Supplementary Material
Confidence Weighted Multitask Learning

Peng Yang[1], Peilin Zhao[2], Jiayu Zhou[3], Xin Gao[1]
[1]King Abdullah University of Science and Technology, Saudi Arabia
[2]Tencent AI Lab, China [3]Michigan State University, USA
yangpeng1985521@gmail.com

Proof of Lemma 1

Proof. Since $\mathbf{u} \sim \mathcal{N}(\mathbf{p}, \mathbf{A})$ and $\mathbf{v} \sim \mathcal{N}(\mathbf{q}, \mathbf{B})$ are mutually independent normal random variables, they are jointly normal: we define a random vector $\mathbf{X} = [\mathbf{u}\ \mathbf{v}]^\top$ that has a multivariate normal distribution with mean and covariance matrix:

$$\mathbb{E}[\mathbf{X}] = [\mathbf{p}\ \mathbf{q}]^\top, \quad \text{Var}[\mathbf{X}] = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$$

Given $\mathbf{B} = [\mathbf{1}\ \mathbf{1}]$, we can write

$$\mathbb{E}[\mathbf{w}] = \mathbf{p} + \mathbf{q} = \mathbf{B}\mathbb{E}[\mathbf{X}].$$

According to the proposition on linear transformations, $\mathbf{w}$ has a normal distribution with mean and variance:

$$\mathbb{E}[\mathbf{w}] = \mathbf{B}\mathbb{E}[\mathbf{X}] = \mathbf{p} + \mathbf{q}, \quad \text{Var}[\mathbf{w}] = \mathbf{B}\text{Var}[\mathbf{X}]\mathbf{B}^\top = \mathbf{A} + \mathbf{B}.$$

□

Proof of Lemma 2

Proof. Let $\mathbf{w} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ and $\mathbf{w} = [w_1, \ldots, w_d]$ be d mutually independent normal random variable, having means $\mu = [\mu_1, \ldots, \mu_d]$ and variances $\Sigma = \begin{bmatrix} \sigma_1^2 & \ldots & \sigma_1\sigma_d \\ \ldots & \ldots & \ldots \\ \sigma_d\sigma_1 & \ldots & \sigma_d^2 \end{bmatrix}$
Given an instance-label pair $(\mathbf{x}, y)$ with $\mathbf{x} = [x_1, \ldots, x_d]$ and $y \in \{\pm 1\}$, then the predicted margin is given by

$$M = y(\mathbf{w} \cdot \mathbf{x}) = y \sum_i w_i x_i$$

has a normal distribution with mean and variance:

$$\mathbb{E}[M] = y(\mathbf{x} \cdot \mathbb{E}[\mathbf{w}]) = y(\mu \cdot \mathbf{x}), \quad \text{Var}[M] = (y\mathbf{x})^\top \text{Var}[\mathbf{w}](y\mathbf{x}) = \mathbf{x}^\top \Sigma \mathbf{x}.$$

The constraint in the objective can be reformulated as

$$\Pr_{\mathbf{w}\sim\mathcal{N}(\mu,\mathbf{\Sigma})}[y_t(\mathbf{w}\cdot\mathbf{x}_t)\leq 0]\leq 1-\eta. \tag{1}$$

The prediction on $(\mathbf{x}_t, y_t)$ with $\mathbf{w}\sim\mathcal{N}(\mu,\Sigma)$ follows the Gaussian distribution with mean $\mu_A = y_t(\mu\cdot\mathbf{x}_t)$ and variance $\sigma_A^2 = \mathbf{x}_t^\top\Sigma\mathbf{x}_t$. Thus the probability of a *wrong* classification is

$$\Pr[A\leq 0] = \Pr\left[\frac{A-\mu_A}{\sigma_A}\leq\frac{-\mu_A}{\sigma_A}\right]$$

Since $\frac{A-\mu_A}{\sigma_A}$ is a normally distributed random variable, the probability $\Pr[A\leq 0]$ equals $\Phi(-\frac{\mu_A}{\sigma_A})\leq 1-\eta$, where Φ is the cumulative function of the normal distribution. Thus we can rewrite (1) as

$$-\frac{\mu_A}{\sigma_A}\leq\Phi^{-1}(1-\eta) = -\Phi^{-1}(\eta).$$

Substituting μ_A and σ_A by their definitions and rearranging terms we obtain:

$$y_t(\mu\cdot\mathbf{x}_t) - \phi\sqrt{\mathbf{x}_t^\top\Sigma\mathbf{x}_t}\geq 0,$$

where $\phi = \Phi^{-1}(\eta)$. □

Proof of Lemma 3

Proof. 1). Firstly, we solve the problem with squared hinge loss. Similar to [1], we take the derivative of the above problem and set it to zero, we have

$$\begin{aligned}\mathbf{q}_t &= \mathbf{q}_{t-1} - \frac{1}{2\epsilon}\left[\frac{d}{d\mathbf{q}}\ell_{h^2}(\mathbf{p}_{t-1}+\mathbf{q};\mathbf{x}_t,y_t)\right]_{\mathbf{q}=\mathbf{q}_t}\mathbf{B}_t\\ &= \mathbf{q}_{t-1} + \frac{y_t}{\epsilon}\left[1-y_t(\mathbf{p}_{t-1}+\mathbf{q}_t)\cdot\mathbf{x}_t\right]\mathbf{B}_t\mathbf{x}_t.\end{aligned} \tag{2}$$

By taking the dot product of each side of the equation with $y_t\mathbf{x}_t$, Eq. (2) yields

$$y_t(\mathbf{q}_t\cdot\mathbf{x}_t) = y_t(\mathbf{q}_{t-1}\cdot\mathbf{x}_t) + \frac{(y_t)^2}{\epsilon}\left(1-y_t(\mathbf{p}_{t-1}+\mathbf{q}_t)\cdot\mathbf{x}_t\right)\mathbf{x}_t^\top\mathbf{B}_t\mathbf{x}_t.$$

Solving for $y_t(\mathbf{q}_t\cdot\mathbf{x}_t)$, we get

$$y_t(\mathbf{q}_t\cdot\mathbf{x}_t) = \frac{\epsilon y_t(\mathbf{q}_{t-1}\cdot\mathbf{x}_t) + (1-y_t\mathbf{p}_{t-1}\cdot\mathbf{x}_t)\mathbf{x}_t^\top\mathbf{B}_t\mathbf{x}_t}{\epsilon+\mathbf{x}_t^\top\mathbf{B}_t\mathbf{x}_t}.$$

Thus,

$$\frac{1-y_t(\mathbf{p}_{t-1}+\mathbf{q}_t)\cdot\mathbf{x}_t}{\epsilon} = \frac{1-y_t(\mathbf{p}_{t-1}+\mathbf{q}_{t-1})\cdot\mathbf{x}_t}{\epsilon+\mathbf{x}_t^\top\mathbf{B}_t\mathbf{x}_t}. \tag{3}$$

Substituting Eq. (3) in Eq. (2), we obtain our solution.

2). We next solve $\mathbf{q}$ with the hinge loss:

$$\min_{\mathbf{q}}(\mathbf{q}-\mathbf{q}_{t-1})\mathbf{B}_t^{-1}(\mathbf{q}-\mathbf{q}_{t-1}) + \frac{1}{\epsilon}\max\{0, 1-y_t(\mathbf{q}+\mathbf{p}_{t-1})\cdot\mathbf{x}_t\}.$$

Next, we change variables:

$$\mathbf{q} = \mathbf{B}_t^{1/2}\mathbf{v}, \;\; \mathbf{q}_{t-1} = \mathbf{B}_t^{1/2}\mathbf{v}_{t-1}, \;\; \mathbf{p}_{t-1} = \mathbf{B}_t^{1/2}\mathbf{u}_{t-1}, \;\; \mathbf{x}_t = \mathbf{B}_t^{-1/2}\mathbf{z}_t.$$

Substituting in the minimization problem, we get

$$\min_{\mathbf{v}} (\mathbf{v} - \mathbf{v}_{t-1})^\top (\mathbf{v} - \mathbf{v}_{t-1}) + \frac{1}{\epsilon} \max\{0, 1 - y_t(\mathbf{v} + \mathbf{u}_{t-1}) \cdot \mathbf{z}_t\}. \tag{4}$$

When $1 - y_t(\mathbf{v} + \mathbf{u}_{t-1}) \cdot \mathbf{z}_t > 0$, the solution is given by

$$\mathbf{v} = \mathbf{v}_{t-1} + \frac{1}{2\epsilon} y_t \mathbf{z}_t. \tag{5}$$

Substituting the solution back to Eq. (4), we have that

$$-\frac{1}{4\epsilon^2}\mathbf{z}_t^\top \mathbf{z}_t + \frac{1}{\epsilon}(1 - y_t(\mathbf{v}_{t-1} + \mathbf{u}_{t-1}) \cdot \mathbf{z}_t).$$

Solving the above objective with respect to $\frac{1}{\epsilon}$, we obtain that

$$\frac{1}{2\epsilon} = \frac{1 - y_t(\mathbf{v}_{t-1} + \mathbf{u}_{t-1}) \cdot \mathbf{z}_t}{\mathbf{z}_t^\top \mathbf{z}_t}.$$

Motivated by [2], setting $\frac{1}{2\epsilon}$ as the upper bound of learning rate, yielding

$$a_t = \min\left\{\frac{1}{2\epsilon}, \frac{1 - y_t(\mathbf{u}_{t-1} + \mathbf{v}_{t-1}) \cdot \mathbf{z}_t}{\mathbf{z}_t^\top \mathbf{z}_t}\right\}.$$

Substituting back the original variables into Eq. (5), we obtain

$$\mathbf{B}_t^{-1/2}\mathbf{q} = \mathbf{B}_t^{-1/2}\mathbf{q}_{t-1} + a_t y_t \mathbf{B}_t^{1/2}\mathbf{x}_t \quad \Rightarrow \quad \mathbf{q} = \mathbf{q}_{t-1} + a_t y_t \mathbf{B}_t \mathbf{x}_t,$$

where

$$a_t = \min\left\{\frac{1}{2\epsilon}, \max\left\{0, \frac{1 - y_t \mathbf{x}_t(\mathbf{p}_{t-1} + \mathbf{q}_{t-1})}{\mathbf{x}_t^\top \mathbf{B}_t \mathbf{x}_t}\right\}\right\}.$$

□

Proof of Theorem 1

Proof. Assume a task model $\mathbf{w} \sim (\mu, \Sigma)$, where $\mu = \mathbf{p} + \mathbf{q}$ and $\Sigma = \mathbf{A} + \mathbf{B}$. We can verify that $\mu_{t+1} = \arg\min_{\mu} h_t(\mu)$, where

$$h_t(\mu) = \frac{1}{2}\|\mu_t - \mu\|^2_{\Sigma^{-1}_{t+1}} + \frac{1}{2\epsilon}\mathbf{g}_t^\top \mu,$$

where $\mathbf{g}_t = \nabla_{\mu_t} \ell_h(\cdot)$ is the gradient descent of the hinge loss function. Because h_t is convex, we have

$$\partial h_t(\mu_{t+1})^\top (\mu - \mu_{t+1}) = \left[(\mu_{t+1} - \mu_t)^\top \Sigma^{-1}_{t+1} + \frac{1}{2\epsilon}\mathbf{g}_t^\top\right](\mu - \mu_{t+1}) \geq 0, \ \forall \mu.$$

Re-arranging the above inequality will result in

$$\begin{aligned}\frac{1}{2\epsilon}\mathbf{g}_t^\top(\mu_{t+1} - \mu) \leq& (\mu_{t+1} - \mu_t)^\top \Sigma^{-1}_{t+1}(\mu - \mu_{t+1}) \\ =& \frac{1}{2}\left[\|\mu - \mu_t\|^2_{\Sigma^{-1}_{t+1}} - \|\mu_{t+1} - \mu_t\|^2_{\Sigma^{-1}_{t+1}} - \|\mu - \mu_{t+1}\|^2_{\Sigma^{-1}_{t+1}}\right],\end{aligned}$$

where the last equality is motivated by $ab = \frac{1}{2}[(a+b)^2 - a^2 - b^2]$. For the left side of the inequality above,

$$\begin{aligned}\mathbf{g}_t^\top(\mu_{t+1} - \mu) &= \mathbf{g}_t^\top(\mu_t - \mu + \mu_{t+1} - \mu_t) \\ &= \mathbf{g}_t^\top(\mu_t - \mu) + \mathbf{g}_t^\top(\mu_{t+1} - \mu_t).\end{aligned}$$

Combining the above two formulas will give the following important inequality

$$\mathbf{g}_t^\top(\mu_t - \mu) \leq \epsilon\left(\|\mu - \mu_t\|^2_{\Sigma^{-1}_{t+1}} - \|\mu_{t+1} - \mu_t\|^2_{\Sigma^{-1}_{t+1}} - \|\mu - \mu_{t+1}\|^2_{\Sigma^{-1}_{t+1}}\right) - \mathbf{g}_t^\top(\mu_{t+1} - \mu_t).$$

Summing the above inequality over $t = 1, 2, \ldots, T$, gives

$$\begin{aligned}\sum_{t \in U_T}^{T}(\mathbf{g}_t^\top \mu_t - \mathbf{g}_t^\top \mu) \leq& \epsilon \sum_{t=1}^{T}\left[\|\mu - \mu_t\|^2_{\Sigma^{-1}_{t+1}} - \|\mu - \mu_{t+1}\|^2_{\Sigma^{-1}_{t+1}}\right] \\ &-\epsilon\sum_{t=1}^{T}\|\mu_{t+1} - \mu_t\|^2_{\Sigma^{-1}_{t+1}} - \sum_{t=1}^{T}\mathbf{g}_t^\top(\mu_{t+1} - \mu_t).\end{aligned} \tag{6}$$

Since the $\ell_t(\mu)$ is convex, $\mathbf{g}_t^\top(\mu_t - \mu) \geq \ell_t(\mu_t) - \ell_t(\mu)$. According to the regret definition, the left side $\sum_t(\ell_t(\mu_t) - \ell_t(\mu))$ is the regret.

Next we bound the right hand side of the first term. According to the proof of Theorem 1 in [3], for all $t \in U_T$,

$$\sum_{t=1}^{T}[\|\mu - \mu_t\|^2_{\Sigma^{-1}_{t+1}} - \|\mu - \mu_{t+1}\|^2_{\Sigma^{-1}_{t+1}}] \leq \max_{t \in U_T}\|\mu_t - \mu\|^2 \mathrm{Tr}(\Sigma^{-1}_{U_T}), \tag{7}$$

For the second term, we notice that the following inequality holds according to the update rule of μ,

$$(\mu_{t+1} - \mu_t)^\top \Sigma^{-1}_{t+1} + \frac{1}{2\epsilon}\mathbf{g}_t^\top = 0,$$

so that

$$\|\mu_{t+1}-\mu_t\|^2_{\Sigma^{-1}_{t+1}} = (\mu_{t+1}-\mu_t)^\top \Sigma^{-1}_{t+1}\Sigma_{t+1}\Sigma^{-1}_{t+1}(\mu_{t+1}-\mu_t) = \frac{1}{4\epsilon^2}\mathbf{g}_t^\top \Sigma_{t+1}\mathbf{g}_t.$$

For the third term,

$$\mathbf{g}_t^\top(\mu_{t+1}-\mu_t) = -\frac{1}{2\epsilon}\mathbf{g}_t^\top\Sigma_{t+1}\mathbf{g}_t.$$

Combining the above two inequalities will result in

$$\begin{aligned}
&-\epsilon\sum_{t=1}^{T}\|\mu_{t+1}-\mu_t\|^2_{\mathbf{\Sigma}^{-1}_{t+1}} - \sum_{t=1}^{T}\mathbf{g}_t^\top(\mu_{t+1}-\mu_t)\\
&= -\frac{1}{4\epsilon}\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{\Sigma}_{t+1}\mathbf{g}_t + \frac{1}{2\epsilon}\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{\Sigma}_{t+1}\mathbf{g}_t = \frac{1}{4\epsilon}\sum_{t=1}^{T}\mathbf{x}_{t+1}^\top\mathbf{\Sigma}_{t+1}\mathbf{x}_{t+1}.
\end{aligned}$$

where the last equality is hold when $\mathbf{g}_t = -y_{t+1}\mathbf{x}_{t+1}$. According to the definition of $\mathbf{\Sigma}^k$ ($k \in [K]$) in multitask setting,

$$\sum_t \mathbf{x}_t^{k\top}\mathbf{\Sigma}_t^k\mathbf{x}_t^k = \sum_t\left(\mathbf{x}_t^{k\top}\mathbf{A}_t\mathbf{x}_t^k + \mathbf{x}_t^{k\top}\mathbf{B}_t^k\mathbf{x}_t^k\right)$$

Assume that $\|\mathbf{x}_t\| \le 1$ and $0 \le \frac{1}{\lambda} \le 1$, we obtain

$$\begin{aligned}
&\sum_t \mathbf{x}_t^\top\mathbf{B}_t\mathbf{x}_t = \lambda\sum_t\left(1-\frac{|\mathbf{B}_{t-1}^{-1}|}{|\mathbf{B}_t^{-1}|}\right)\\
\le &-\lambda\sum_t\log\left(\frac{|\mathbf{B}_{t-1}^{-1}|}{|\mathbf{B}_t^{-1}|}\right) = \lambda\log(|\mathbf{B}_T^{-1}|) \le \lambda\log(1+T),
\end{aligned}$$

where the first equality is inferred from

$$\mathbf{B}_t^{-1} = \mathbf{B}_{t-1}^{-1} + \frac{1}{\lambda}\mathbf{x}_t\mathbf{x}_t^\top \quad\Rightarrow\quad \frac{1}{\lambda}\mathbf{x}_t^\top\mathbf{B}_t\mathbf{x}_t = 1-\frac{|\mathbf{B}_{t-1}^{-1}|}{|\mathbf{B}_t^{-1}|},$$

while the second inequality is due to

$$1-1/x \le \log(x), \quad \text{for all} \quad x \ge 1,$$

and $\mathbf{B}_{t-1}^{-1} \preceq \mathbf{B}_t^{-1}$ for $t \ge 1$. Finally, the last inequality is inferred from $\mathbf{B}_T^{-1} = \mathbf{I} + \frac{1}{\lambda}\sum_t^T \mathbf{x}_t\mathbf{x}_t^\top$ with $\|\mathbf{x}_t\| \le 1$. Similarly, we have a bound for

$$\sum_t \mathbf{x}_t^\top\mathbf{A}_t^k\mathbf{x}_t \le \lambda\log(1+KT),$$

given $\mathbf{A}_T^{-1} = \mathbf{I} + \frac{1}{\lambda}\sum_{k=1}^K\sum_t^T \mathbf{x}_t^k\mathbf{x}_t^{k\top}$. To summarize, the third term in Eq.(6) is bounded by

$$\frac{1}{4\epsilon}\sum_t \mathbf{x}_t^{k\top}(\mathbf{A}_t+\mathbf{B}_t^k)\mathbf{x}_t^k \le \frac{\lambda}{4\epsilon}\log(1+KT). \tag{8}$$

Plugging Eq. (7), (8) into the Eq. (6) concludes the proof.

□

Proof of Theorem 2

Proof. According to Eq. (6), (7), (8) in the proof of Theorem 1,

$$\sum_{t\in U_T} \mathbf{g}_t^\top \mu_t - \mathbf{g}_t^\top \mu \leq \frac{1}{4\epsilon}\sum_t \mathbf{g}_t^\top \mathbf{\Sigma}_t \mathbf{g}_t + \max_{t\in U_T} \|\mu_t - \mu\|^2 \mathrm{Tr}(\Sigma_{U_T}^{-1}). \tag{9}$$

In active learning setting with query/update decision Q_t/Z_t, $\mathbf{g}_t = \nabla_{\mu_t}\ell(\cdot) = -Q_t Z_t y_t \mathbf{x}_t$, where $Q_t Z_t = 1$ if $\ell(\mu_t) > 0$, and $Q_t Z_t = 0$, otherwise. Thus, we rearrange Eq. (9) with some manipulations,

$$\sum_{t=1}^T Q_t Z_t \left(-y_t \mathbf{x}_t^\top \mu_t - \frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t\right) \leq \sum_{t=1}^T -Q_t Z_t y_t \mathbf{x}_t^\top \mu + \max_{t\in U_T} \|\mu_t - \mu\|^2 \mathrm{Tr}(\Sigma_T^{-1}).$$

When an error occurs, i.e., $y_t \mathbf{x}_t^\top \mu_t \leq 0$, we have $-y_t \mathbf{x}_t^\top \mu_t = |\Delta_t|$. Since μ is a random variable, we use $h\mu$ to replace μ. We add a positive scalar $Q_t Z_t h > 0$ on both sides of the above inequality, which introduce a upper bound for $\Theta_t + h$:

$$\sum_{t=1}^T Q_t Z_t \left(\Theta_t + h\right) \leq h\sum_{t=1}^T Q_t Z_t \ell(\mu; \mathbf{x}_t, y_t) + \lambda \max_{t\in U_T} \|\mu_t - h\mu\|^2 \mathrm{Tr}(\Sigma_T^{-1}), \tag{10}$$

where $\Theta_t = |\Delta_t| - \frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t$, and

$$Q_t Z_t (h - h y_t \mathbf{x}_t^\top \mu) \leq Q_t Z_t \max(0, h - h y_t \mathbf{x}_t^\top \mu) = h Q_t Z_t \ell_h(\mu; \mathbf{x}_t, y_t).$$

When an error occurs at trial $t \in \mathcal{M}$, the function Θ_t can be positive in randomized query set ($t \in \mathcal{M} \cap \mathcal{S}$) or negative in deterministic query set ($t \in \mathcal{M} \cap \mathcal{D}$). In the former case, Q_t is a random variable with $\mathbb{E}[Q_t] = \frac{h}{h+\Theta_t}$, we have

$$\mathbb{E}[Q_t Z_t(\Theta_t + h)] = \mathbb{E}[Z_t]\mathbb{E}[Q_t(\Theta_t + h)] = h\mathbb{E}[Z_t].$$

In the later case, $\mathbb{E}[Q_t] = 1$, yielding

$$\mathbb{E}[Q_t Z_t(|\Delta_t| - \frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t + h)] \geq \mathbb{E}[Z_t(h - \frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t)] \geq h\mathbb{E}[Z_t] - \mathbb{E}[\frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t],$$

where the first inequality is due to $|\Delta_t| \geq 0$. To summarize,

$$\begin{aligned}\sum_{t=1}^T Q_t Z_t \left(\Theta_t + h\right) &\geq \sum_{t\in\mathcal{M}\cap\mathcal{S}} h\mathbb{E}[Z_t] + \sum_{t\in\mathcal{M}\cap\mathcal{D}} \left(h\mathbb{E}[Z_t] - \mathbb{E}[\frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t]\right) \\ &= h\mathbb{E}[M] - \sum_{t\in\mathcal{M}\cap\mathcal{D}} \mathbb{E}[\frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t].\end{aligned} \tag{11}$$

Plugging Eq. (11) into Eq. (10), give

$$\begin{aligned}\mathbb{E}[M] \leq \sum_{t=1}^T \mathbb{E}[\ell(\mu; \mathbf{x}_t, y_t)] + \frac{1}{4h\epsilon}\sum_{t\in\mathcal{M}\cap\mathcal{D}} \mathbb{E}[\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t] \\ + \frac{\epsilon}{h}\mathbb{E}\left[\max_{t\leq T} \|\mu_t - h\mu\|^2 \mathrm{Tr}(\Sigma_T^{-1})\right].\end{aligned} \tag{12}$$

Plugging Eq. (8) into Eq. (12) can conclude the proof. ∎

Proof of Theorem 3

Proof. The update trials in Algorithm 2 can be divided into three groups,

$$\sum_t Z_t = |\mathcal{M} \cap \mathcal{S}| + |\mathcal{M} \cap \mathcal{D}| + |\mathcal{V} \cap \mathcal{D}|.$$

In the first case of the randomized query $t \in \mathcal{M} \cap \mathcal{S}$, Q_t is a random variable with $E[Q_t] = \frac{h}{h+\Theta_t}$, we bound,

$$\sum_{t \in \mathcal{M} \cap \mathcal{S}} \mathbb{E}[Q_t Z_t(|\Delta_t| - \frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t + h)] = h \sum_{t \in \mathcal{M} \cap \mathcal{S}} \mathbb{E}[Z_t].$$

In the second case of deterministic query $t \in \mathcal{M} \cap \mathcal{D}$, $\mathbb{E}[Q_t] = 1$,

$$\sum_{t \in \mathcal{M} \cap \mathcal{D}} \mathbb{E}[Q_t Z_t(|\Delta_t| - \frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t + h)] \geq h \sum_{t \in \mathcal{M} \cap \mathcal{D}} \mathbb{E}[Z_t] - \frac{1}{4\epsilon} \sum_{t \in \mathcal{M} \cap \mathcal{D}} \mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t.$$

Finally, we consider the third case where the update is performed with no mistake ($t \in \mathcal{V} \cap \mathcal{D}$), i.e., $y_t \Delta_t > 0$ and $\mathbb{E}[Q_t] = 1$, we have that

$$\sum_{t \in \mathcal{V} \cap \mathcal{D}} \mathbb{E}[Q_t Z_t(|\Delta_t| - \frac{1}{4\epsilon}\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t + h)] \geq h \sum_{t \in \mathcal{V} \cap \mathcal{D}} \mathbb{E}[Z_t] - \frac{1}{4\epsilon} \sum_{t \in \mathcal{V} \cap \mathcal{D}} \mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t.$$

To summarize,

$$\mathbb{E}[\sum_t Q_t Z_t(\Theta_t + h)] \geq h \sum_{t \in \mathcal{M}} \mathbb{E}[Z_t] + h \sum_{t \in \mathcal{V} \cap \mathcal{D}} \mathbb{E}[Z_t] - \frac{1}{4\epsilon} \sum_{t \in \mathcal{D}} \mathbf{x}_t^\top \mathbf{\Sigma}_{t+1} \mathbf{x}_t. \quad (13)$$

Since $\mathcal{V} \subset \mathcal{D}$, we have $\sum_{t \in \mathcal{V} \cap \mathcal{D}} \mathbb{E}[Z_t] = \mathbb{E}[V]$. Plugging the upper bound (8), (10) into (13) can conlude the proof. □

References

[1] K. Crammer, A. Kulesza, and M. Dredze, "Adaptive regularization of weight vectors," in *NIPS*, 2009, pp. 414–422.

[2] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *JMLR*, vol. 7, pp. 551–585, 2006.

[3] P. Zhao, F. Zhuang, M. Wu, X.-L. Li, and S. C. Hoi, "Cost-sensitive online classification with adaptive regularization and its applications," in *ICDM*. IEEE, 2015, pp. 649–658.