Appendix A: Proof of Lemma 2.2

In this section, we prove Lemma 2.2. We should note that our deviations below also provide insights for the developments of online BayesPA algorithms with the averaging classifiers.

Proof. We prove for the more generalized soft-margin version of BayesPA learning, which can be reformulated using a slack variable ξ:

$$\begin{aligned} q_{t+1}(\boldsymbol{w}) = \quad & \underset{q(\boldsymbol{w})\in\mathcal{P}}{\operatorname{argmin}} \ \mathrm{KL}[q(\boldsymbol{w})||q_t(\boldsymbol{w})] + 2c\xi_t \\ & \text{s.t.} : y_t\mathbb{E}_q[\boldsymbol{w}^\top \boldsymbol{x}_t] \geq \epsilon - \xi_t), \ \ \xi_t \geq 0. \end{aligned} \tag{24}$$

Similar to Corollary 5 in (Zhu et al., 2012), the optimal solution $q^*(\boldsymbol{w})$ of the above problem can be derived from its functional Lagrangian and has the following form:

$$q^*(\boldsymbol{w}) = \frac{1}{\Gamma(\tau_t)} q_t(\boldsymbol{w}) \exp(\tau_t y_t \boldsymbol{w}^\top \boldsymbol{x}_t) \tag{25}$$

where $\Gamma(\tau_t)$ is a normalization term and τ_t is the optimal solution to the dual problem:

$$\begin{aligned} & \max_{\tau_t} \ \tau_t\epsilon - \log\Gamma(\tau_t) \\ & \text{s.t.} \ \ 0 \leq \tau_t \leq 2c \end{aligned} \tag{26}$$

Using this primal-dual interpretation, we first prove that for prior $p_0(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}_0, I)$, $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_t, I)$ for some $\boldsymbol{\mu}_t$ in each round $t = 0, 1, 2, ...$ This can be shown by induction. Assume for round t, the distribution $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_t, I)$. Then for round $t+1$, (25) suggests the distribution

$$q_{t+1}(\boldsymbol{w}) = \frac{\mathcal{C}}{\Gamma(\tau_t)} \exp\left(-\frac{1}{2}||\boldsymbol{w} - (\boldsymbol{\mu}_t + \tau_t y_t \boldsymbol{x}_t)||^2\right) \tag{27}$$

where the constant $\mathcal{C} = \exp(y_t\tau_t\boldsymbol{\mu}_t^\top \boldsymbol{x}_t + \frac{1}{2}\tau_t^2 \boldsymbol{x}_t^\top \boldsymbol{x}_t)$. Therefore, the distribution $q_{t+1}(\boldsymbol{w}) = \mathcal{N}(\mu_t + \tau_t \boldsymbol{x}_t, I)$ and as a by-product, the normalization term $\Gamma(\tau_t) = \sqrt{2\pi}^K \exp(\tau_t y_t \boldsymbol{x}_t^\top \boldsymbol{\mu}_t + \frac{1}{2}\tau_t^2 \boldsymbol{x}_t^\top \boldsymbol{x}_t)$.

Next, we show that $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \tau y_t \boldsymbol{x}_t$ is the optimal solution of the online Passive-Aggressive update rule (Crammer et al., 2006). To see this, we plug the derived $\Gamma(\tau_t)$ into (26), ignore constant terms and obtain

$$\begin{aligned} & \max_{\tau_t} \ \epsilon\tau_t - \tfrac{1}{2}\tau_t^2 \boldsymbol{x}_t^\top \boldsymbol{x}_t - y_t\tau_t \boldsymbol{\mu}_t^\top \boldsymbol{x}_t \\ & \text{s.t.} \ \ 0 \leq \tau_t \leq 2c \end{aligned} \tag{28}$$

which is exactly the dual form of the online Passive-Aggressive update equation:

$$\begin{aligned} \boldsymbol{\mu}_{t+1}^* = \arg\min \ & \tfrac{1}{2}||\boldsymbol{\mu} - \boldsymbol{\mu}_t||^2 + 2c\xi \\ \text{s.t.} \ \ & y_t \boldsymbol{\mu}^\top \boldsymbol{x}_t \geq \epsilon - \xi, \ \ \xi \geq 0, \end{aligned} \tag{29}$$

the optimal solution to which is $\boldsymbol{\mu}_{t+1}^* = \boldsymbol{\mu}_t + \tau_t y_t \boldsymbol{x}_t$. It is then clear that $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_{t+1}^*$. □

Appendix B:

We show the objective in (11) is an upper bound of that in (6), that is,

$$\begin{aligned} & \mathcal{L}(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)) - \mathbb{E}_q[\log(\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w}))] \\ & \geq \mathcal{L}(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)) + 2c \sum_{d\in B_t} \mathbb{E}_q[(\xi_d)_+] \end{aligned} \tag{30}$$

where $\mathcal{L}(q) = \mathrm{KL}[q||q_t(\boldsymbol{w}, \boldsymbol{\Phi})q_0(\boldsymbol{Z}_t)]$.

Proof. We first have

$$\mathcal{L}(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)) = \mathbb{E}_q[\log \frac{q(\boldsymbol{\lambda}_t \mid \boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t) q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)}{q_t(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)}],$$

and

$$\mathcal{L}(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)) = \mathbb{E}_q[\log \frac{q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)}{q_t(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)}]$$

Comparing these two equations and canceling out common factors, we know that in order for (30) to make sense, it suffices to prove

$$\mathbb{H}[q'] - \mathbb{E}_{q'}[\log(\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w})] \geq 2c \sum_{d\in B_t} \mathbb{E}_{q'}[(\xi_d)_+] \tag{31}$$

is uniformly true for any given $(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)$, where $\mathbb{H}(\cdot)$ is the entropy operator and $q' = q(\boldsymbol{\lambda}_t \mid \boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)$. The inequality (31) can be reformulated as

$$\mathbb{E}_{q'}[\log \frac{q'}{\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w})}] \geq 2c \sum_{d\in B_t} \mathbb{E}_{q'}[(\xi_d)_+] \tag{32}$$

Exploiting the convexity of the function $\log(\cdot)$, i.e.

$$-\mathbb{E}_{q'}[\log \frac{\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w})}{q'}] \geq -\log \int_{\boldsymbol{\lambda}_t} \psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w}) \, d\boldsymbol{\lambda}_t,$$

and utilizing the equality (10), we then have (32) and therefore prove (30). □