Supervised Learning: No Loss No Cry
— Supplementary Material —

Abstract

This is the Supplementary Material to Paper ”Supervised Learning: No Loss No Cry” by R. Nock and A.-K. Menon. To differentiate with the numberings in the main file, the numbering of Theorems is letter-based (A, B, ...).

I Table of contents

II Factsheet on Bregman divergences

We summarize in this section the results we use (both in the main file and in this SI) related to Bregman divergence with convex generator F,

$$D_F(z\|z') \doteq F(z) - F(z') - (z - z')F'(z'), \quad (1)$$

where we assume for the sake of simplicity that F is twice differentiable.

$\triangleright$ **General properties –** D_F is always non-negative, convex in its left parameter, but not always in its right parameter. Only the divergences corresponding to $F(z) \propto z^2$ are symmetric (Boissonnat et al., 2010).

$\triangleright$ **D_F is locally proportional to the square loss –** assuming second order differentiability, we have (Nock et al., 2008):

$$\forall z, z', \exists c \in [z \wedge z', z \vee z'] : D_F(z\|z') = \frac{F''(c)}{2} \cdot (z - z')^2. \quad (2)$$

$\triangleright$ **Bregman triangle equality –** also called the three points property (Nock et al., 2008, 2016),

$$\forall z, z', z'', D_F(z\|z'') = D_F(z\|z') + D_F(z'\|z'') + (F'(z'') - F'(z'))(z' - z). \quad (3)$$

$\triangleright$ **Invariance to affine terms –** for any affine function $G(z)$ (Boissonnat et al., 2010),

$$\forall z, z', D_{F+G}(z\|z') = D_F(z\|z'). \quad (4)$$

$\triangleright$ **Dual symmetry –** letting $F^\star$ denote the convex conjugate of F, we have (Nock et al., 2016),

$$\forall z, z', D_F(z\|z') = D_{F^\star}(F'(z')\|F'(z)). \quad (5)$$

$\triangleright$ **The right population minimizer is the mean –** we have (Banerjee et al., 2004),

$$\arg\min_z \mathbb{E}_{\mathsf{Z}}[D_F(\mathsf{Z}\|z)] = \mathbb{E}_{\mathsf{Z}}[\mathsf{Z}] \doteq \mu(\mathsf{Z}). \quad (6)$$

$\triangleright$ **Bregman information –** the Bregman information of random variable Z, defined as $I_F(\mathsf{Z}) \doteq \min_z \mathbb{E}_{\mathsf{Z}}[D_F(\mathsf{Z}\|z)]$, satisfies (Banerjee et al., 2004)

$$I_F(\mathsf{Z}) = \mathbb{E}_{\mathsf{Z}}[D_F(\mathsf{Z}\|\mu(\mathsf{Z}))]. \quad (7)$$

III Proof of Theorem 1

($\Rightarrow$) The proof assumes basic knowledge about proper losses as in Reid & Williamson (2010) (and references therein) for example. It comes from Reid & Williamson (2010, Theorem 1, Corollary 3) and Shuford et al. (1966) that a differentiable function defines a proper loss iff there exists a Riemann integrable (eventually improper in the integrability sense) function $w : (0, 1) \to \mathbb{R}_+$ such that:

$$w(c) = \frac{\ell'_{-1}(c)}{c} = -\frac{\ell'_1(c)}{1 - c} \ , \forall c \in (0, 1). \quad (8)$$

To simplify notations, we slightly abuse notations and let $\underline{L}'' \doteq -w$ and define $\underline{L}'(u) \doteq \int_a^u \underline{L}''(z)\mathrm{d}z$ for some adequately chosen constant a (for example, $a = 1/2$ for symmetric proper canonical losses Nock & Nielsen (2009, 2008)). We denote such a representation of loss functions their integral representation (Reid & Williamson, 2010, eq. (5)), as it gives:

$$\ell_1(c) = \int_c^1 -(1-u)\underline{L}''(u)\mathrm{d}u, \tag{9}$$

from which we derive by integrating by parts,

$$\begin{aligned}
\ell_1(c) &= -\left[(1-u)\underline{L}'(u)\right]_c^1 - \int_c^1 \underline{L}'(u)\mathrm{d}u \\
&= (1-c)\underline{L}'(c) - \underline{L}(1) + \underline{L}(c) \qquad (10)\\
&= (-\underline{L})(1) - (-\underline{L})(c) - (1-c)(-\underline{L})'(c) \qquad (11)\\
&= D_{-\underline{L}}(1\|c), \qquad (12)
\end{aligned}$$

Where $D_{-\underline{L}}$ is the Bregman divergence with generator $-\underline{L}$ (we remind that the conditional Bayes risk of a proper loss is concave (Reid & Williamson, 2010, Section 3.2)). We get similarly for the partial loss ℓ_{-1} (Reid & Williamson, 2010, eq. (5)):

$$\begin{aligned}
\ell_{-1}(c) &= -\int_0^c u\underline{L}''(u)\mathrm{d}u \\
&= -\left[u\underline{L}'(u)\right]_0^c + \int_0^c \underline{L}'(u)\mathrm{d}u \\
&= -c\underline{L}'(c) + \underline{L}(c) - \underline{L}(0) \qquad (13)\\
&= (-\underline{L})(0) - (-\underline{L})(c) - (0-c)(-\underline{L})'(c) \qquad (14)\\
&= D_{-\underline{L}}(0\|c). \qquad (15)
\end{aligned}$$

We now replace c by the inverse of the link chosen, ψ, and we get for any proper composite loss:

$$\begin{aligned}
\ell(y^*, z) &\doteq [\![y^* = 1]\!] \cdot \ell_1(\psi^{-1}(z)) + [\![y^* = -1]\!] \cdot \ell_{-1}(\psi^{-1}(z)) \\
&= D_{-\underline{L}}(y\|\psi^{-1}(z)), \qquad (16)
\end{aligned}$$

as claimed for the implication $\Rightarrow$. The identity

$$D_{-\underline{L}}(y\|\psi^{-1}(z)) = D_{(-\underline{L})^\star}(-\underline{L}' \circ \psi^{-1}(z)\|-\underline{L}'(y)) \tag{17}$$

follows from the dual symmetry property of Bregman divergences (Boissonnat et al., 2010; Nock et al., 2016).

($\Leftarrow$) Let $\ell(y^*, z) \doteq D_{-F}(y\|g^{-1}(z))$, some Bregman divergence, where $g : [0,1] \rightarrow \mathbb{R}$ is invertible. Let $\ell_p(y^*, c) : \mathcal{Y} \times [0,1] \rightarrow \overline{\mathbb{R}}$ defined by $\ell_p(y^*, c) \doteq \ell(y^*, g(c))$. We know that the right population minimizer of any Bregman divergence is the expectation (Banerjee et al., 2004; Nock et al., 2016), so $\pi \in \arg\inf_u \mathsf{E}_{\mathsf{Y}\sim\pi}\ell_p(\mathsf{Y}, u), \forall \pi \in [0,1]$ and ℓ_p is proper. Therefore ℓ is proper composite since g is invertible. The conditional Bayes risk of ℓ_p is therefore by definition:

$$\begin{aligned}
\underline{L}(\pi) &\doteq \mathsf{E}_{\mathsf{Y}\sim\pi}\ell_p(\mathsf{Y}, \pi) \qquad (18)\\
&= F(\pi) + G(\pi) \qquad (19)
\end{aligned}$$

where $G(\pi) \doteq -\pi F(1) - (1-\pi)F(0)$ is affine. Since a Bregman divergence is invariant by addition of an affine term to its generator (4), we get

$$\ell_p(y^*, c) = D_{-F}(y\|c) \quad (20)$$
$$= D_{-\underline{L}}(y\|c). \quad (21)$$

We now check that if $g = -F'$ then ℓ is proper canonical. It comes from (19) $(-F')^{-1}(z) = (-\underline{L}')^{-1}(z + K)$ where $K \doteq -(F(1) - F(0))$ is a constant, which is still the inverse of the canonical link since it is defined up to multiplication or addition by a scalar (Buja et al., 2005). Hence, if $g = -F'$ then $\ell(y^*, z)$ is proper canonical. Otherwise as previously argued it is proper composite with link g in the more general case. This completes the proof for the implication $\Leftarrow$, and ends the proof of Theorem 1.

Remark: symmetric proper canonical losses (such as the logistic, square or Matsushita losses) admit $\underline{L}(0) = \underline{L}(1)$ Nock & Nielsen (2009, 2008). Hence (19) enforces $\forall \pi \in [0, 1]$

$$\pi(F(0) - F(1)) = \underline{L}(0) = \underline{L}(1) = (1-\pi)(F(1) - F(0)), \quad (22)$$

resulting in $F(1) = F(0)$ and therefore enforcing the constraint $K = 0$ above.

IV Proof of Theorem 5

IV.1 Helper results about BREGMANTRON and FIT

To prove the Theorem, we first show several simple helper results. The first is a simple consequence of the design of u_t. We prove it for the sake of completeness.

Lemma A *Let u_t be the function output by* FIT *in* BREGMANTRON. *Let $z_m \doteq u_t^{-1}(0)$ and $z_m \doteq u_t^{-1}(1)$. Let U_t be defined as in* (14) *(main body, with $u \leftarrow u_t$). The following holds true on u_t*

$$n_{t-1} \cdot (z - z') \leq u_t(z) - u_t(z') \leq N_{t-1} \cdot (z - z') \ , \quad (23)$$
$$\frac{1}{N_{t-1}} \cdot (p - p') \leq u_t^{-1}(p) - u_t^{-1}(p') \leq \frac{1}{n_{t-1}} \cdot (p - p') \ , \quad (24)$$

$\forall z_m \leq z' \leq z \leq z_M$, $\forall 0 \leq p' \leq p \leq 1$, *and the following holds true on U_t:*

$$\frac{(p-p')^2}{2N_{t-1}} \leq D_{U_t^\star}(p\|p') \leq \frac{(p-p')^2}{2n_{t-1}}. \quad (25)$$

Proof We show the right-hand side of ineq. (23). The left hand side of (23) follows by symmetry and ineq (24) follow after a variable change from ineq (23). The proof is a rewriting of the mean-value Theorem for subdifferentials: consider for example the case $u_t(b) - u_t(a) = N'(b - a)$ with $N' > N_{t-1}$ for some $z_m < a < b < z_M$. Let

$$v(z) \doteq u_t(z) - u_t(b) + N'(b - z) \ , \quad (26)$$

and since $v(a) = v(b) = 0$, let $z_* \doteq \arg\min_z v(z)$, assuming wlog that the min exists. Then $v(z) \geq v(z_*)$, and equivalently $u_t(z) - u_t(b) + N'(b - z) \geq u_t(z_*) - u_t(b) + N'(b - z_*)$ ($\forall z \in [a, b]$), which, after reorganising, gives $u_t(z) \geq u_t(z_*) + N'(z - z_*)$, implying $N' \in \partial u_t(z_*)$. Pick now $a \leq z'_* < z_* < z''_* \leq b$ that are linked to z_* by a line segment in u_t. At least one of the two segments has slope $\geq N'$, which is impossible since $N' > N_{t-1}$ and yields a contradiction. The case $a = z_m$ xor $b = z_M$ reduces to a single segment with slope $\geq N'$, also impossible.

We now show (25). Let

$$V(p) \doteq U_t^\star(b) - U_t^\star(p) - (b-p)u_t^{-1}(z) - A(b-p)^2, \quad (27)$$

(remind that $(U_t^\star)' = u_t^{-1}$) where A is chosen so that $V(a) = 0$, which implies[1] since $V(b) = 0$ that $\exists c \in (a, b), 0 \in \partial V(c)$. We have $\partial V(c) \ni -(b-c)c' - 2A(c-b)$ for any $c' \in \partial u_t^{-1}(c)$, implying $A = c'/2$ for some $c' \in \partial u_t^{-1}(c)$. Solving for $V(a) = 0$ yields $D_{U_t^\star}(b\|a) = (c'/2)(b-a)^2$ for some $c' \in \partial u_t^{-1}$ and since $\mathrm{Im}\partial u_t^{-1} \subset [1/N_{t-1}, 1/n_{t-1}]$ from (24), we get

$$\frac{(b-a)^2}{2N_{t-1}} \leq D_{U_t^\star}(b\|a) \leq \frac{(b-a)^2}{2n_{t-1}}, \quad (28)$$

as claimed. ■

Note that we indeed have $\hat{y}_1 \doteq u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_1)$ by the design of Step 4 in BREGMANTRON. The second result we need is a direct consequence of Step 3 in BREGMANTRON.

Lemma B *The following holds for any $t \geq 1, i \in [m]$,*

$$u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) \in u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) \cdot \left[\min\left\{1 - \beta_t, \frac{n_t}{N_t}\right\}, \max\left\{1 + \alpha_t, \frac{N_t}{n_t}\right\}\right], \quad (29)$$

where $\alpha_t, \beta_t \geq 0$ are the stability property parameters at the current iteration of BREGMANTRON*, as defined in Definition 3 (main file).*

Proof We prove the upperbound in (29) by induction. Assuming the property holds for $\boldsymbol{x}_i$ and considering $\boldsymbol{x}_{i+1}$ (recall that indexes are ordered in increasing value of $\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i$, see Step 2 in BREGMANTRON), we obtain

$$\begin{aligned} u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) &\leq u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) + N_t \boldsymbol{w}_{t+1}^\top(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) & (30)\\ &\leq u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) + \frac{N_t}{n_t} \cdot \left(u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)\right) & (31)\\ &= \frac{N_t}{n_t} u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) + u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - \frac{N_t}{n_t} \cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)\ . & (32) \end{aligned}$$

The first inequality comes from the right interval constraint in problem (10) applied to u_{t+1}, ineq. (31) comes from Lemma A applied to u_t. We now have two cases.

Case 1 If $N_t/n_t > 1 + \alpha_t$, using the induction hypothesis (29) yields $u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) \leq (N_t/n_t) \cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)$ and so (32) becomes

$$\begin{aligned} u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) &\leq \frac{N_t}{n_t} u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) + \frac{N_t}{n_t} u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - \frac{N_t}{n_t} u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)\\ &= \frac{N_t}{n_t} u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}). & (33) \end{aligned}$$

[1] This is a simple application of Rolle's Theorem to subdifferentials.

Case 2 If $N_t/n_t \leq 1+\alpha_t$, we have this time from the induction hypothesis $u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) \leq (1+\alpha_t)\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)$, and so we get from (32),

$$
\begin{aligned}
u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) &\leq \frac{N_t}{n_t} u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) + \left(1+\alpha_t - \frac{N_t}{n_t}\right)\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)\\
&\leq \frac{N_t}{n_t} u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) + \left(1+\alpha_t - \frac{N_t}{n_t}\right)\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) \quad (34)\\
&= (1+\alpha_t) u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) \ , \quad (35)
\end{aligned}
$$

where (34) holds because $\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i \leq \boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}$ (by assumption) and u_t is non-decreasing.

The proof of the lowerbound in (29) follows from the following "symmetric" induction, noting first that the second constraint in problem (10) (main file) implies the base case, $u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_1) \geq (1-\beta_t) u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_1)$, and then, for the general index $i > 1$,

$$
\begin{aligned}
u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) &\geq u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) + n_t \boldsymbol{w}_{t+1}^\top (\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) \quad (36)\\
&\geq u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) + \frac{n_t}{N_t}\cdot \big(u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)\big) \quad (37)\\
&= \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) + u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i). \quad (38)
\end{aligned}
$$

The first inequality comes from the left interval constraint in problem (10) applied to u_{t+1}, ineq. (37) comes from Lemma A applied to u_t. Similarly to the upperbound in (29), we now have two cases.

Case 1 If $n_t/N_t \leq 1-\beta_t$, using the induction hypothesis (29) yields $u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) \geq (n_t/N_t)\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)$ and so (38) becomes

$$
\begin{aligned}
u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) &\geq \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) + \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)\\
&= \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}). \quad (39)
\end{aligned}
$$

Case 2 If $n_t/N_t > 1-\beta_t$, using the induction hypothesis (29) yields $u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) \geq (1-\beta_t)\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)$ and so (38) becomes

$$
\begin{aligned}
u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) &\geq \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}) + (1-\beta_t)\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)\\
&\geq \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) + (1-\beta_t)\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - \frac{n_t}{N_t}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) \quad (40)\\
&= (1-\beta_t)\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i). \quad (41)
\end{aligned}
$$

(40) holds because $\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i \leq \boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i+1}$ (by assumption) and u_t is non-decreasing. This achieves the proof of Lemma B. ■

We now analyze the following Bregman loss for $r, t, t' = 1, 2, ...$:

$$
\ell_t^r(S, \boldsymbol{w}_{t'}) \doteq \mathbb{E}_S[D_{U_r^\star}(y \| u_t(\boldsymbol{w}_{t'}^\top \boldsymbol{x}))] = \mathbb{E}_S\left[D_{U_r}(u_r^{-1}\circ u_t(\boldsymbol{w}_{t'}^\top \boldsymbol{x}) \| u_r^{-1}(y))\right] \ , \quad (42)
$$

The key to the proof of Theorem 5 is the following Theorem which breaks down the bound that we have to analyze into several parts.

Theorem C *For any* $t \geq 1$,

$$\ell_{t+1}^{t+1}(S, \boldsymbol{w}_{t+1}) \quad \leq \quad \ell_t^t(S, \boldsymbol{w}_t) - \mathbb{E}_S[D_{U_t}(\boldsymbol{w}_t^\top \boldsymbol{x} \| u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] - L_{t+1} - Q_{t+1},$$

where

$$\begin{aligned} L_{t+1} &\doteq \mathbb{E}_S[(\boldsymbol{w}_t^\top \boldsymbol{x} - u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})) \cdot (u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))], \\ Q_{t+1} &\doteq \mathbb{E}_S[(\boldsymbol{w}_t^\top \boldsymbol{x} - u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})) \cdot (u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - y)] - \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \ell_{t+1}^t(S, \boldsymbol{w}_{t+1}). \end{aligned}$$

Proof We have the following derivations:

$$\begin{aligned} \ell_t^t(S, \boldsymbol{w}_t) &= \mathbb{E}_S[D_{U_t^\star}(y \| u_t(\boldsymbol{w}_t^\top \boldsymbol{x}))] \\ &= \mathbb{E}_S[D_{U_t^\star}(y \| u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] + \mathbb{E}_S[D_{U_t^\star}(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) \| u_t(\boldsymbol{w}_t^\top \boldsymbol{x}))] \\ &\quad + \mathbb{E}_S[((U_t^\star)'(u_t(\boldsymbol{w}_t^\top \boldsymbol{x})) - (U_t^\star)'(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))) \cdot (u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - y)] \quad (43) \\ &= \ell_{t+1}^t(S, \boldsymbol{w}_{t+1}) + \mathbb{E}_S[D_{U_t}(\boldsymbol{w}_t^\top \boldsymbol{x} \| u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] \\ &\quad + \underbrace{\mathbb{E}_S[(\boldsymbol{w}_t^\top \boldsymbol{x} - u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})) \cdot (u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - y)]}_{\doteq \Delta_{t+1}}. \quad (44) \end{aligned}$$

(43) follows from the Bregman triangle equality (3). (44) follows from $(U_t^\star)' \doteq u_t^{-1}$ and (5). Reordering, we get:

$$\ell_{t+1}^t(S, \boldsymbol{w}_{t+1}) \quad = \quad \ell_t^t(S, \boldsymbol{w}_t) - \mathbb{E}_S[D_{U_t}(\boldsymbol{w}_t^\top \boldsymbol{x} \| u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] - \Delta_{t+1}, \quad (45)$$

and we further split Δ_{t+1} in two: $\Delta_{t+1} \doteq F_{t+1} + L_{t+1}$, where

$$\begin{aligned} F_{t+1} &\doteq \mathbb{E}_S[(\boldsymbol{w}_t^\top \boldsymbol{x} - u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})) \cdot (u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - y)], \quad (46) \\ L_{t+1} &\doteq \mathbb{E}_S[(\boldsymbol{w}_t^\top \boldsymbol{x} - u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})) \cdot (u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))]. \quad (47) \end{aligned}$$

We now have the following Lemma.

Lemma D *The following holds for any* $t \geq 0$*:*

$$\ell_{t+1}^{t+1}(S, \boldsymbol{w}_{t+1}) \quad \leq \quad \frac{N_{t-1}}{n_t} \cdot \ell_{t+1}^t(S, \boldsymbol{w}_{t+1}). \quad (48)$$

Proof We use Lemma A and we get:

$$\begin{aligned} D_{U_{t+1}^\star}(p \| p') &\leq \frac{1}{2n_t} \cdot (p - p')^2 \\ &\leq \frac{N_{t-1}}{n_t} \cdot D_{U_t^\star}(p \| p') \ , \quad (49) \end{aligned}$$

from which we just compute the expectation in $\ell_{t+1}(S, \boldsymbol{w}_{t+1})$ and get the result as claimed. ■

Putting altogether (45), (46), (47) and Lemma D yields, $\forall t \geq 1$,

$$\begin{aligned} \ell_{t+1}^{t+1}(S, \boldsymbol{w}_{t+1}) &\leq \ell_{t+1}^t(S, \boldsymbol{w}_{t+1}) + \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \ell_{t+1}^t(S, \boldsymbol{w}_{t+1}) \\ &= \ell_t^t(S, \boldsymbol{w}_t) - \mathbb{E}_S[D_{U_t}(\boldsymbol{w}_t^\top \boldsymbol{x} \| u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] - L_{t+1} \\ &\quad - \left(F_{t+1} - \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \ell_{t+1}^t(S, \boldsymbol{w}_{t+1})\right), \quad (50) \end{aligned}$$

as claimed. This ends the proof of Theorem C. ■

Last, we provide a simple result about the gradient step in Step 1.

Lemma E *Let* $\hat{\boldsymbol{\mu}}_y \doteq \mathbb{E}_S[y \cdot \boldsymbol{x}]$ *and* $\hat{\boldsymbol{\mu}}_t \doteq \mathbb{E}_S[\hat{y}_t \cdot \boldsymbol{x}]$. *The gradient update for* (9) *in Step 1 of the* BREGMANTRON *yields the following update to get* $\boldsymbol{w}_{t+1}$, *for some learning rate* $\eta > 0$*:*

$$\boldsymbol{w}_{t+1} \quad \leftarrow \quad \boldsymbol{w}_t + \eta \cdot (\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t) \ . \tag{51}$$

Proof We trivially have $\nabla_{\boldsymbol{w}} \mathbb{E}_S[D_{U_t}(\boldsymbol{w}^\top \boldsymbol{x} \| u_t^{-1}(y))] = \mathbb{E}_S[u_t(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \boldsymbol{x} - y \cdot \boldsymbol{x}] = \mathbb{E}_S[u_t(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \boldsymbol{x}] - \hat{\boldsymbol{\mu}}_y$, from which we get, for some $\eta > 0$ the gradient update:

$$\boldsymbol{w}_{t+1} \quad \leftarrow \quad \boldsymbol{w}_t - \eta \cdot \nabla_{\boldsymbol{w}} \mathbb{E}_S[D_{U_t}(\boldsymbol{w}^\top \boldsymbol{x} \| u_t^{-1}(y))]_{|\boldsymbol{w}=\boldsymbol{w}_t} = \boldsymbol{w}_t + \eta \cdot (\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t) \ , \tag{52}$$

as claimed. ■

IV.2 Proof of Theorem 5

Lemma F $L_{t+1} \geq 0$, $\forall t$.

Proof We show that the Lemma is a consequence of the fitting of u_{t+1} by FIT from Step 3 in BREGMANTRON. The proof elaborates on the proofsketch of Lemma 2 of Kakade et al. (2011). Denote for short $N_i \doteq N_t \boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i$ and $n_i \doteq n_t \boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i$. We introduce two $(m-1)$-dim vectors of Lagrange multipliers $\boldsymbol{\lambda}_{\mathrm{l}}$ and $\boldsymbol{\lambda}_{\mathrm{r}}$ for the top left and right interval constraints and two multipliers ρ_1 and ρ_m for the additional bounds on $\hat{y}_1$ and $\hat{y}_m$ respectively. This gives the Lagrangian,

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{y}}, S | \boldsymbol{\lambda}_{\mathrm{l}}, \boldsymbol{\lambda}_{\mathrm{r}}, \rho_1, \rho_m) \quad &\doteq \quad \mathbb{E}_{\mathbb{S}}[D_{U_t^\star}(\hat{y} \| y_t)] + \sum_{i=1}^{m-1} \lambda_{\mathrm{l}i} \cdot (\hat{y}_i - \hat{y}_{i+1} + n_{i+1} - n_i) \\ &\quad + \sum_{i=1}^{m-1} \lambda_{\mathrm{r}i} \cdot (\hat{y}_{i+1} - \hat{y}_i - N_{i+1} + N_i) + \rho_1 \cdot -\hat{y}_1 + \rho_m \cdot (\hat{y}_m - 1) \ ,\end{aligned}$$

where we let $q_i \doteq u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)$ for readability and we adopt the convention of Boyd & Vandenberghe (2004, Chapter 5) for constraints. Letting $\boldsymbol{\omega} \in \triangle_m$ (the m-dim probability simplex) denote the weight vector of the examples ins S, we get the following KKT conditions for the optimum:

$$\begin{aligned}
\omega_i(u_t^{-1}(\hat{y}_i) - u_t^{-1}(\hat{y}_{ti})) + \lambda_{\mathrm{l}i} - \lambda_{\mathrm{l}(i-1)} + \lambda_{\mathrm{r}(i-1)} - \lambda_{\mathrm{r}i} &= 0 \ , \forall i = 2, 3, ..., m-1 \ , & (53)\\
\omega_i(u_t^{-1}(\hat{y}_1) - u_t^{-1}(\hat{y}_{t1})) + \lambda_{\mathrm{l}1} - \lambda_{\mathrm{r}1} - \rho_1 &= 0 \ , & (54)\\
\omega_i(u_t^{-1}(\hat{y}_m) - u_t^{-1}(\hat{y}_{tm})) - \lambda_{\mathrm{l}(m-1)} + \lambda_{\mathrm{r}(m-1)} + \rho_m &= 0 \ , & (55)\\
\hat{y}_{i+1} - \hat{y}_i &\in [n_{i+1} - n_i, N_{i+1} - N_i] \ , \forall i \in [m-1] \ , & (56)\\
\hat{y}_1 &\geq 0 \ , & (57)\\
\hat{y}_m &\leq 1 \ , & (58)\\
\lambda_{\mathrm{l}i} \cdot (\hat{y}_i - \hat{y}_{i+1} + n_{i+1} - n_i) &= 0 \ , \forall i \in [m-1] \ , & (59)\\
\lambda_{\mathrm{r}i} \cdot (\hat{y}_{i+1} - \hat{y}_i - N_{i+1} + N_i) &= 0 \ , \forall i \in [m-1] \ , & (60)\\
\rho_1 \cdot -\hat{y}_1 &= 0 \ , & (61)\\
\rho_m \cdot (1 - \hat{y}_m) &= 0 \ , & (62)\\
\boldsymbol{\lambda}_{\mathrm{l}}, \boldsymbol{\lambda}_{\mathrm{r}} &\succeq \mathbf{0} \ , \\
\rho_1, \rho_m &\geq 0 \ .
\end{aligned}$$

For $i = 1, 2, ..., m$, we define

$$\sigma_i \doteq \sum_{j=i}^{m} \omega_j(u_t^{-1}(\hat{y}_{tj}) - u_t^{-1}(\hat{y}_j)).$$

We note that by summing the corresponding subset of (117 — 119), we get

$$\begin{aligned}
\sigma_i &= \lambda_{\mathrm{r}(i-1)} - \lambda_{\mathrm{l}(i-1)} + \rho_m \ , \forall i \in \{2, 3, ..., m\} \ , & (63)\\
\sigma_1 &= -\rho_1 + \rho_m \ . & (64)
\end{aligned}$$

Letting $\hat{y}_0$ and q_0 denote any identical reals, we obtain:

$$\sum_{i=1}^{m} \omega_i(u_t^{-1}(\hat{y}_{ti}) - u_t^{-1}(\hat{y}_i)) \cdot (\hat{y}_i - q_i) = \sum_{i=1}^{m} \sigma_i \cdot ((\hat{y}_i - q_i) - (\hat{y}_{(i-1)} - q_{(i-1)})) \ , \quad (65)$$

which we are going to show is non-negative, which is the statement of the Lemma, in two steps:
Step 1 – We show, for any $i \geq 1$,

$$(\sigma_i - \rho_m) \cdot ((\hat{y}_i - q_i) - (\hat{y}_{(i-1)} - q_{(i-1)})) \geq 0. \quad (66)$$

We have four cases:
Case 1.1 $i > 1$, $\sigma_i - \rho_m > 0$. In this case, $\lambda_{\mathrm{r}(i-1)} > \lambda_{\mathrm{l}(i-1)}$, implying $\lambda_{\mathrm{r}(i-1)} > 0$ and so from eq. (123), $\hat{y}_i - \hat{y}_{(i-1)} - N_i + N_{(i-1)} = 0$, and so $\hat{y}_i - \hat{y}_{(i-1)} = N_t \boldsymbol{w}_{t+1}^\top(\boldsymbol{x}_i - \boldsymbol{x}_{(i-1)})$. Lemma A applied to u_t gives

$$u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{(i-1)}) \leq N_t \boldsymbol{w}_{t+1}^\top(\boldsymbol{x}_i - \boldsymbol{x}_{(i-1)}) \ , \quad (67)$$

and so $\hat{y}_i - \hat{y}_{(i-1)} \geq q_i - q_{(i-1)}$, that is, $(\hat{y}_i - q_i) - (\hat{y}_{(i-1)} - q_{(i-1)}) \geq 0$.
Case 1.2 $i > 1$, $\sigma_i - \rho_m < 0$. In this case, $\lambda_{\mathrm{l}(i-1)} > \lambda_{\mathrm{r}(i-1)}$, implying $\lambda_{\mathrm{l}(i-1)} > 0$, and so from eq.

(122), $\hat{y}_{(i-1)} - \hat{y}_i + n_i - n_{(i-1)} = 0$ and so $\hat{y}_i - \hat{y}_{(i-1)} = n_t \boldsymbol{w}_{t+1}^\top(\boldsymbol{x}_i - \boldsymbol{x}_{(i-1)})$. Lemma A applied to u_t also gives

$$u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{(i-1)}) \quad \geq \quad n_t \boldsymbol{w}_{t+1}^\top(\boldsymbol{x}_i - \boldsymbol{x}_{(i-1)}) \ , \tag{68}$$

and so $q_i - q_{(i-1)} \geq \hat{y}_i - \hat{y}_{(i-1)}$, or, equivalently, $(\hat{y}_i - q_i) - (\hat{y}_{(i-1)} - q_{(i-1)}) \leq 0$.
Case 1.3 $i = 1, \rho_1 > 0$. The case $i = 1$ yields $\sigma_1 - \rho_m = -\rho_1$. It comes from KKT condition (124) that $\hat{y}_1 = 0$, and since $q_1 \geq 0$ (because of FIT), we get $\sigma_1 - \rho_m < 0, \hat{y}_1 - q_1 \leq 0$ and since $\hat{y}_0 = q_0$, we get the statement of (66).
Case 1.4 $i = 1, \rho_1 = 0$. We obtain $\sigma_1 - \rho_m = 0$ and so (66) immediately holds.

Step 2 – We sum (66) for $i \in [m]$, getting

$$\begin{aligned}\sum_{i=1}^m \sigma_i \cdot ((\hat{y}_i - q_i) - (\hat{y}_{(i-1)} - q_{(i-1)})) \quad &\geq \quad \sum_{i=1}^m \rho_m \cdot ((\hat{y}_i - q_i) - (\hat{y}_{(i-1)} - q_{(i-1)})) \\ &= \rho_m \cdot (\hat{y}_m - q_m).\end{aligned} \tag{69}$$

We show that the right-hand side of (69) is non-negative. Indeed, it is immediate if $\rho_m = 0$, and if $\rho_m > 0$, then it comes from KKT condition (125) that $\hat{y}_1 = 1$, and since $q_m \leq 1$ (because of FIT), we get $\rho_m \cdot (\hat{y}_m - q_m) = \rho_m \cdot (1 - q_m) \geq 0$.

To summarize our two steps, we have shown that

$$\sum_{i=1}^m \sigma_i \cdot ((\hat{y}_i - q_i) - (\hat{y}_{(i-1)} - q_{(i-1)})) \quad \geq \quad \rho_m \cdot (\hat{y}_m - q_m) \geq 0, ,$$

which brings from (65) that

$$\mathbb{E}_{\mathcal{S}}[(u_t^{-1}(\hat{y}_t) - u_t^{-1}(\hat{y}_{t+1})) \cdot (\hat{y}_{t+1} - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] \quad \geq \quad 0 \ , \tag{70}$$

which after using the fact that FIT guarantees $\hat{y}_{t+1} = u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}), \forall t$, yields

$$\mathbb{E}_{\mathcal{S}}[(\boldsymbol{w}_t^\top \boldsymbol{x} - u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})) \cdot (u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] \quad \geq \quad 0 \ , \tag{71}$$

which is the statement of Lemma F. ■

We recall

$$\begin{aligned}\hat{\boldsymbol{\mu}}_y \quad &\doteq \quad \mathbb{E}_{\mathcal{S}}[y \cdot \boldsymbol{x}], \qquad (72)\\ \hat{\boldsymbol{\mu}}_t \quad &\doteq \quad \mathbb{E}_{\mathcal{S}}[\hat{y}_t \cdot \boldsymbol{x}], \forall t \geq 1, \qquad (73)\end{aligned}$$

Finally, we let

$$p_t^* \quad \doteq \quad \max\{\mathbb{E}_S[y], \mathbb{E}_S[u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})]\} \ \ (\in [0, 1]). \tag{74}$$

Lemma G *Fix any lowerbound $\delta_t > 0$ such that*

$$\frac{\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2}{X} \geq 2\sqrt{p_t^* \delta_t}. \quad (75)$$

Fix any γ_t satisfying:

$$\gamma_t \in \left[0, \sqrt{\frac{\delta_t}{2(2+\delta_t)}}\right], \quad (76)$$

and learning rate

$$\eta = \frac{1-\gamma_t}{2N_t X^2} \cdot \left(1 - \frac{\delta_t(2+\delta_t)}{(1+\delta_t)^2} \cdot \frac{p_t^* X}{\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2}\right). \quad (77)$$

Suppose $\alpha_t, \beta_t \leq \delta_t/(1+\delta_t)$ and

$$\frac{N_t}{n_t}, \frac{N_{t-1}}{n_t} \leq 1 + \frac{\delta_t}{1+\delta_t}. \quad (78)$$

Then

$$F_{t+1} \geq \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \frac{p_t^*}{2n_t} + \frac{p_t^* \delta_t}{n_t(1+\delta_t)}. \quad (79)$$

Remark: it can be shown from (75) (see also 109) that η belongs to the following interval:

$$\eta \in \frac{1-\gamma_t}{2N_t X^2} \cdot \left(1 - \frac{\sqrt{\delta_t p_t^*}(2+\delta_t)}{2(1+\delta_t)^2} \cdot \left[\sqrt{\delta_t p_t^*}, 1\right]\right).$$

Also, since $\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \leq 2X$, (75) implies

$$\delta_t \leq \frac{1}{p_t^*}. \quad (80)$$

Proof The following two facts are consequences of Lemmata E, A and the continuity of u_t: $\forall i \in [m]$,

$$\begin{aligned}
\exists p_i \in [N^{-1}, n^{-1}] : u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) &= u_t^{-1} \circ u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) \\
&\quad + p_i \cdot (u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)) \\
&= \boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i + p_i \cdot (u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i)) \quad (81) \\
\exists r_i \in [n, N] : u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_i) &= u_t(\boldsymbol{w}_t^\top \boldsymbol{x}_i) + r_i \cdot (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top \boldsymbol{x}_i \\
&= u_t(\boldsymbol{w}_t^\top \boldsymbol{x}_i) + \eta r_i \cdot (\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{x}_i \ . \quad (82)
\end{aligned}$$

Folding (81) and (82) in F_{t+1}, we get:

$$\begin{aligned}
F_{t+1} &= \mathbb{E}_{\mathcal{S}}\left[(u_t^{-1}\circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - \boldsymbol{w}_t^\top \boldsymbol{x})\cdot(y - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\left\{\begin{array}{l} (\boldsymbol{w}_{t+1}^\top \boldsymbol{x} - \boldsymbol{w}_t^\top \boldsymbol{x} + p\cdot(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))) \\ \cdot (y - u_t(\boldsymbol{w}_t^\top \boldsymbol{x}) - \eta r\cdot(\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{x}) \end{array}\right.\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\left\{\begin{array}{l} (\eta\cdot(\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{x} + p\cdot(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))) \\ \cdot (y - u_t(\boldsymbol{w}_t^\top \boldsymbol{x}) - \eta r\cdot(\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{x}) \end{array}\right.\right] \quad (83) \\
&= \eta\cdot(\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \mathbb{E}_{\mathcal{S}}\left[y\cdot\boldsymbol{x} - u_t(\boldsymbol{w}_t^\top \boldsymbol{x})\cdot\boldsymbol{x}\right] \\
&\quad + \mathbb{E}_{\mathcal{S}}\left[p\cdot(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))\cdot(y - u_t(\boldsymbol{w}_t^\top \boldsymbol{x}))\right] \\
&\quad - \eta^2\cdot\mathbb{E}_{\mathcal{S}}\left[r\cdot((\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{x})^2\right] \\
&\quad - \eta\cdot(\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \mathbb{E}_{\mathcal{S}}\left[pr(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))\cdot\boldsymbol{x}\right] \\
&= \eta\cdot\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2 + \underbrace{\mathbb{E}_{\mathcal{S}}\left[p\cdot(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))\cdot(y - u_t(\boldsymbol{w}_t^\top \boldsymbol{x}))\right]}_{\doteq A} \\
&\quad - \underbrace{\eta^2\cdot\mathbb{E}_{\mathcal{S}}\left[r\cdot((\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{x})^2\right]}_{\doteq B} \\
&\quad - \underbrace{\eta\cdot(\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t)^\top \mathbb{E}_{\mathcal{S}}\left[pr(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))\cdot\boldsymbol{x}\right]}_{\doteq C}\ . \quad (84)
\end{aligned}$$

We now bound lowerbound A and upperbound B, C. Lemma B brings

$$\min\left\{-\beta_t, \frac{n_t}{N_t} - 1\right\}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) \leq u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}), \quad (85)$$

and

$$u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) \leq \max\left\{\alpha_t, \frac{N_t}{n_t} - 1\right\}\cdot u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}), \quad (86)$$

and so we get

$$\begin{aligned}
A &\doteq \mathbb{E}_{\mathcal{S}}\left[p\cdot(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))\cdot(y - u_t(\boldsymbol{w}_t^\top \boldsymbol{x}))\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[py\cdot(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}}\left[pu_t(\boldsymbol{w}_t^\top \boldsymbol{x})\cdot(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))\right] \\
&\geq -\frac{1}{n_t}\cdot\max\left\{\beta_t, 1 - \frac{n_t}{N_t}\right\}\mathbb{E}_{\mathcal{S}}\left[u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})\right] - \frac{1}{n_t}\cdot\max\left\{\alpha_t, \frac{N_t}{n_t} - 1\right\}\mathbb{E}_{\mathcal{S}}\left[u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})u_t(\boldsymbol{w}_t^\top \boldsymbol{x})\right] \\
&\geq -\frac{1}{n_t}\cdot\max\left\{\alpha_t, \beta_t, 1 - \frac{n_t}{N_t}, \frac{N_t}{n_t} - 1\right\}\mathbb{E}_{\mathcal{S}}\left[u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})\right] \\
&\geq -\frac{1}{n_t}\cdot\max\left\{\alpha_t, \beta_t, \frac{N_t}{n_t} - 1\right\}p_t^*, \quad (87)
\end{aligned}$$

since $y \in \{0, 1\}$, $U_t \leq 1$ and $1 - (1/z) \leq z - 1$ for $z \geq 0$. Cauchy-Schwartz inequality and (82) yield

$$\begin{aligned}
B &\leq \eta^2\cdot\mathbb{E}_{\mathcal{S}}\left[r\cdot\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2\|\boldsymbol{x}\|_2^2\right] \\
&\leq \eta^2 N X^2\cdot\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2. \quad (88)
\end{aligned}$$

We also have successively because of Cauchy-Schwartz inequality, the triangle inequality, Lemma A and (86)

$$\begin{aligned}
C &\leq \eta \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \cdot \|\mathbb{E}_{\mathcal{S}}\left[pr(u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})) \cdot \boldsymbol{x}\right]\|_2 \\
&\leq \eta \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \cdot \mathbb{E}_{\mathcal{S}}[pr|u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})| \cdot \|\boldsymbol{x}\|_2] \\
&\leq \frac{\eta N_t}{n_t} \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \cdot \mathbb{E}_{\mathcal{S}}[|u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}) - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})| \cdot \|\boldsymbol{x}\|_2] \\
&\leq \frac{\eta N_t \max\left\{\alpha_t, \frac{N_t}{n_t} - 1\right\} X}{n_t} \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \cdot \mathbb{E}_{\mathcal{S}}\left[u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})\right] \\
&\leq \frac{\eta N_t \max\left\{\alpha_t, \frac{N_t}{n_t} - 1\right\} X p_t^*}{n_t} \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \ .
\end{aligned} \tag{89}$$

We thus get

$$\begin{aligned}
& F_{t+1} - \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \frac{p_t^*}{n_t} \\
&\geq \eta \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2 - \frac{1}{n_t} \cdot \max\left\{\alpha_t, \beta_t, \frac{N_t}{n_t} - 1\right\} p_t^* - \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \frac{p_t^*}{n_t} \\
&\quad -\eta^2 N_t X^2 \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2 - \frac{\eta N_t \max\left\{\alpha_t, \frac{N_t}{n_t} - 1\right\} X p_t^*}{n_t} \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \\
&\geq \eta \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2 - \frac{2\max\left\{\alpha_t, \beta_t, \frac{N_t}{n_t} - 1, \frac{N_{t-1}}{n_t} - 1\right\} p_t^*}{n_t} - \eta^2 N_t X^2 \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2 \\
&\quad - \frac{\eta N_t \max\left\{\alpha_t, \frac{N_t}{n_t} - 1\right\} X p_t^*}{n_t} \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \\
&\quad = \eta \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2 - \frac{2\max\left\{\alpha_t, \beta_t, \frac{N_t}{n_t} - 1, \frac{N_{t-1}}{n_t} - 1\right\} p_t^*}{n_t} - \eta^2 N_t X^2 \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2 \\
&\quad - \frac{\eta N_t \max\left\{n_t\alpha_t, N_t - n_t\right\} X p_t^*}{n_t^2} \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \\
&= \tilde{\eta} \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 - \frac{2\max\left\{\alpha_t, \beta_t, \frac{N_t}{n_t} - 1, \frac{N_{t-1}}{n_t} - 1\right\} p_t^*}{n_t} - \tilde{\eta}^2 N_t X^2 \\
&\quad - \frac{\tilde{\eta} N_t \max\left\{n_t\alpha_t, N_t - n_t\right\} X p_t^*}{n^2} \\
&\geq \underbrace{-a\tilde{\eta}^2 + b\tilde{\eta} + c}_{\doteq J(\tilde{\eta})},
\end{aligned} \tag{90}$$

with $\tilde{\eta} \doteq \eta \cdot \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2$ and:

$$a \doteq N_t X^2, \tag{91}$$

$$b \doteq \|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 - \varepsilon_t(1 + \varepsilon_t) \cdot p_t^* X, \tag{92}$$

$$c \doteq -\frac{2\varepsilon_t p_t^*}{n_t}, \tag{93}$$

where ε_t is any real satisfying

$$\varepsilon_t \geq \max\left\{\alpha_t, \beta_t, \frac{N_t}{n_t} - 1, \frac{N_{t-1}}{n_t} - 1\right\}. \quad (94)$$

Remark that

$$2\sqrt{a(1+\varepsilon_t)\cdot -c} = 2\sqrt{2\varepsilon_t}(1+\varepsilon_t)\sqrt{p_t^*}X,$$

so if we can guarantee that $b^2 \geq 4a(1+\varepsilon_t)\cdot -c$, then fixing $\tilde{\eta} \doteq (1-\gamma_t)b/(2a)$ for some $\gamma_t \in [0,1]$ yields from (90)

$$\begin{aligned} J(\tilde{\eta}) &= \frac{b^2(1-\gamma_t^2)}{4a} + c \\ &\geq -\varepsilon_t c + \gamma_t^2(1+\varepsilon_t)c \end{aligned} \quad (95)$$

The condition on b is implied by the following one, since $p_t^* \leq 1$:

$$\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \geq 2\sqrt{2\varepsilon_t}(1+\varepsilon_t)\sqrt{p_t^*}X + \varepsilon_t(1+\varepsilon_t)\sqrt{p_t^*}X. \quad (96)$$

Fix any $K_t > 1$. It is easy to check that for any

$$\varepsilon_t \leq \sqrt{K_t} - 1, \quad (97)$$

we have $\varepsilon_t \leq 2(\sqrt{K_t} - \sqrt{2})\sqrt{\varepsilon_t}$, so a sufficient condition to get (96) is

$$\sqrt{\varepsilon_t}(1+\varepsilon_t) \leq \frac{\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2}{2\sqrt{K_t}\sqrt{p_t^*}X}. \quad (98)$$

Letting $f(z) \doteq \sqrt{z}(1+z)$, it is not hard to check that if we pick $z = \min\{\sqrt{K_t} - 1, u^2/K_t\}$ then $f(z) \leq u$: indeed,

- if the $\min$ is u^2/K_t, implying $u \leq \sqrt{K_t(\sqrt{K_t} - 1)}$, then $f(z)$ being increasing we observe $f(z) \leq f(u^2/K_t) \leq u$, which simplifies for the rightmost inequality into $u \leq \sqrt{K_t(\sqrt{K_t} - 1)}$, which is our assumption;

- if the $\min$ is $\sqrt{K_t} - 1$, implying $u \geq \sqrt{K_t(\sqrt{K_t} - 1)}$, then this time we directly get $f(z) = \sqrt{\sqrt{K_t} - 1}(1 + \sqrt{K_t} - 1) = \sqrt{K_t(\sqrt{K_t} - 1)} \leq u$, as claimed.

To summarize, if we pick

$$\varepsilon_t \doteq \min\left\{\sqrt{K_t} - 1, \frac{\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2}{4K_t^2 p_t^* X^2}\right\}, \quad (99)$$

then we check that our precondition (97) holds and we obtain from (90) and (95),

$$F_{t+1} - \left(\frac{N_{t-1}}{n_t} - 1\right)\cdot\frac{p_t^*}{n} \geq \frac{2\varepsilon_t^2 p_t^*}{n_t} - \frac{2\gamma_t^2\varepsilon_t(1+\varepsilon_t)p_t^*}{n_t}. \quad (100)$$

Suppose γ_t satisfies

$$(1+\varepsilon_t)\gamma_t^2 \quad \leq \quad \frac{\varepsilon_t}{2}. \tag{101}$$

In this case, we further lowerbound (100) as

$$\begin{aligned} F_{t+1} - \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \frac{p_t^*}{n} \quad &\geq \quad \frac{\varepsilon_t^2 p_t^*}{n_t} \\ &= \frac{p_t^*}{n_t} \cdot \left(\min\left\{\sqrt{K_t} - 1, \frac{\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2}{4K_t^2 p_t^* X^2}\right\}\right)^2. \end{aligned} \tag{102}$$

To simplify this bound and make it more readable, suppose we fix a lowerbound

$$\frac{\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2^2}{4p_t^* X^2} \quad \geq \quad \delta_t, \tag{103}$$

for some $\delta_t > 0$. Some simple calculation shows that if we pick

$$K_t \quad \doteq \quad \left(1 + \frac{\delta_t}{1+\delta_t}\right)^2, \tag{104}$$

then the $\min$ in (102) is achieved in $\sqrt{K_t} - 1$, which therefore guarantees

$$F_{t+1} - \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \frac{p_t^*}{n_t} \quad \geq \quad \frac{p_t^* \delta_t}{n_t(1+\delta_t)}, \tag{105}$$

and therefore gives the choice $\varepsilon_t = \delta_t/(1+\delta_t)$. The constraint on γ_t from (101) becomes

$$\gamma_t \quad \leq \quad \sqrt{\frac{\delta_t}{2(2+\delta_t)}}, \tag{106}$$

and it comes from (94) that $\alpha_t, \beta_t \leq \delta_t/(1+\delta_t)$ and

$$\frac{N_t}{n_t}, \frac{N_{t-1}}{n_t} \quad \leq \quad 1 + \frac{\delta_t}{1+\delta_t}, \tag{107}$$

as claimed. This ends the proof of Lemma G, after having remarked that the learning rate η is then fixed to be (from (90))

$$\begin{aligned} \eta \quad &\doteq \quad \frac{\tilde{\eta}}{\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2} \\ &= \quad \frac{1-\gamma_t}{2\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 N_t X^2} \cdot \left(\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 - \frac{\delta_t(2+\delta_t)}{(1+\delta_t)^2} p_t^* X\right) \\ &= \quad \frac{1-\gamma_t}{2N_t X^2} \cdot \left(1 - \frac{\delta_t(2+\delta_t)}{(1+\delta_t)^2} \cdot \frac{p_t^* X}{\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2}\right), \end{aligned} \tag{108}$$

and it satisfies, because of (103),

$$\eta \quad \geq \quad \frac{1-\gamma_t}{2N_t X^2} \cdot \left(1 - \frac{\sqrt{\delta_t p_t^*}(2+\delta_t)}{2(1+\delta_t)^2}\right) \tag{109}$$

and since $\|\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_t\|_2 \leq 2X$,

$$\eta \quad \leq \quad \frac{1-\gamma_t}{2N_t X^2} \cdot \left(1 - \frac{\delta_t p_t^*(2+\delta_t)}{2(1+\delta_t)^2}\right) \tag{110}$$

(we note that (103) implies $\delta_t p_t^* \leq 1$) This ends the proof of Lemma G. ■

We now show a lowerbound on Q_{t+1} in Theorem C.

Lemma H *Suppose the setting of Lemma G holds. Then*

$$Q_{t+1} \quad \geq \quad \frac{p_t^* \delta_t}{n_t(1+\delta_t)}. \tag{111}$$

Proof Remind that it comes from Theorem C

$$Q_{t+1} \quad \doteq \quad F_{t+1} - \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \ell_{t+1}^t(S, \boldsymbol{w}_{t+1}).$$

We have using Lemma A,

$$\begin{aligned}
\ell_{t+1}^t(S, \boldsymbol{w}_{t+1}) \quad & \doteq \quad \mathbb{E}_S[D_{U_t^\star}(y \| u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] \\
& \leq \quad \frac{1}{2n_t} \cdot \mathbb{E}_S[(y - u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))^2] \\
& \quad = \frac{1}{2n_t} \cdot (\mathbb{E}_S[y] - 2\mathbb{E}_S[y u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})] + \mathbb{E}_S[u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})^2]) \\
& \leq \quad \frac{\mathbb{E}_S[y] + \mathbb{E}_S[u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x})]}{2n_t} \\
& \leq \quad \frac{p_t^*}{n_t},
\end{aligned} \tag{112}$$

because $u_t(z) \leq 1$. We get

$$\left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \ell_{t+1}^t(S, \boldsymbol{w}_{t+1}) \quad \leq \quad \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \frac{p_t^*}{n_t}, \tag{113}$$

so using Lemma G, we get

$$\begin{aligned}
Q_{t+1} \quad & \geq \quad F_{t+1} - \left(\frac{N_{t-1}}{n_t} - 1\right) \cdot \frac{p_t^*}{n_t} \\
& \geq \quad \frac{p_t^* \delta_t}{n_t(1+\delta_t)},
\end{aligned} \tag{114}$$

as claimed. ■

Remind from Theorem C that

$$\ell_{t+1}^{t+1}(S, \boldsymbol{w}_{t+1}) \quad \leq \quad \ell_t^t(S, \boldsymbol{w}_t) - \mathbb{E}_S[D_{U_t}(\boldsymbol{w}_t^\top \boldsymbol{x} \| u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] - L_{t+1} - Q_{t+1},$$

and we know that

- $\mathbb{E}_S[D_{U_t}(\boldsymbol{w}_t^\top \boldsymbol{x} \| u_t^{-1} \circ u_{t+1}(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}))] \geq 0$, because a Bregman divergence cannot be negative;
- $L_{t+1} \geq 0$ from Lemma F;
- $Q_{t+1} \geq p_t^* \delta_t / (n_t(1+\delta_t))$ from Lemma H (assuming the conditions of Lemma G).

Putting this altogether, we get

$$\ell_{t+1}^{t+1}(S, \boldsymbol{w}_{t+1}) \quad \leq \quad \ell_t^t(S, \boldsymbol{w}_t) - \frac{p_t^* \delta_t}{n_t(1+\delta_t)},$$

which then easily translates into the statement of Theorem 5.

V Proof of Corollary 6

To make things explicit, we replace Step 3 in the BREGMANTRON by the following new Step 3:

Step 3 fit $\hat{\boldsymbol{y}}_{t+1}$ by solving for global optimum:

$$\hat{\boldsymbol{y}}_{t+1} \doteq \arg\min_{\hat{\boldsymbol{y}}} \mathbb{E}_S[D_{U_t^\star}(\hat{y} \| \hat{\boldsymbol{y}}_t)] \quad \text{//proper composite fitting of } \hat{\boldsymbol{y}}_{t+1} \text{ given } \boldsymbol{w}_{t+1}, u_t$$

$$\text{s.t.} \left\{ \begin{array}{l} \hat{y}_{i+1} - \hat{y}_i \in [n_t \cdot (\boldsymbol{w}_{t+1}^\top(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)), N_t \cdot (\boldsymbol{w}_{t+1}^\top(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i))] \ , \forall i \in [m-1] \\ \hat{y}_1 \in [(1-\beta_t) u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_1), (1+\alpha_t) u_t(\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_1)] \\ \hat{y}_m \leq 1 \end{array} \right. \quad (115)$$

The only step that needs update in the proof of Theorem 5 is Lemma F. We now show that the property still holds for this new Step 3.

Lemma I *$L_{t+1} \geq 0$, $\forall t$.*

Proof The proof proceeds from the same steps as for Lemma F. We reuse the same notations. This time, we get the Lagrangian,

$$\begin{aligned} \mathcal{L}(\hat{\boldsymbol{y}}, S | \boldsymbol{\lambda}_{\mathrm{l}}, \boldsymbol{\lambda}_{\mathrm{r}}, \rho_1, \rho_m) \quad \doteq \quad & \mathbb{E}_{\mathcal{S}}[D_{U_t^\star}(\hat{y} \| y_t)] + \sum_{i=1}^{m-1} \lambda_{\mathrm{l}i} \cdot (\hat{y}_i - \hat{y}_{i+1} + n_{i+1} - n_i) \\ & + \sum_{i=1}^{m-1} \lambda_{\mathrm{r}i} \cdot (\hat{y}_{i+1} - \hat{y}_i - N_{i+1} + N_i) + \rho_1 \cdot ((1-\beta_t) q_1 - \hat{y}_1) \\ & + \rho_1' \cdot (\hat{y}_1 - (1+\alpha_t) q_1) + \rho_m \cdot (\hat{y}_m - 1) \ , \end{aligned} \quad (116)$$

and the following KKT conditions for the optimum:

$$
\begin{aligned}
\omega_i(u_t^{-1}(\hat{y}_i) - u_t^{-1}(\hat{y}_{ti})) + \lambda_{\mathrm{l}i} - \lambda_{\mathrm{l}i-1} + \lambda_{\mathrm{r}i-1} - \lambda_{\mathrm{r}i} &= 0 \ , \forall i = 2, 3, ..., m-1 \ , & (117)\\
\omega_i(u_t^{-1}(\hat{y}_1) - u_t^{-1}(\hat{y}_{t1})) + \lambda_{\mathrm{l}1} - \lambda_{\mathrm{r}1} - \rho_1 + \rho_1' &= 0 \ , & (118)\\
\omega_i(u_t^{-1}(\hat{y}_m) - u_t^{-1}(\hat{y}_{tm})) - \lambda_{\mathrm{l}m-1} + \lambda_{\mathrm{r}m-1} - \rho_m &= 0 \ , & (119)\\
\hat{y}_{i+1} - \hat{y}_i &\in [n_{i+1} - n_i, N_{i+1} - N_i] \ , \forall i \in [m-1] & (120)\\
\hat{y}_1 &\in q_1 \cdot [1 - \beta_t, 1 + \alpha_t] \ , & (121)\\
\lambda_{\mathrm{l}i} \cdot (\hat{y}_i - \hat{y}_{i+1} + n_{i+1} - n_i) &= 0 \ , \forall i \in [m-1] \ , & (122)\\
\lambda_{\mathrm{r}i} \cdot (\hat{y}_{i+1} - \hat{y}_i - N_{i+1} + N_i) &= 0 \ , \forall i \in [m-1] \ , & (123)\\
\rho_1 \cdot ((1-\beta)q_1 - \hat{y}_1) &= 0 \ , & (124)\\
\rho_1' \cdot (\hat{y}_1 - (1+\alpha)q_1) &= 0 \ , & (125)\\
\rho_m \cdot (\hat{y}_m - 1) &= 0 \ , & (126)\\
\boldsymbol{\lambda}_{\mathrm{l}}, \boldsymbol{\lambda}_{\mathrm{r}} &\succeq \mathbf{0} \ , & \\
\rho_1, \rho_1', \rho_m &\geq 0 \ . & (127)
\end{aligned}
$$

Letting again $\sigma_i \doteq \sum_{j=i}^m \omega_j (u_t^{-1}(\hat{y}_{tj}) - u_t^{-1}(\hat{y}_j))$ (for $i = 1, 2, ..., m$) and $\hat{y}_0$ and q_0 any identical reals, we obtain this time:

$$
\begin{aligned}
\sigma_i &= \lambda_{\mathrm{r}i-1} - \lambda_{\mathrm{l}i-1} + \rho_m \ , \forall i \in \{2, 3, ..., m\} \ , & (128)\\
\sigma_1 &= -\rho_1 + \rho_1' + \rho_m \ . & (129)
\end{aligned}
$$

We now remark that just like in (66), we still get

$$
(\sigma_i - \rho_m) \cdot ((\hat{y}_i - q_i) - (\hat{y}_{(i-1)} - q_{(i-1)})) \ \ \geq \ \ 0, \forall i > 1, \qquad (130)
$$

since the expression of the corresponding σs does not change. The proof changes for σ_1 as this time,

$$
(\sigma_1 - \rho_m) \cdot ((\hat{y}_1 - q_1) - (\hat{y}_0 - q_0)) \ \ = \ \ (-\rho_1 + \rho_1') \cdot (\hat{y}_1 - q_1), \qquad (131)
$$

and we have the following possibilities:

- suppose $\rho_1 > 0$. In this case, KKT condition (124) implies $\hat{y}_1 = (1 - \beta_t)q_1$, implying $\hat{y}_1 - q_1 = -\beta_t q_1 \leq 0$, and also $\hat{y}_1 \neq (1 + \alpha_t)q_1$, implying from KKT condition (125) $\rho_1' = 0$, which gives us $(-\rho_1 + \rho_1') \cdot (\hat{y}_1 - q_1) = -\rho_1 \cdot (\hat{y}_1 - q_1) \geq 0$.

- suppose $\rho_1' > 0$. In this case, the KKT condition (125) implies $\hat{y}_1 = (1 + \alpha)q_1$ and so $\hat{y}_1 - q_1 = \alpha q_1 \geq 0$, but also so $\hat{y}_1 \neq (1-\beta)q_1$, so $\rho_1 = 0$, which gives us $(-\rho_1 + \rho_1') \cdot (\hat{y}_1 - q_1) = \rho_1' \cdot (\hat{y}_1 - q_1) \geq 0$.

- If both $\rho_1 = \rho_1' = 0$, we note $(-\rho_1 + \rho_1') \cdot (\hat{y}_1 - q_1) = 0$,

and so (66) also holds for $i = 1$, which allows us to conclude in the same way as we did for Lemma F, and ends the proof of Lemma I. ■

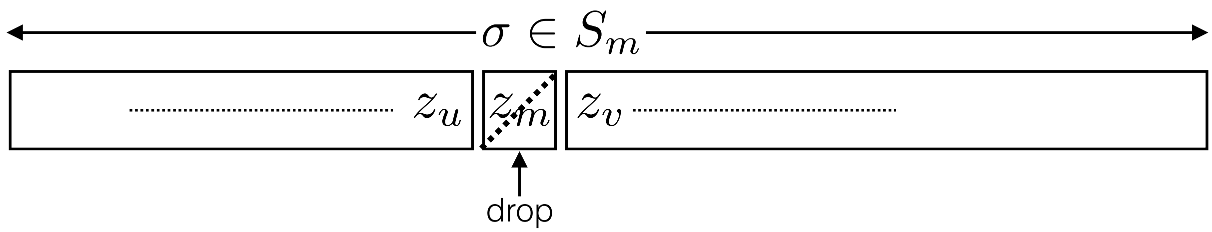

Figure 1: Crafting from $\sigma \in S_m$ a subset of $m-1$ reals for which the induction hypothesis can be applied in the proof of Lemma 7 (see text).

VI Proof of Lemma 7

Let us drop the iteration index, thus letting $z_i \doteq z_{Ti}$ for $i = 0, 1, ..., m+1$ (with $z_0 \doteq z_{T_{\text{MIN}}}$ and $z_{m+1} \doteq z_{T_{\text{MAX}}}$). We thus have $z_i \leq z_{i+1}, \forall i$. We now pick one specific element in $\mathcal{U}(\boldsymbol{w}, S)$, such that

$$u(z_i) \quad = \quad (-\underline{L}')^{-1}(z_i), \tag{132}$$

for $i \in [d]$, which complies with the definition of $\mathcal{U}$ as both u and $(-\underline{L}')^{-1}$ are non decreasing. We then have

$$\begin{aligned} \int_{z_1}^{z_m} |(-\underline{L}')^{-1}(z) - u(z)|\mathrm{d}z &= \sum_{i=1}^{m-1} \int_{z_i}^{z_{i+1}} |(-\underline{L}')^{-1}(z) - u(z)|\mathrm{d}z \\ &\leq \sum_{i=1}^{m-1} (u(z_{i+1}) - u(z_i))(z_{i+1} - z_i) \\ &\leq N \sum_{i=1}^{m-1} (z_{i+1} - z_i)^2, \end{aligned} \tag{133}$$

where the first inequality holds because of (132) and u is non decreasing, and the second inequality holds because of the constraint in Step 3. Let $S_m \ni \sigma : [m] \rightarrow [m]$ be a permutation of the indices. We now show

$$\sum_{i=1}^{m-1} (z_{\sigma(i+1)} - z_{\sigma(i)})^2 \quad \geq \quad \sum_{i=1}^{m-1} (z_{i+1} - z_i)^2, \forall m > 1, \forall \sigma \in S_m. \tag{134}$$

We show this by induction on m. The result is trivially true for $m = 2$. Considering any $m > 2$ and any permutation $\sigma \in S_m$, suppose the order of the zs in the permutation is as in Figure 1. Let $\Sigma_{\text{tot}} \doteq \sum_{i=1}^{m-1} (z_{\sigma(i+1)} - z_{\sigma(i)})^2$, which therefore includes term $(z_m - z_u)^2 + (z_v - z_m)^2$. Now, drop z_m. This gives us a partial sum, Σ_{partial}, over $\{z_1, z_2, ..., z_{m-1}\}$ described by a permutation $\sigma \in S_{m-1}$ for which the induction hypothesis applies. We then have two cases:

Case 1: $1 < \sigma(m) < m$, which implies that z_m is "inside" the ordering given by σ and is in fact the case depicted in Figure 1. In this case and using notations from Figure 1, we get:

$$\Sigma_{\text{tot}} \quad = \quad \Sigma_{\text{partial}} + (z_m - z_u)^2 + (z_v - z_m)^2 - (z_v - z_u)^2, \tag{135}$$

and the induction hypothesis yields

$$\Sigma_{\text{partial}} \;\geq\; \sum_{i=1}^{m-2}(z_{i+1}-z_i)^2. \tag{136}$$

So to show (134) we just need to show

$$\underbrace{\sum_{i=1}^{m-2}(z_{i+1}-z_i)^2+(z_m-z_u)^2+(z_v-z_m)^2-(z_v-z_u)^2}_{\text{Lowerbound on }\Sigma_{\text{tot}}\text{ from (135) and (140)}} \;\geq\; \sum_{i=1}^{m-1}(z_{i+1}-z_i)^2, \tag{137}$$

which equivalently gives

$$(z_m-z_u)^2+(z_m-z_v)^2 \;\geq\; (z_v-z_u)^2+(z_m-z_{m-1})^2. \tag{138}$$

After putting $(z_v-z_u)^2$ in the LHS and simplifying, we get equivalently that the induction holds if

$$2z_m^2-2z_mz_u-2z_mz_v+2z_vz_u \;\geq\; (z_m-z_{m-1})^2. \tag{139}$$

The LHS factorizes conveniently as $2z_m^2-2z_mz_u-2z_mz_v+2z_vz_u = 2(z_m-z_u)(z_m-z_v)$. Since by hypothesis $z_1 \leq z_2 ... \leq z_{m-1} \leq z_m$, we get $2(z_m-z_u)(z_m-z_v) \geq 2(z_m-z_{m-1})^2$, which implies (139) holds and the induction is proven.

Case 2: $\sigma(m)=m$ (the case $\sigma(m)=1$ give the same proof). In this case, z_m is at the "right" of the permutation's ordering. Using notations from Figure 1, we get in lieu of (135),

$$\Sigma_{\text{tot}} \;=\; \Sigma_{\text{partial}}+(z_m-z_u)^2, \tag{140}$$

and leaves us with the following result to show:

$$\sum_{i=1}^{m-2}(z_{i+1}-z_i)^2+(z_m-z_u)^2 \;\geq\; \sum_{i=1}^{m-1}(z_{i+1}-z_i)^2, \tag{141}$$

which simplifies in $(z_m-z_u)^2 \geq (z_m-z_{m-1})^2$, which is true by assumption ($z_u \leq z_{m-1} \leq z_m$).

To summarize, we have shown that $\forall \sigma : [m] \rightarrow [m]$,

$$\int_{z_1}^{z_m} |(-\underline{L}')^{-1}(z)-u(z)|\mathrm{d}z \;\leq\; \sum_{i=1}^{m-1}(z_{\sigma(i+1)}-z_{\sigma(i)})^2. \tag{142}$$

Assuming the ε-NN graph is 2-vertex-connected, we square the graph. Because of the triangle inequality on norm $\|.\|$, every edge has now length at most 2ε and the graph is Hamiltonian, a result known as Fleischner's Theorem (Fleischner, 1974), (Gross & Yellen, 2004, p. 265, F17). Consider any Hamiltonian path and the permutation $\boldsymbol{\sigma}$ of $[m]$ it induces. We thus get $\|\boldsymbol{x}_{\sigma(i+1)}-\boldsymbol{x}_{\sigma(i)}\| \leq 2\varepsilon, \forall i$, and so Cauchy-Schwarz inequality yields:

$$\begin{aligned}\sum_{i=1}^{m-1}(z_{\sigma(i+1)}-z_{\sigma(i)})^2 &\doteq \sum_{i=1}^{m-1}(\boldsymbol{w}^\top\boldsymbol{x}_{\sigma(i+1)}-\boldsymbol{w}^\top\boldsymbol{x}_{\sigma(i)})^2\\ &\leq \|\boldsymbol{w}\|_*^2\sum_{i=1}^{m-1}\|\boldsymbol{x}_{\sigma(i+1)}-\boldsymbol{x}_{\sigma(i)}\|^2\\ &\leq 2m\varepsilon^2\cdot\|\boldsymbol{w}\|_*^2,\end{aligned} \tag{143}$$

as claimed, where $\|.\|_*$ is the dual norm of $\|.\|$. We assemble (133) and (143) and get:

$$\int_{z_1}^{z_m} |(-\underline{L}')^{-1}(z) - u(z)|\mathrm{d}z \quad \leq \quad 2Nm\varepsilon^2 \cdot \|\boldsymbol{w}\|_*^2,$$

which is the statement of the Lemma.

Remark: had we measured the ℓ_1 discrepancy using the loss and not its link (and adding a second order differentiability condition), we could have used the fact that a Bregman divergence between two points is proportional to the square loss to get a result similar to the Lemma (see Section II).

References

Banerjee, A., Merugu, S., Dhillon, I., and Ghosh, J. Clustering with bregman divergences. In *Proc. of the* 4^{th} *SIAM International Conference on Data Mining*, pp. 234–245, 2004.

Boissonnat, J.-D., Nielsen, F., and Nock, R. Bregman voronoi diagrams. *DCG*, 44(2):281–307, 2010.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Buja, A., Stuetzle, W., and Shen, Y. Loss functions for binary class probability estimation ans classification: structure and applications, 2005. Technical Report, University of Pennsylvania.

Fleischner, H. The square of every two-connected graph is Hamiltonian. *Journal of Combinatorial Theory, Series B*, 16:29–34, 1974.

Gross, J.-L. and Yellen, J. *Handbook of graph theory*. CRC press, 2004. ISBN 1-58488-090-2.

Kakade, S., Kalai, A.-T., Kanade, V., and Shamir, O. Efficient learning of generalized linear and single index models with isotonic regression. In *NIPS*24*, pp. 927–935, 2011.

Nock, R. and Nielsen, F. On the efficient minimization of classification-calibrated surrogates. In *NIPS*21*, pp. 1201–1208, 2008.

Nock, R. and Nielsen, F. Bregman divergences and surrogates for learning. *IEEE Trans.PAMI*, 31: 2048–2059, 2009.

Nock, R., Luosto, P., and Kivinen, J. Mixed Bregman clustering with approximation guarantees. In *Proc. of the 19* th *ECML*, pp. 154–169, 2008.

Nock, R., Menon, A.-K., and Ong, C.-S. A scaled Bregman theorem with applications. In *NIPS*29*, pp. 19–27, 2016.

Reid, M.-D. and Williamson, R.-C. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

Shuford, E., Albert, A., and Massengil, H.-E. Admissible probability measurement procedures. *Psychometrika*, pp. 125–145, 1966.