Supplementary Material
Efficient Online Multi-Task Learning via Adaptive Kernel Selection

Peng Yang and Ping Li
Cognitive Computing Lab, Baidu Research
10900 NE 8th ST, Bellevue WA 98004

Proof of Lemma 1

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 $\mu \in \mathbb{R}^d$ and variances $\Sigma \in \mathbb{R}^{d \times d}$. 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

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

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

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

Since $\frac{A - \mu_A}{\sigma_A}$ is a normally distributed random variable, the probability $\text{Pr}[A \leq 0]$ 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 $\phi = \Phi^{-1}(\eta)$. $\square$

Proof of Lemma 3

Proof. The parameters $\mathbf{\Sigma}^k$ can be solved as below,

$$f(\mathbf{\Sigma}^k) = \log\left(\frac{|\mathbf{\Sigma}_{t-1}^k|}{|\mathbf{\Sigma}^k|}\right) + \mathrm{Tr}\left(\frac{\mathbf{\Sigma}^k}{\mathbf{\Sigma}_{t-1}^k}\right) + \frac{1}{\lambda}\phi_t^{k\top}\mathbf{\Sigma}^k\phi_t^k.$$

By applying the KKT condition on $\mathbf{\Sigma}$, we have that $(\mathbf{\Sigma}_t^k)^{-1} = (\mathbf{\Sigma}_{t-1}^k)^{-1} + \frac{1}{\lambda}\phi_t^k\phi_t^{k\top}$. By using the Sherman$-$Morrison formula [**?**], $\mathbf{\Sigma}_t^k$ can be updated efficiently with time complexity $O(D^2)$,

$$\mathbf{\Sigma}_t^k = \mathbf{\Sigma}_{t-1}^k - \frac{\mathbf{\Sigma}_{t-1}^k\phi_t^k\phi_t^{k\top}\mathbf{\Sigma}_{t-1}^k}{\lambda + \phi_t^{k\top}\mathbf{\Sigma}_{t-1}^k\phi_t^k}. \tag{2}$$

Let $\mu^0 = \mu_{t-1}^0$, the μ^k is solved under the hinge loss and squared hinge loss, respectively.Let $\hat{e}_t^k = \langle \mu_{t-1}^0 + \mu_{t-1}^k, \phi_t^k\rangle$. Whenever $y_t^k \neq \mathrm{sgn}(\hat{e}_t^k)$, we solve the problem

$$f(\mu^k) = \|\mu^k - \mu_{t-1}^k\|^2_{(\mathbf{\Sigma}_t^k)^{-1}} + \frac{1}{\epsilon}\ell_t(\mu_{t-1}^0 + \mu^k),$$

where the optimal solution of μ^k is given by,

$$\mu_t^k = \mu_{t-1}^k + g_t^k y_t^k \mathbf{\Sigma}_t^k \phi_t^k, \tag{3}$$

where

$$g_t^k = \frac{\max\{0, 1 - y_t^k\hat{e}_t^k\}}{\epsilon + \phi_t^{k\top}\mathbf{\Sigma}_t^k\phi_t^k} \quad \text{(squared hinge)}$$

$$g_t^k = \min\left\{\frac{1}{2\epsilon}, \max\left\{0, \frac{1 - y_t^k\hat{e}_t^k}{\phi_t^{k\top}\mathbf{\Sigma}_t^k\phi_t^k}\right\}\right\} \quad \text{(hinge)}$$

The global parameter $\mathbf{\Sigma}^0$ can be optimized:

$$f(\mathbf{\Sigma}^0) = \log\left(\frac{|\mathbf{\Sigma}_{t-1}^0|}{|\mathbf{\Sigma}^0|}\right) + \mathrm{Tr}\left(\frac{\mathbf{\Sigma}^0}{\mathbf{\Sigma}_{t-1}^0}\right) + \frac{1}{\lambda}\mathrm{Tr}\left(\Phi_t^\top\mathbf{\Sigma}^0\Phi_t\right),$$

where $\Phi_t = [\phi_t^1, \phi_t^2, \ldots, \phi_t^K] \in \mathbb{R}^{D\times K}$. By using Woodbury matrix identity, $\mathbf{A}$ can be updated by

$$\mathbf{\Sigma}_t^0 = \mathbf{\Sigma}_{t-1}^0 - \mathbf{\Sigma}_{t-1}^0\Phi_t\mathbf{C}_{t-1}^{-1}\Phi_t^\top\mathbf{\Sigma}_{t-1}^0, \tag{4}$$

where $\mathbf{C}_{t-1} = \lambda\mathbf{I}_K + \Phi_t^\top\mathbf{\Sigma}_{t-1}^0\Phi_t$ is positive-definite and $\mathbf{I}_K \in \mathbb{R}^{K\times K}$ is an identity matrix. The matrix inverse in Eq. (4) takes $O(K^3 + d^2K)$ complexity, which is acceptable when the task number K is small.

Let $z_t^k = \mathcal{I}(y_t^k \neq \hat{y}_t^k)$ where $\mathcal{I}(\cdot)$ is an indicator function, μ^0 is solved by

$$f(\mu^0) = \|\mu^0 - \mu_{t-1}^0\|^2_{(\mathbf{\Sigma}_t^0)^{-1}} + \frac{1}{\epsilon}\sum_{k=1}^K z_t^k\ell_t(\mu^0 + \mu_{t-1}^k).$$

Taking the derivative of the above problem, i.e. $\nabla_{\mu_{t-1}^0} f(\mu^0)$, μ^0 is solved by

$$\mu_t^0 = \mu_{t-1}^0 + \frac{1}{2\epsilon}\mathbf{\Sigma}_t^0\sum_{k=1}^K z_t^k y_t^k\phi_t^k. \tag{5}$$

□

Proof of Theorem 1

Proof. Assume a task model $\mathbf{w}^k \sim (\widehat{\mu}^k, \widehat{\Sigma}^k)$, where $\widehat{\mu}^k = \mu^0 + \mu^k$ and $\widehat{\Sigma}^k = \Sigma^0 + \Sigma^k$. 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 [**?**], 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 $\mathbf{g}_t = -y_{t+1}\phi_{t+1}$. According to the definition of $\widehat{\mathbf{\Sigma}}^k$ ($k \in [K]$) in multitask setting,

$$\sum_t \phi_t^{k\top}\widehat{\mathbf{\Sigma}}_t^k\phi_t^k = \sum_t \left(\phi_t^{k\top}\mathbf{\Sigma}_t^0\phi_t^k + \phi_t^{k\top}\mathbf{\Sigma}_t^k\phi_t^k\right)$$

Assume that $\mathcal{K}_{t,t} = \langle\phi_t,\phi_t\rangle \le 1$ and $0 \le \frac{1}{\lambda} \le 1$, we obtain

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

where the first equality is inferred from

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

while the second inequality is due to

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

and $\mathbf{\Sigma}_{t-1}^{-1} \preceq \mathbf{\Sigma}_t^{-1}$ for $t \ge 1$. Finally, the last inequality is inferred from $(\mathbf{\Sigma}_T^k)^{-1} = \mathbf{I} + \frac{1}{\lambda}\sum_t^T \phi_t^k\phi_t^{k\top}$ with $\|\phi_t^k\| \le 1$.

Similarly, we have a bound for $\sum_t \phi_t^\top\mathbf{\Sigma}_t^0\phi_t$:

$$\sum_t \phi_t^\top\mathbf{\Sigma}_t^0\phi_t \le \lambda\log(1+KT),$$

given $(\mathbf{\Sigma}_T^0)^{-1} = \mathbf{I} + \frac{1}{\lambda}\sum_{k=1}^K\sum_{t=1}^T \phi_t^k\phi_t^{k\top}$. To summarize, the third term in Eq.(6) is bounded by

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

Plugging Eq. (7), (8) into the Eq. (6), we have

$$Regret \leq \frac{\lambda \log(1+KT)}{4\epsilon} + \epsilon(\max_{t\in U_T} \|\mu_t - \mu\|^2 \mathrm{Tr}\left((\mathbf{\Sigma}_T^0 + \mathbf{\Sigma}_T^k)^{-1}\right).$$

Let $\mathcal{D}(\mu) = \max_{t\in U_T} \|\mu_t - \mu\|^2$ and set $\epsilon = \frac{1}{2}\sqrt{\frac{\lambda \log(1+KT)}{(\mathcal{D}(\mu))^2 \mathrm{Tr}\left((\mathbf{\Sigma}_T^0+\mathbf{\Sigma}_T^k)^{-1}\right)}}$, the algorithm satisfies:

$$Regret \leq \frac{1}{2}\sqrt{\lambda}\mathcal{D}(\mu)\sqrt{\mathrm{Tr}((\mathbf{\Sigma}_T^0 + \mathbf{\Sigma}_T^k)^{-1})\log(1+KT)}.$$

□

Proof of Theorem 2

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

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

In active learning setting with query/update decision Q_t/Z_t, $\mathbf{g}_t = \nabla_{\mu_t}\ell(\cdot) = -Q_t Z_t y_t \phi_t$, 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,

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

When an error occurs, i.e., $y_t\phi_t^\top \mu_t \leq 0$, we have $-y_t\phi_t^\top \mu_t = |\widehat{f}_t|$. 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 $\Theta_t + h$:

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

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

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

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

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

where the first inequality is due to $|\widehat{f}_t| \geq 0$. To summarize,

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

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

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

Plugging Eq. (8) into Eq. (12) can conclude the proof. $\square$