Fast Generalized Distillation for Semi-supervised Domain Adaptation

Theorem 1 *Let $L_c(\lambda)$ be a strongly convex function and λ^* be its optimal solution.Let $\lambda_1, ..., \lambda_{T+1}$ be a sequence such that $\lambda_t \in B$ where B is a closed convex set. For $t > 1$, we have $\lambda_{t+1} = \lambda_t - \eta_t \Delta_t$, where Δ_t is the sub-gradient of $L(\lambda_t)$ and $\eta_t = 1/t$. Assume we have $||\Delta_t|| \leq G$ for all t. Then we have:*

$$L_c(\lambda_{T+1}) \leq L_c(\lambda^*) + \frac{G^2(1+\ln(T))}{2T} \tag{1}$$

As $L_c(\lambda)$ is strongly convex and Δ_t is in its sub-gradient set at λ_t, according to the definition of strong convexity, the following inequality holds:

$$\langle \lambda_t - \lambda^*, \Delta_t \rangle \geq L(\lambda_t) - L(\lambda^*) + \frac{1}{2}||\lambda_t - \lambda^*||^2 \tag{2}$$

For the term $\langle \lambda_t - \lambda^*, \Delta_y \rangle$, it can be written as:

$$\begin{aligned}\langle \lambda_t - \lambda^*, \Delta_t \rangle &= \left\langle \lambda_t - \frac{1}{2}\eta_t \Delta_t + \frac{1}{2}\eta_t\Delta_t - \lambda^*, \Delta_t \right\rangle \\ &= \frac{1}{2}\left\langle [(\lambda_t - \eta_t\Delta_t) - \lambda^*] + (\lambda_t - \lambda^*) + \eta_t\Delta_t, \Delta_t \right\rangle \\ &= \frac{1}{2}\left\langle (\lambda_{t+1} - \lambda^*) + (\lambda_t - \lambda^*), \Delta_t \right\rangle + \frac{1}{2}\eta_t\Delta_t^2 \\ &= \frac{1}{2}\left\langle \lambda_{t+1} + \lambda_t - 2\lambda^*, \Delta_t \right\rangle + \frac{1}{2}\eta_t\Delta_t^2\end{aligned} \tag{3}$$

Then we have:

$$\begin{aligned}||\lambda_t - \lambda^*||^2 - ||\lambda_{t+1} - \lambda^*||^2 &= (\lambda_t - \lambda_{t+1})(\lambda_t + \lambda_{t+1} - 2\lambda^*) \\ &= \langle \lambda_{t+1} + \lambda_t - 2\lambda^*, \eta_t\Delta_t \rangle\end{aligned} \tag{4}$$

Using the assumption $||\Delta_t|| \leq G$, we can rearrange (2) and plug (3) and (4) into it, we have:

$$\begin{aligned}&Diff_t = L_c(\lambda_t) - L_c(\lambda^*) \\ &\leq \frac{||\lambda_t - \lambda^*||^2 - ||\lambda_{t+1} - \lambda^*||^2}{2\eta_t} - \frac{1}{2}||\lambda_t - \lambda^*||^2 + \frac{1}{2}\eta_t\Delta_t^2 \\ &\leq \frac{||\lambda_t - \lambda^*||^2 - ||\lambda_{t+1} - \lambda^*||^2}{2\eta_t} - \frac{\lambda}{2}||\lambda_t - \lambda^*||^2 + \frac{1}{2}\eta_t G^2 \\ &= \frac{(t-1)}{2}||\lambda_t - \lambda^*||^2 - \frac{t}{2}||\lambda_{t+1} - \lambda^*||^2 + \frac{1}{2}\eta_t G^2\end{aligned} \tag{5}$$

Due to the strong convexity, for each pair of $L_c(\lambda_t)$ and $L_c(\lambda_{t+1})$ and $t = 1, ..., T$, according to (2), we have:

$$\begin{aligned}L_c(\lambda_{t+1}) - L_c(\lambda_t) &\leq \langle \lambda_{t+1} - \lambda_t, \Delta_t \rangle - \frac{1}{2}||\lambda_{t+1} - \lambda_t||^2 \\ &= -\eta_t\Delta_t^2(1 - \frac{1}{2t}) \leq 0\end{aligned} \tag{6}$$

Therefore, we have the following sequence $L_c(\lambda^*) \leq L_c(\lambda_T) \leq L_c(\lambda_{T-1}) \leq ... \leq L_c(\lambda_1)$. For the sequence $Diff_t$ for $t = 1, ..., T$, we have:

$$\sum_{t=1}^{T} Diff_t = \sum_{t=1}^{T} L_c(\lambda_t) - TL_c(\lambda^*) \geq T\left[L_c(\lambda_T) - L_c(\lambda^*)\right] \tag{7}$$

Next, we show that

$$\begin{aligned}&\sum_{t=1}^{T} Diff_t = \\ &\sum_{t=1}^{T}\left\{\frac{(t-1)}{2}||\lambda_t - \lambda^*||^2 - \frac{t}{2}||\lambda_{t+1} - \lambda^*||^2 + \frac{1}{2}\eta_t G^2\right\} \\ &= -\frac{T}{2}||\lambda_{T+1} - \lambda^*||^2 + \frac{G^2}{2}\sum_{t=1}^{T}\frac{1}{t} \\ &\leq \frac{G^2}{2}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{G^2}{2}(1 + \ln(T))\end{aligned} \tag{8}$$

Combining (7) and rearranging the result, we have:

$$L_c(\lambda_{T+1}) \leq L_c(\lambda^*) + \frac{G^2(1+\ln(T))}{2T}$$