Algorithm 1 Meta-algorithm

Require: A step size α, and a set $\mathcal{H}$ containing step sizes for experts and $\mathbf{w}_0$

Activate a set of experts $\{E^\eta | \eta \in \mathcal{H}\}$ by invoking Algorithm 2 for each step size $\eta \in \mathcal{H}$

for $t = 1, \ldots, T$ **do**

Receive $\mathbf{x}_t^\eta$ from each expert E^η

Output

$$\mathbf{x}_t = \sum_{\eta \in \mathcal{H}} w_t^\eta \mathbf{x}_t^\eta$$

Query the gradient of $f_t(\cdot)$ at $\mathbf{x}_t$

Construct the surrogate loss $l_t(\cdot)$

$$l_t(\mathbf{x}_t^\eta, \eta) = \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^\eta - \mathbf{x}_t \rangle + \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\|$$

Construct the new target function

$$F_t(\mathbf{w}_t) = \sum_{\eta \in \mathcal{H}} w_t^\eta l_t(\mathbf{x}_t^\eta, \eta)$$

Update the weight of each expert by

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\left[\mathbf{w}_t - \alpha_t \nabla F_t(w_t)\right]$$

$\mathcal{W}$ is the probability simplex of which dimension is N(the number of experts)

Send gradient $\nabla f_t(\mathbf{x}_t)$ to each expert E^η

end for

Algorithm 2 Expert-algorithm(As same as Ader)

Require: The step size η

Let $\mathbf{x}_1^\eta$ be any point in $\mathcal{X}$

for $t = 1, \ldots, T$ **do**

Submit $\mathbf{x}_t^\eta$ to the meta-algorithm

Receive gradient $\nabla f_t(\mathbf{x}_t)$ from the meta-algorithm

$$\mathbf{x}_{t+1}^\eta = \Pi_{\mathcal{X}}\left[\mathbf{x}_t^\eta - \eta \nabla f_t(\mathbf{x}_t)\right]$$

end for

$N = \lceil \frac{1}{2} \log_2(1 + 4T/7) \rceil + 1$

1. Meta-algorithm

$$l_t(\mathbf{x}_t^\eta, \eta) = \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^\eta - \mathbf{x}_t \rangle + \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\| \tag{1}$$

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*) \leq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_t \rangle - \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t \rangle \tag{2}$$

$$\|\nabla F_t(\mathbf{w}_t)\| = \sqrt{\sum_{\eta\in\mathcal{H}} l_t(\mathbf{x}_t^\eta,\eta)^2} \le (G+1)D\sqrt{N} = G_{meta} \tag{3}$$

$$\|\mathbf{w}-\mathbf{y}\|_{\mathbf{w},\mathbf{y}\in\mathcal{W}} \le 2 = D_{meta} \tag{4}$$

Lemma 1 *With assumption $\|\nabla f_t(\mathbf{x})\| \le G$ and $\|\mathbf{x}-\mathbf{y}\|_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \le D$ Online gradient descent with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$ guarantees the following for all $T \ge 1$:*

$$regret_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \le \frac{3}{2} GD\sqrt{T}$$

So

$$\sum_{t=1}^{T} F_t(\mathbf{w}_t) - F_t(\mathbf{w}^*) \le \frac{3}{2} D_{meta} G_{meta}\sqrt{T} \tag{5}$$

when $\alpha_t = \frac{D_{meta}}{G_{meta}\sqrt{t}}$

$$\begin{aligned}
F_t(\mathbf{w}_t) - F_t(\mathbf{w}^*) &= \sum_{\eta\in\mathcal{H}} w_t^\eta l_t(\mathbf{x}_t^\eta) - l_t(\mathbf{x}_t^{\eta^*}) \\
&\overset{(1)}{=} \sum_{\eta\in\mathcal{H}} w_t^\eta \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^\eta - \mathbf{x}_t\rangle + \sum_{\eta\in\mathcal{H}} w_t^\eta \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\| - \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^{\eta^*} - \mathbf{x}_t\rangle - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&= \langle \nabla f_t(\mathbf{x}_t), \sum_{\eta\in\mathcal{H}} w_t^\eta \mathbf{x}_t^\eta - \mathbf{x}_t\rangle + \sum_{\eta\in\mathcal{H}} w_t^\eta \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\| - \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^{\eta^*} - \mathbf{x}_t\rangle - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&= \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_t\rangle + \sum_{\eta\in\mathcal{H}} w_t^\eta \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\| - \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^{\eta^*} - \mathbf{x}_t\rangle - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&\overset{(2)}{\ge} f_t(\mathbf{x}_t) + \sum_{\eta\in\mathcal{H}} w_t^\eta \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\| - f_t(\mathbf{x}_t^{\eta^*}) - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\|
\end{aligned} \tag{6}$$

The regret with switching cost about $f_t(\mathbf{x})$ is

$$
\begin{aligned}
& f_t(\mathbf{x}_t) + \|\mathbf{x}_t - \mathbf{x}_{t-1}\| - f_t(\mathbf{x}_t^{\eta^*}) - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&= f_t(\mathbf{x}_t) + \|\sum_{\eta\in\mathcal{H}} w_t^\eta \mathbf{x}_t^\eta - \sum_{\eta\in\mathcal{H}} w_{t-1}^\eta \mathbf{x}_{t-1}^\eta\| - f_t(\mathbf{x}_t^{\eta^*}) - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&\le f_t(\mathbf{x}_t) + \|\sum_{\eta\in\mathcal{H}} w_t^\eta \mathbf{x}_t^\eta - \sum_{\eta\in\mathcal{H}} w_t^\eta \mathbf{x}_{t-1}^\eta\| + \|\sum_{\eta\in\mathcal{H}} w_t^\eta \mathbf{x}_{t-1}^\eta - \sum_{\eta\in\mathcal{H}} w_{t-1}^\eta \mathbf{x}_{t-1}^\eta\| - f_t(\mathbf{x}_t^{\eta^*}) - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&\le f_t(\mathbf{x}_t) + \sum_{\eta\in\mathcal{H}} w_t^\eta \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\| + \sum_{\eta\in\mathcal{H}} |w_t^\eta - w_{t-1}^\eta| \|\mathbf{x}_{t-1}^\eta\| - f_t(\mathbf{x}_t^{\eta^*}) - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&\le f_t(\mathbf{x}_t) + \sum_{\eta\in\mathcal{H}} w_t^\eta \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\| + \sqrt{N}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|D - f_t(\mathbf{x}_t^{\eta^*}) - \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&\overset{(6)}{\le} F_t(\mathbf{w}_t) - F_t(\mathbf{w}^*) + \sqrt{N}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|D
\end{aligned} \tag{7}
$$

So

$$
\begin{aligned}
& \sum_{t=1}^{T} f_t(\mathbf{x}_t) + \|\mathbf{x}_t - \mathbf{x}_{t-1}\| - \sum_{t=1}^{T} f(x_t^{\eta^*}) + \|\mathbf{x}_t^{\eta^*} - \mathbf{x}_{t-1}^{\eta^*}\| \\
&\le 3D_{meta}G_{meta}\sqrt{T} + \sum_{t=1}^{T} \sqrt{N}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|D \\
&\le 3(G+1)D\sqrt{NT} + \sum_{t=1}^{T} \alpha_t \|\nabla F_{t-1}(\mathbf{w}_{t-1})\|\sqrt{N}D \\
&\le 3(G+1)D\sqrt{NT} + \sum_{t=1}^{T} \frac{\sqrt{N}DD_{meta}}{\sqrt{t}} \\
&\le 3(G+1)D\sqrt{NT} + 2D\sqrt{NT}
\end{aligned} \tag{8}
$$

So for any expert we have

$$
\sum_{t=1}^{T} f_t(\mathbf{x}_t) + \|\mathbf{x}_t - \mathbf{x}_{t-1}\| - \sum_{t=1}^{T} f(x_t^\eta) + \|\mathbf{x}_t^\eta - \mathbf{x}_{t-1}^\eta\| \le 3(G+1)D\sqrt{NT} + 2D\sqrt{NT} \tag{9}
$$

2. Expert-algorithm(As same as Ader)

$$
\eta^*(P_T) = \sqrt{\frac{7D^2 + 4DP_T}{2TG^2}}. \tag{10}
$$

for any possible value of P_T, there exists a step size $\eta_k \in \mathcal{H}$, such that

$$
\eta_k = \frac{2^{k-1}D}{G}\sqrt{\frac{7}{2T}} \le \eta^*(P_T) \le 2\eta_k \tag{11}
$$

$$
\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t^{\eta_k}) + \|\mathbf{x}_t^{\eta_k} - \mathbf{x}_{t-1}^{\eta_k}\| - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \\
\leq &\frac{7D^2}{4\eta_k} + \frac{DP_T}{\eta_k} + \frac{\eta_k TG^2}{2} + \sum_{t=1}^{T} \|\mathbf{x}_t^{\eta_k} - \mathbf{x}_{t-1}^{\eta_k}\| \\
\leq &\frac{7D^2}{2\eta^*(P_T)} + \frac{2DP_T}{\eta^*(P_T)} + \frac{\eta^*(P_T)TG^2}{2} + T\eta^* G \\
\leq &\frac{3G}{4}\sqrt{2T(7D^2 + 4DP_T)} + \sqrt{\frac{T(7D^2 + 4DP_T)}{2}}
\end{aligned} \tag{12}
$$

3. Algorithm

Combine (9) and(12)

$$
\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) + \|\mathbf{x}_t - \mathbf{x}_{t-1}\| - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \\
\leq &\frac{3G}{4}\sqrt{2T(7D^2 + 4DP_T)} + \sqrt{\frac{T(7D^2 + 4DP_T)}{2}} + 3(G+1)D\sqrt{NT} + 2D\sqrt{NT}
\end{aligned} \tag{13}
$$