:*

$$\sum_{i=1}^{n} \langle q_i - a_i^\square, \nabla_{a_i^\square} d_\phi(p_i, a_i^\square) \rangle \geq 0$$

Proof. This is pretty much like the proof of Lemma 3 except we use the Karush-Kuhn-Tucker method to add our extra set of inequality constraints.

$$\begin{aligned} F(a, \theta, \lambda, \psi) = &\sum_{i=1}^{n} d_\phi(p_i, a_i) + \sum_{i=1}^{n} \theta_i (1 - \sum_{j=1}^{k} a_{i,j}) \\ &+ \sum_{j=1}^{k} \lambda_j (\pi - \frac{1}{n} \sum_{i=1}^{n} a_{i,j}) + \sum_{i=1}^{n} \sum_{j=1}^{k} \psi_{i,j} (-a_{i,j}) \end{aligned}$$

Minimum is when

$$\nabla F(a, \theta, \lambda, \psi) = \mathbf{0}.$$

Let's expand the gradient.

$$\nabla F(a, \theta, \lambda, \psi) = (\nabla_a F(a, \theta, \lambda), \nabla_\theta F(a, \theta, \lambda), \nabla_\lambda F(a, \theta, \lambda), \nabla_\psi F(a, \theta, \lambda))$$

Let's expand the first term. For simplicity's sake we will represent ∇_a as a matrix, but it is a vector in actuality.

$$\nabla_a F(a, \theta, \lambda) = \begin{bmatrix} \frac{\partial}{\partial a_{1,1}} F(a, \theta, \lambda) & \dots & \frac{\partial}{\partial a_{1,k}} F(a, \theta, \lambda) \\ \dots & \dots & \dots \\ \frac{\partial}{\partial a_{n,1}} F(a, \theta, \lambda) & \dots & \frac{\partial}{\partial a_{n,k}} F(a, \theta, \lambda) \end{bmatrix}$$

$$= \begin{bmatrix} \theta_1 + \lambda_1 - \psi_{1,1} + \frac{\partial}{\partial a_{1,1}} d_\phi(p_1, a_1) & \dots & \theta_1 + \lambda_k - \psi_{1,k} + \frac{\partial}{\partial a_{1,k}} d_\phi(p_1, a_1) \\ \dots & \dots & \dots \\ \theta_n + \lambda_1 - \psi_{n,1} + \frac{\partial}{\partial a_{n,1}} d_\phi(p_n, a_n) & \dots & \theta_n + \lambda_k - \psi_{n,k} + \frac{\partial}{\partial a_{n,k}} d_\phi(p_n, a_n) \end{bmatrix}$$

We can now see that for each entry (i, j) in $\nabla_a F(a, \theta, \lambda, \psi)$ to equal 0, then each

$$\frac{\partial}{\partial a_{i,j}} d_\phi(p_i, a_i) = \psi_{i,j} - \theta_i - \lambda_j$$

Since $a^\square$ is at the minimum, this implies

$$\nabla_{a_i^\square} d_\phi(p_i, a_i^\square) = (\psi_{i,1} - \theta_i - \lambda_1, ..., \psi_{i,k} - \theta_i - \lambda_k)$$

We can write out

$$\begin{aligned}\sum_{i=1}^{n}\langle q_i - a_i^{\square}, \nabla_{a_i^{\square}} d_\phi(p_i, a_i^{\square})\rangle &= \sum_{i=1}^{n}\sum_{j=1}^{k}(q_{i,j} - a_{i,j}^{\square})(\psi_{i,j} - \theta_i - \lambda_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k}\psi_{i,j}(q_{i,j} - a_{i,j}^{\square}) \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{k}(q_{i,j} - a_{i,j}^{\square})(-\theta_i) \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{k}(q_{i,j} - a_{i,j}^{\square})(-\lambda_j)\end{aligned}$$

We know from the earlier proof of Lemma 3 that the last two terms equal 0, which leaves us

$$\sum_{i=1}^{n}\langle q_i - a_i^{\square}, \nabla_{a_i^{\square}} d_\phi(p_i, a_i^{\square})\rangle = \sum_{i=1}^{n}\sum_{j=1}^{k}\psi_{i,j}(q_{i,j} - a_{i,j}^{\square})$$

Now let's look at what each $\psi_{i,j}$ actually is. The KKT conditions require that each $\psi_{i,j} \geq 0$ and that $\psi_{i,j}(-a_{i,j}^{\square}) = 0$. This implies that the only times that $\psi_{i,j} \neq 0$ is when $a_{i,j}^{\square} = 0$ in which case $\psi_{i,j} \geq 0$.
In our double sum, we only have to be concerned with the terms that have an $a_{i,j}^{\square} = 0$ (all the other terms will be 0, since if $a_{i,j}^{\square} \neq 0$ then $\psi_{i,j} = 0$). In these cases, $q_{i,j} - a_{i,j}^{\square} > 0$ since $q_{i,j} \geq 0$ by the constraint. $q_{i,j} - a_{i,j}^{\square} \geq 0$ and $\psi_{i,j} \geq 0$, so $\sum_{i=1}^{n}\sum_{j=1}^{k}\psi_{i,j}(q_{i,j} - a_{i,j}^{\square}) \geq 0$.

Theorem 4. *Let d_ϕ be a Bregman divergence, let p be a set of predictions, and π be a class distribution over k classes. Then for any $y \in Q_\pi^{\square}$ the following holds:*

$$\begin{aligned}&\sum_{i=1}^{n}(d_\phi(p_{i\cdot}, y_{i\cdot}) - d_\phi(a_{i\cdot}^{\square}, y_{i\cdot})) \geq \\ &\geq \sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_{i\cdot}^{\square}) \geq \sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_{i\cdot}^{\star}) = \sum_{i=1}^{n}(d_\phi(p_{i\cdot}, y_{i\cdot}) - d_\phi(a_{i\cdot}^{\star}, y_{i\cdot}))\end{aligned}$$

Proof. Writing out the difference and using the previous Lemma 4 with $q = y$ gives:

$$\begin{aligned}\sum_{i=1}^{n}(d_\phi(p_{i\cdot}, y_{i\cdot}) - d_\phi(a_{i\cdot}^{\square}, y_{i\cdot})) &= \sum_{i=1}^{n}(\langle a_{i\cdot}^{\square} - y_{i\cdot}, \nabla_a d_\phi(p_{i\cdot}, a_{i\cdot}^{\square})\rangle + d_\phi(p_{i\cdot}, a_{i\cdot}^{\square})) \\ &= \sum_{i=1}^{n}\langle a_{i\cdot}^{\square} - y_{i\cdot}, \nabla_a d_\phi(p_{i\cdot}, a_{i\cdot}^{\square})\rangle + \sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_{i\cdot}^{\square}) \\ &\geq \sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_{i\cdot}^{\square})\end{aligned}$$

We know $\sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_{i\cdot}^{\square}) \geq \sum_{i=1}^{n} d_\phi(p_{i\cdot}, a_i^\star)$ since $a^\star$ and $a^{\square}$ are either equal or $a^{\square}$ would have been chosen over $a^\star$ in $\alpha^\star$'s minimization task. The rest follows from Lemma 4.

References

1. Kull, M., Flach, P.: Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In: Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. pp. 68–85. Springer (2015)