Recursive Exponential Weighting for Online non-convex optimization

Speaker: Yi-Shan Wu

Institute of Information Science

Academia Sinica

Taiwan

Feb, 02, 2018

Online Convex Optimization

For $t = 1$ to T

- player chooses a point x_t from $\mathcal{K} \subset \mathbb{R}^n$(bounded convex set)
- cost function $f_t : \mathcal{K} \mapsto \mathbb{R}$ is revealed (bounded convex function)

Regret:

$$R_T = \sup_{f_1,\cdots,f_T \in \mathcal{F}} \left\{ \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x) \right\}$$

- OCO lower bound: $O(\sqrt{T})$ obtained by OGD, SGD,$\cdots$
- If further assumes that f_t is strictly convex, OCO lower bound: $O(\ln T)$

A natural question: Is it possible to obtain $O(\sqrt{T})$ regret bound if the convexity assumption on the cost function is relaxed?

Recursive Exponential Weighting Algorithm

Assumption: cost function $f_t \in \mathcal{F}$, $f_t(x) : \mathcal{K} \mapsto [0, B]$, and it is L-Lipschitz

REW Algorithm:

1. **Set Discretization from $\mathcal{K}$ to $\mathcal{I}$**

   Idea: Group highly correlated decisions into one set.

   $\Longrightarrow$ reduce the original problem to expert problem with finite experts

2. **Set partition according to layers**

   Idea: Divide and conquer, recursively digging into subsets that are chosen.

Analysis:

$$R_{ImC} = \sup_{f_1,\cdots,f_T \in \mathcal{F}} \left\{ \sum_{t=1}^{T} c_t(I_t) - \min_{i \in \mathcal{I}} \sum_{t=1}^{T} c_t(i) \right\}$$

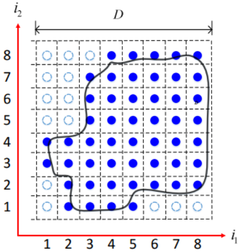
$$R_{ImD} = \sup_{f_1,\cdots,f_T \in \mathcal{F}} \left\{ \min_{i \in \mathcal{I}} \sum_{t=1}^{T} c_t(i) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x) \right\}$$

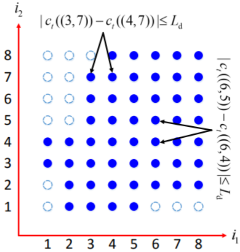
Set Discretization

- Dimension of $\mathcal{K}$: n
- edge length of sub-cube: $D/2^m$
- index of sub-cube $D_{\mathbf{i}}$: $\mathbf{i} = (i_1, i_2, \cdots, i_n), \quad 1 \le i_j \le 2^m, \forall j \in [n]$
- $\mathcal{I} = \{\mathbf{i} : D_{\mathbf{i}} \bigcap \mathcal{K} \ne \emptyset\} \Rightarrow |\mathcal{I}|$ experts
- cost of $\mathbf{i} \in \mathcal{I}$ at time t : $c_t(\mathbf{i})$

$$|c_t(\mathbf{p}) - c_t(\mathbf{q})| \le L_d \|\mathbf{p} - \mathbf{q}\|_1, \quad \mathbf{p}, \mathbf{q} \in \mathcal{I}$$

where $L_d = L\sqrt{n} 2D/2^m$





Set Partition for $\mathcal{I}$

$$\mathcal{I}_\ell(\mathbf{i}) = \{\mathbf{p} \in \mathcal{I} : 1+(i_j-1)2^{m-\ell} \le p_j \le 2^{m-\ell}+(i_j-1)2^{m-\ell}, j = 1, 2, \cdots, n\}.$$

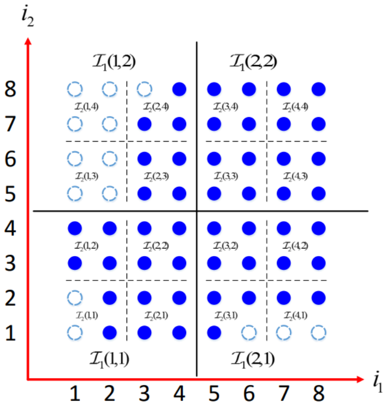


Figure: Set partition when $n = 2$ and $m = 3$

Set Partition for $\mathcal{I}$

- $\mathcal{U}_\ell(\mathbf{i})$: the layer-ℓ subset containing lower-layer subset $\mathcal{I}_{\ell+1}(\mathbf{i})$

  $$\mathcal{U}_\ell(\mathbf{i}) = \{\mathbf{j} : \mathcal{I}_{\ell+1}(\mathbf{i}) \subset \mathcal{I}_\ell(\mathbf{j}) : \mathcal{I}_{\ell+1}(\mathbf{i}) \neq \emptyset\}.$$

  Ex: $\mathcal{U}_1(3,2) = \{(2,1)\}, \qquad (\mathcal{I}_2(3,2) \subset \mathcal{I}_1(2,1))$
- $\mathcal{M}_\ell(\mathbf{k})$: the layer-ℓ subsets that contain point $\mathbf{k}$
  Ex: $\mathcal{M}_2(5,3) = \{(3,2)\}$
- $\mathcal{D}_\ell(\mathbf{i})$: the index set of non-empty layer-$(\ell+1)$ subsets within $\mathcal{I}_\ell(\mathbf{i})$

  $$\mathcal{D}_\ell(\mathbf{i}) = \{\mathbf{j} : \mathcal{I}_{\ell+1}(\mathbf{j}) \subset \mathcal{I}_\ell(\mathbf{i}) : \mathcal{I}_{\ell+1}(\mathbf{j}) \neq \emptyset\}.$$

  Ex: $\mathcal{D}_2(3,2) = \{(5,3),(6,3),(5,4),(6,4)\}$

REW-selection

$$\bar{C}_{\ell+1,t}(\mathbf{i}) = \sum_{\tau=1}^{t} \bar{c}_{\ell+1,\tau}(\mathbf{i})$$

for $t = 1$ to T **do**
    $\boldsymbol{s} = \mathbf{1}$
    `//Recursively select the subsets in all layers`
    for $l = 0$ to $m - 1$ **do**
        Select a non-empty subset $\mathcal{I}_{l+1}(\boldsymbol{i}) \subset \mathcal{I}_l(\boldsymbol{s})$ with probability

$$p_t(\mathcal{I}_{l+1}(\boldsymbol{i})) = \frac{\exp\left(-\eta_t \bar{C}_{l+1,t-1}(\boldsymbol{i})\right)}{\sum_{\boldsymbol{i} \in \mathcal{D}_l(\boldsymbol{s})} \exp\left(-\eta_t \bar{C}_{l+1,t-1}(\boldsymbol{i})\right)}$$

        if subset $\mathcal{I}_{l+1}(\boldsymbol{i})$ is selected **then**
            Set $\boldsymbol{s} = \boldsymbol{i}$
        end if
    end for

Figure: Selection

$$\Pr[\mathbf{I}_t = \mathbf{k} | \mathbf{I}_t \in \mathcal{I}_{\ell+1}(\mathbf{i})] \\ = \Pi_{i=\ell+2}^{m} \Pr[\mathbf{I}_t \in \mathcal{M}_i(\mathbf{k}) | \mathbf{I}_t \in \mathcal{M}_{i-1}(\mathbf{k})]$$

`//Get the expected normalized cost for all subsets in all layers`

for $l = 0$ to $m-1$ **do**

  for each non-empty subset $\mathcal{I}_{l+1}(\boldsymbol{i})$ in layer-$(l+1)$ **do**

    Calculate $\Pr[\boldsymbol{I}_t = \boldsymbol{k} | \boldsymbol{I}_t \in \mathcal{I}_{l+1}(\boldsymbol{i})]$ based on Equation (3.3)

    Calculate

$$\bar{c}_{l+1,t}(\boldsymbol{i}) \\ = \sum_{\boldsymbol{k} \in \mathcal{I}_{l+1}(\boldsymbol{i})} \frac{c_t(\boldsymbol{k}) - \min_{\boldsymbol{k} \in \mathcal{U}_l(\boldsymbol{i})} c_t(\boldsymbol{k})}{n 2^{m-l} L_{\mathsf{d}}} \cdot \Pr[\boldsymbol{I}_t = \boldsymbol{k} | \boldsymbol{I}_t \in \mathcal{I}_{l+1}(\boldsymbol{i})]$$

  end for

end for

end for

Figure: cumulative cost

Analysis

$R_T = R_{ImC} + R_{ImD}$

$$
\begin{aligned}
R_{ImC} &= \sup_{f_1,\cdots,f_T \in \mathcal{F}} \left\{ \sum_{t=1}^{T} c_t(I_t) - \min_{i \in \mathcal{I}} \sum_{t=1}^{T} c_t(i) \right\} \\
&\leq (4n^2 + \frac{1}{2}n) 2^m L_d \sqrt{T} + 2n^2 \cdot 2^m L_d.
\end{aligned}
$$

$$
\begin{aligned}
R_{ImD} &= \sup_{f_1,\cdots,f_T \in \mathcal{F}} \left\{ \min_{i \in \mathcal{I}} \sum_{t=1}^{T} c_t(i) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x) \right\} \\
&\leq \frac{1}{2} L_d T
\end{aligned}
$$

Note that $L_d = \frac{2L\sqrt{n}D}{2^m}$
Choose $m = \log_2 \sqrt{T}$, then we obtain $O(n^2\sqrt{T})$

R_{ImC}

$$
\begin{aligned}
&= \sum_{t \in T} \mathbb{E}[c_t(I_t)] - \sum_{t \in T} c_t(i^*) \\
&= \sum_{t \in T} \mathbb{E}[c_t(I_t) | I_t \in \mathcal{M}_0(i^*)] - \sum_{t \in T} c_t(i^*) \\
&\leq \sum_{t \in T} \mathbb{E}[c_t(I_t) | I_t \in \mathcal{M}_1(i^*)] + (2n + \frac{1}{4})\sqrt{T} n 2^m L_d + \frac{\ln |\mathcal{D}_\ell(0)|}{\eta_1} - \sum_{t \in T} c_t(i^*)
\end{aligned}
$$

$$\vdots$$

$$\leq (4n^2 + \frac{n}{2}) 2^m L_d \sqrt{T} + 2n^2 2^m L_d$$

REW-cumulative expected normalized cost

$$
\begin{aligned}
&\sum_{t\in\mathcal{T}} \mathbb{E}[c_t(\boldsymbol{I}_t)|\boldsymbol{I}_t \in \mathcal{I}_l(\boldsymbol{i})] \\
=&\sum_{t\in\mathcal{T}} \sum_{\boldsymbol{j}\in\mathcal{D}_l(\boldsymbol{i})} \Pr[\boldsymbol{I}_t \in \mathcal{I}_{l+1}(\boldsymbol{j})|\boldsymbol{I}_t \in \mathcal{I}_l(\boldsymbol{i})]\mathbb{E}\left[c_t(\boldsymbol{I}_t)|\boldsymbol{I}_t \in \mathcal{I}_{l+1}(\boldsymbol{j})\right] \\
\overset{(a)}{=}&\sum_{t\in\mathcal{T}} \sum_{\boldsymbol{j}\in\mathcal{D}_l(\boldsymbol{i})} \frac{\exp\left(-\eta_t \bar{C}_{l+1,t-1}(\boldsymbol{j})\right)}{\sum_{\boldsymbol{j}\in\mathcal{D}_l(\boldsymbol{i})} \exp\left(-\eta_t \bar{C}_{l+1,t-1}(\boldsymbol{j})\right)} \cdot \left[\bar{c}_{l+1,t}(\boldsymbol{j}) \cdot n2^{m-l}L^{\mathsf{d}} + \min_{\boldsymbol{k}\in\mathcal{I}_l(\boldsymbol{i})} c_t(\boldsymbol{k})\right] \\
\overset{(b)}{\leq}&\sum_{t\in\mathcal{T}} \left[-\frac{1}{\eta_t}\ln \sum_{\boldsymbol{j}\in\mathcal{D}_l(\boldsymbol{i})} \frac{\exp\left(-\eta_t \bar{C}_{l+1,t-1}(\boldsymbol{j})\right)}{\sum_{\boldsymbol{j}\in\mathcal{D}_l(\boldsymbol{i})} \exp\left(-\eta_t \bar{C}_{l+1,t-1}(\boldsymbol{j})\right)} \exp\left(-\eta_t \bar{c}_{l+1,t}(\boldsymbol{j})\right) + \frac{\eta_t}{8}\cdot 1^2\right] \cdot n2^{m-l}L_{\mathsf{d}} \\
&\qquad + \sum_{t\in\mathcal{T}} \min_{\boldsymbol{k}\in\mathcal{I}_l(\boldsymbol{i})} c_t(\boldsymbol{k}) \\
=&\sum_{t\in\mathcal{T}} \left[-\frac{1}{\eta_t}\ln \sum_{\boldsymbol{j}\in\mathcal{D}_l(\boldsymbol{i})} \frac{\exp\left(-\eta_t \bar{C}_{l+1,t}(\boldsymbol{j})\right)}{\sum_{\boldsymbol{j}\in\mathcal{D}_l(\boldsymbol{i})} \exp\left(-\eta_t \bar{C}_{l+1,t-1}(\boldsymbol{j})\right)} + \frac{\eta_t}{8}\cdot 1^2\right] \cdot n2^{m-l}L_{\mathsf{d}} + \sum_{t\in\mathcal{T}} \min_{\boldsymbol{k}\in\mathcal{I}_l(\boldsymbol{i})} c_t(\boldsymbol{k}) \\
\overset{(c)}{=}&\sum_{t\in\mathcal{T}} \left[\Phi_t(\eta_t) - \Phi_{t-1}(\eta_t) + \frac{\eta_t}{8}\cdot 1^2\right] \cdot n2^{m-l}L_{\mathsf{d}} + \sum_{t\in\mathcal{T}} \min_{\boldsymbol{k}\in\mathcal{I}_l(\boldsymbol{i})} c_t(\boldsymbol{k}) \\
\overset{(d)}{\leq}&\left\{\Phi_T(\eta_T) + \sum_{t=2}^{T}\left[n\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{\eta_t}{8}\right] + \frac{1}{\eta_1}\ln|\mathcal{D}_l(\boldsymbol{i})|\right\} \cdot n2^{m-l}L_{\mathsf{d}} + \sum_{t\in\mathcal{T}} \min_{\boldsymbol{k}\in\mathcal{I}_l(\boldsymbol{i})} c_t(\boldsymbol{k}) \\
\overset{(e)}{\leq}&\left\{\sum_{t\in\mathcal{T}} \mathbb{E}\left[\frac{c_t(\boldsymbol{I}_t) - \min_{\boldsymbol{k}\in\mathcal{I}_l(\boldsymbol{i})} c_t(\boldsymbol{k})}{n2^{m-l}L_{\mathsf{d}}} |\boldsymbol{I}_t \in \mathcal{I}_{l+1}(\boldsymbol{j})\right]\right. \\
&\qquad \left. + \frac{1}{\eta_T}n + \sum_{t=2}^{T}\left[n\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{\eta_t}{8}\right] + \frac{1}{\eta_1}\ln|\mathcal{D}_l(\boldsymbol{i})|\right\} \cdot n2^{m-l}L_{\mathsf{d}} + \sum_{t\in\mathcal{T}} \min_{\boldsymbol{k}\in\mathcal{I}_l(\boldsymbol{i})} c_t(\boldsymbol{k}) \\
\overset{(f)}{\leq}&\sum_{t\in\mathcal{T}} \mathbb{E}\left[c_t(\boldsymbol{I}_t)|\boldsymbol{I}_t \in \mathcal{I}_{l+1}(\boldsymbol{j})\right] + \left(2n + \frac{1}{4}\right)\sqrt{T}\cdot n2^{m-l}L_{\mathsf{d}} + \frac{1}{\eta_1}\ln|\mathcal{D}_l(\boldsymbol{i})| \cdot n2^{m-l}L_{\mathsf{d}}
\end{aligned}
$$

log-sum-exp approximation

$$\mathbf{MWC}: \quad \max_{f \in \mathcal{F}} \sum_{r \in R} x_r(f). \qquad (1)$$

An equivalent formulation is

$$\mathbf{MWC-EQ}: \quad \max_{p \geq 0} \sum_{f \in \mathcal{F}} p_f \sum_{r \in R} x_r(f) \qquad (2)$$
$$\text{s.t.} \quad \sum_{f \in \mathcal{F}} p_f = 1,$$

MWC: maximum weighted configuration

$$\max_{f \in \mathcal{F}} \sum_{r \in R} x_r(f) \approx \frac{1}{\beta} \log \left(\sum_{f \in \mathcal{F}} \exp\left(\beta \sum_{r \in R} x_r(f) \right) \right) \triangleq g_\beta(\mathbf{x}), \qquad (3)$$

where β is a positive constant and $\mathbf{x} \triangleq [\sum_{r \in R} x_i(f), f \in \mathcal{F}]$.

Theorem 1: For the log-sum-exp function $g_\beta(\boldsymbol{x})$, we have

- its conjugate function[3] is given by

$$g_\beta^*(\boldsymbol{p}) = \begin{cases} \frac{1}{\beta}\sum_{f\in\mathcal{F}} p_f \log p_f & \text{if } \boldsymbol{p} \geq 0 \text{ and } \boldsymbol{1}^T\boldsymbol{p} = 1 \\ \infty & \text{otherwise.} \end{cases} \tag{7}$$

- it is a convex and closed function; hence, the conjugate of its conjugate $g_\beta^*(\boldsymbol{p})$ is itself, i.e., $g_\beta(\boldsymbol{x}) = g_\beta^{**}(\boldsymbol{x})$. Specifically,

$$\begin{aligned} g_\beta(\boldsymbol{x}) \quad = \quad & \max_{\boldsymbol{p}\geq 0} \sum_{f\in\mathcal{F}} p_f \sum_{r\in R} x_r(f) - \frac{1}{\beta}\sum_{f\in\mathcal{F}} p_f \log p_f \\ & \text{s.t.} \sum_{f\in\mathcal{F}} p_f = 1. \end{aligned} \tag{8}$$

By solving KKT condition,

we have

$$p_f^*(\boldsymbol{x}) = \frac{\exp\left(\beta\sum_{r\in R} x_r(f)\right)}{\sum_{f'\in\mathcal{F}} \exp\left(\beta\sum_{r\in R} x_r(f')\right)}, \forall f \in \mathcal{F}. \tag{12}$$