Fixed-Point Decoding

First Author and Second Author and Third Author
Singapore University of Technology and Design
{blah_bleh, ..., bleh_blue}@mymail.sutd.edu.sg

Abstract

Main contributions and findings (200 words).

1 Introduction

Blah.

2 Related Work

Blah.

3 Methodology

Fix a uniform-cost random-access machine model of computation.

Let ℓ denote $\log \Pr$, and juxtaposition denote concatenation.

Let V be a finite vocabulary totally ordered by $\preceq$, and $\Pr$ be a next-token kernel, and $\gamma \in \mathbb{R}_{(0,1)}$ be a discount factor, and $\varepsilon \in \mathbb{R}_{(0,1)}$ be a machine epsilon, and $\delta \in \mathbb{R}_{(0,1)}$ be a confidence tolerance.

We define the update operator

$$\mathcal{U} : \mathbb{R}^{V^*} \to \mathbb{R}^{V^*}$$
$$\mathcal{U}(G)[x] = \max_{y \in V} \left[\ell(y \mid x) + \gamma G(xy)\right] \tag{1}$$

Proposition 1 (Existence and Uniqueness)**.** There exists a unique fixed point of $\mathcal{U}$.

Proof. Trivial. □

We define the decoding rule

$$f : V^* \to V$$
$$f(x) \in \arg\max_{y \in V} \left[\ell(y \mid x) + \gamma G(xy)\right] \tag{2}$$

where $\gamma \in \mathbb{R}_{(0,1)}$, and $G = \mathcal{U}_\gamma(G)$.

We define the lower bounding sequence

$$l_t : V^* \to \mathbb{R}$$
$$l_t(x) = \begin{cases} -\frac{\sup_{x \in V^*, y \in V} |\ell(y|x)|}{1-\gamma}, & t = 0 \\ \mathcal{U}(l_{t-1}), & t \geq 1 \end{cases} \tag{3}$$

where $t \in \mathbb{N}$.

We define the upper bounding sequence

$$u_t : V^* \to \mathbb{R}$$
$$u_t(x) = \begin{cases} 0, & t = 0 \\ \min\left[u_{t-1}, \mathcal{U}(u_{t-1})\right], & t \geq 1 \end{cases} \tag{4}$$

where $t \in \mathbb{N}$.

Proposition 2 (Convergence)**.** Let $G \in \mathbb{R}^{V^*}$ be the fixed point of $\mathcal{U}$. Then,

$$l_t(x) \leq l_{t+1}(x) \leq G(x) \leq u_{t+1}(x) \leq u_t(x)$$

and

$$\lim_{t \to \infty} l_t(x) = \lim_{t \to \infty} u_t(x) = G(x)$$

for each $t \in \mathbb{N}$, and for each $x \in V^*$.

Proof. Trivial. □

Let $Q_G(x, y)$ denote

$$Q_G(x, y) = \ell(y \mid x) + \gamma G(xy) \tag{5}$$

Algorithm 1 DECODE(x)

Input: $x \in V^*$ *# Current prefix*
Output: $\hat{y} \in V$ *# Next token*
 # Initialise step
$t \leftarrow 0$
while true:
  # Get candidate
  $\hat{y} \leftarrow \arg\max_{y \in V} Q_{l_t}(x, y)$
  # Get candidate-competitor gap
  $\delta \leftarrow Q_{l_t}(x, \hat{y}) - \max_{z \neq \hat{y}} Q_{u_t}(x, z)$
  # Return candidate if gap is tolerable
  if $\delta > \varepsilon$:
    return $\hat{y}$
  # Tighten lower bound
  $l_{t+1} \leftarrow \mathcal{U}(l_t)$
  # Tighten upper bound
  $u_{t+1} \leftarrow \min\left[u_t, \mathcal{U}(u_t)\right]$
  # Increment step
  $t \leftarrow t + 1$

Proposition 3 (Soundness)**.** Let $x \in V^*$. If DECODE halts on x, then DECODE$(x) = f(x)$.

Proposition 4 (Completeness)**.** Let $x \in V^*$. If $f(x)$ exists, then DECODE halts on x.

For tractability, we propose a Probably Approximately Correct (PAC) variant of our algorithm.

We define the error decomposition

$$\boldsymbol{e} = \varepsilon(p_{\text{tail}}, p_{\text{stat}}, p_{\text{gap}}) \tag{6}$$

where $(p_{\text{tail}}, p_{\text{stat}}, p_{\text{gap}})$ lies on the 2-simplex.

Algorithm 2 PAC-DECODE(x)

Input: $x \in V^*$ *# Current prefix*
Output: $\hat{y} \in V$ *# Next token*
$S \leftarrow \sup_{x \in V^*, y \in V} |\ell(y \mid x)|$
$H \leftarrow \min\left\{n \in \mathbb{N} \mid \frac{\gamma^{n+1} S}{1-\gamma} \leq e_1\right\}$
for $y \in V$:
  $U_\infty(x, y) \leftarrow \ell(y \mid x) + \frac{\gamma S}{1-\gamma}$
$\tau(x) \leftarrow \max_{z \in V}\left[U(x, z)\right] - e_3$
$C(x) \leftarrow \{y \in V \mid \tau(x) \leq U_\infty(x, y)\}$
for $y \in C(x)$:
  $n(y) \leftarrow 0$
  $\mu_y(x) \leftarrow 0$
$\Delta \leftarrow 0$
while $\Delta \leq \varepsilon$:
  for $y \in C(x)$:
    $n(y) \leftarrow n(y) + 1$
    $x_0 \leftarrow xy$
    $r \leftarrow 0$
    for $t \in \mathbb{N}_{[0, H-1]}$:
      $y_t \leftarrow \arg\max_{v \in V} \ell(v \mid x_t)$
      $r \leftarrow r + \gamma^t \ell(y_t \mid x_t)$
      $x_{t+1} \leftarrow x_t y_t$
    $\mu_y(x) \leftarrow \mu_y(x) + \frac{r - \mu_y(x)}{n(y)}$
  for $y \in C(x)$:
    $\beta(x, y) \leftarrow \frac{S}{1-\gamma}\sqrt{\frac{2\ln(4|C(x)|/\delta)}{n(y)}}$
    $L(x, y) \leftarrow \mu_y(x) - \beta(x, y) - e_2$
    $U(x, y) \leftarrow U_\infty(x, y)$
  $\hat{y} \leftarrow \arg\max_{y \in C(x)} L(x, y)$
  $\Delta \leftarrow L(x, \hat{y}) - \max_{z \in C(x), z \neq \hat{y}} U(x, z)$
return $\hat{y}$

4 Experimentation

5 Conclusion

References

John Doe and Jane Roe. 2025. An example paper. *Journal of Examples*.

A Appendix Title

Appendix content goes here (Doe and Roe, 2025).