# Supplementary Material for "Learning efficient contextual categorical sequence generation models with pseudo-action policy gradient"

**Anonymous Author(s)**
Affiliation
Address
email

## A Qualitative results

### A.1 Pseudo sentences produced by ARS-K algorithm.

Greedy sentence: a man riding a motorcycle with a dog.
Pseudo sentence 1: a man riding a motorcycle with hay on it.
Pseudo sentence 2: a man riding a motorcycle with pack of sheep.

Greedy sentence: a group of people flying kites in a field.
Pseudo sentence 1: a group of teenagers standing in a field flying kites.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

Greedy sentence: a man and woman are standing in a of a table.
Pseudo sentence 1: a man and woman are standing in an market.
Pseudo sentence 2: a man and woman are standing in the street.
Pseudo sentence 3: a man and woman are standing in to a tent.
Pseudo sentence 4: a man and woman are standing in front of a table.


Greedy sentence: a city that has a large white building on it.
Pseudo sentence 1: a city with a red traffic and a large building.
Pseudo sentence 2: a city intersection with a traffic light and a street sign.
Pseudo sentence 3: a city bus is driving down the street.
Pseudo sentence 4: a city street with a city street with cars parked on it.
Pseudo sentence 5: a city road with a traffic light and a street sign.

## B Algorithms

**Algorithm 1:** ARS-K/ARSM(K=V) policy gradient for fine-tuning a contextual categorical sequence generation model with a discrete-action space of $V$ actions.

**input :** MLE pretrained policy parameter $\boldsymbol{\theta}$, $K$
**output :** Fine-tuned policy parameter $\boldsymbol{\theta}$;

**while** *not converged* **do**
  Receive random sample $\boldsymbol{x}, \boldsymbol{y}$;
  $\boldsymbol{\theta}_{\text{update}} = 0$;
  **for** $t = 1 : T$ **do**
    $\boldsymbol{\phi}_t = \mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:t-1}, x)$;
    $\boldsymbol{g}_{\boldsymbol{\phi}_t} = 0$;
    Sample $\boldsymbol{\pi}_t \sim \text{Dirichlet}(\mathbf{1}_V)$;
    Let $z_t = \text{argmin}_{i \in \{1,...,V\}}(\ln \pi_{ti} - \phi_{ti})$;
    Let $j_1, ..., j_K$ be randomly selected K reference categories without replacement;
    **for** $k = 1, \ldots, K$ **do**
      **for** $v = 1, \ldots, V$ *(in parallel)* **do**
        Let $z_t^{v \leftrightharpoons j_k} := \text{argmin}_{i \in \{1,...,V\}}(\ln \pi_i^{v \leftrightharpoons j_k} - \phi_{ti})$ as the $(v, k)$th pseudo action;
      **end for**
    **end for**
    Let $S_t = \text{unique}(\{z_t^{v \leftrightharpoons j}\}_{v,j})$ which means $S_t$ is the set of all unique values in $\{z_t^{v \leftrightharpoons j}\}_{v,j}$, denote the cardinality of $S_t$ as $|S_t|$;
    **for** $s \in \{1, ..., |S_t|\}$ *(in parallel)* **do**
      Let $\tilde{z}_{ts} = S_t(s)$ be the $s$th unique pseudo action at time $t$;
      If $t < T$, sample $\boldsymbol{z}_{t+1:T}^s \sim p_{\boldsymbol{\theta}}(\boldsymbol{z}_{t+1:T} \mid \boldsymbol{z}_{1:t-1}, \tilde{z}_{ts}, \boldsymbol{x})$;
      Let $f(z_t^{v \leftrightharpoons j_k}) = r(\boldsymbol{z}_{1:t-1}, \boldsymbol{z}_{t:T}^s \mid \boldsymbol{x}, \boldsymbol{y})$ if $z_t^{v \leftrightharpoons j_k} = \tilde{z}_{ts}$;
    **end for**
    **for** $k = 1, \ldots, K$ **do**
      Let $\bar{f}_t = \frac{1}{V} \sum_{v=1}^{V} f(z_t^{v \leftrightharpoons j_k})$;
      Let $g_{\phi_{tv}} = g_{\phi_{tv}} + \frac{1}{K}\big(f(z_t^{v \leftrightharpoons j_k}) - \bar{f}_t\big)(1 - V\pi_{j_k})$;
      $\boldsymbol{\theta}_{\text{update}} = \boldsymbol{\theta}_{\text{update}} + \eta_\theta \nabla_{\boldsymbol{\theta}} \boldsymbol{\phi}_t \boldsymbol{g}_{\boldsymbol{\phi}_t}$, with step-size $\eta_\theta$
    **end for**
  **end for**
  $\boldsymbol{\theta} = \boldsymbol{\theta} + \boldsymbol{\theta}_{\text{update}}$
**end while**

**Algorithm 2:** Binary-tree-ARSM policy gradient for fine-tuning a binary-tree contextual categorical sequence generation model.

**input :** MLE pretrained policy parameter $\boldsymbol{\theta}$, $K$
**output :** Fine-tuned policy parameter $\boldsymbol{\theta}$;

**while** *not converged* **do**
  Receive random sample $\boldsymbol{x}, \boldsymbol{y}$;
  $\boldsymbol{\theta}_{\text{update}} = 0$;
  **for** $t = 1 : T$ **do**
    $\boldsymbol{\phi}_t = \mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:t-1}, x)$;
    **for** $l = 1 : D$ **do**
      Sample $\pi_{tl} \sim \text{Uniform}(0, 1)$;
      Let $b_{tl}^{(1)} = \mathbf{1}_{[\pi_{tl} < \sigma(\phi_{t, b_{t(1:l-1)}})]}$;
      Let $b_{tl} = b_{tl}^{(1)}$;
      Let $b_{tl}^{(2)} = \mathbf{1}_{[\pi_{tl} > \sigma(-\phi_{t, b_{t(1:l-1)}})]}$;
      **if** $b_{tl}^{(1)} \neq b_{tl}^{(2)}$ **then**
        If $l < D$, sample $b_{t(l+1:D)}^{(j)} \sim p_{\boldsymbol{\theta}}(b_{t(l+1:D)}^{(1)} \mid \boldsymbol{z}_{1:t-1}, b_{t,1:l-1}, b_{tl}^{(j)}, \boldsymbol{x})$, $j = 1, 2$;
        Let $z_t^{(j)} = \nu(b_{t(1:D)}^{(j)})$, $j = 1, 2$;
        If $t < T$, sample $\boldsymbol{z}_{t+1:T}^{(j)} \sim p_{\boldsymbol{\theta}}(\boldsymbol{z}_{t+1:T} \mid \boldsymbol{z}_{1:t-1}, z_t^{(j)}, \boldsymbol{x})$, $j = 1, 2$;
        Let $f_{tl}^{(j)} = r(\boldsymbol{z}_{1:t-1}, \boldsymbol{z}_{t:T}^{(j)} \mid \boldsymbol{x}, \boldsymbol{y})$;
        Let $g_{\phi_{t, b_{t(1:l-1)}}} = \frac{1}{2}(f_{tl}^{(1)} - f_{tl}^{(2)})(1 - 2\pi_{tl})$;
        $\boldsymbol{\theta}_{\text{update}} = \boldsymbol{\theta}_{\text{update}} + \eta_\theta \nabla_{\boldsymbol{\theta}} \phi_{t, b_{t(1:l-1)}} g_{\phi_{t, b_{t(1:l-1)}}}$, with step-size $\eta_\theta$
      **end if**
    **end for**
    $\boldsymbol{\theta} = \boldsymbol{\theta} + \boldsymbol{\theta}_{\text{update}}$
  **end for**
**end while**