

1 A Optimality of SV's Verification Length Selection

2 **Definition A.1.** Let S be a random variable representing natural divergence of P_d and P_c :

$$S = \sum_{j \in \text{vocab}} \min(P_d(t_j), P_c(t_j)) \quad (1)$$

3 where P_d and P_c are token probability distributions of the draft and companion model, respectively.

4 **Definition A.2.** Let A be a random variable for t_d 's acceptance probability in the companion model:

$$A = \min\left(1, \frac{P_c(t_d)}{P_d(t_d)}\right) \quad (2)$$

5 where t_d is the token generated by the draft model, and P_d and P_c are defined as in A.1.

6 **Definition A.3.** Let T_i be a random variable following the probability distribution of i 'th draft token
7 and $P(T_i)$ is its acceptance probability in the target model:

- 8 • $P(T_i = t)$ is the probability that a generated token t is accepted by the target model and
- 9 • $P(T_i = t|S, A)$ is its conditional probability when S and A are given

10 **Definition A.4.** Let N be a random variable for the number of tokens accepted by the target model.

11 The probability for N given γ is calculated as:

$$P_\gamma(N) = \begin{cases} P(T_{N+1} \neq t_{N+1}) \prod_{i=1}^N P(T_i = t_i) & \text{if } N < \gamma, \\ \prod_{i=1}^\gamma P(T_i = t_i) & \text{if } N = \gamma \end{cases} \quad (3)$$

12 The expected number of accepted tokens when verifying γ tokens is calculated as:

$$E(N|\gamma) = \sum_{i=1}^{\gamma} i \cdot P_\gamma(N = i) \quad (4)$$

13 where $P_\gamma(N = i)$ is the probability of accepting exactly i tokens when verifying γ tokens.

14 **Definition A.5.** Let $\text{goodput}(\gamma)$ be the goodput (i.e., the number of accepted tokens per unit time)
15 when verifying γ number of tokens, which is calculated as:

$$\text{Goodput}(\gamma) = \frac{N}{\text{Latency}(\gamma)} \quad (5)$$

16 where N is the number of tokens accepted when verifying γ tokens.

17 To estimate goodput, we use the profiled latency and expected number of accepted tokens as following:

$$\hat{\text{Goodput}}(\gamma) = \frac{E(N|\gamma)}{\text{Latency}_{\text{profiled}}(\gamma)} \quad (6)$$

18 **Assumption A.6.** The draft and companion models are reasonably aligned with the target model.

19 **Definition A.7.** We define $\hat{P}(T_i)$, i.e., an estimator of $P(T_i)$, to be $P(T_i|S, A)$

20 **Definition A.8.** We obtain $P(T_i|S, A)$ by observing sample data and grouping (i.e., binning) accep-
21 tance probability according to the values of S and A as following:

$$P(T_i|S, A) = \begin{cases} E(P(T_i|s_0 \leq S < s_1, a_0 \leq A < a_1)) \\ E(P(T_i|s_1 \leq S < s_2, a_1 \leq A < a_2)) \\ \dots \end{cases} \quad (7)$$

22 We can reduce the variance of the observed acceptance probabilities by controlling the bin sizes of S
23 and A . Reducing this variance subsequently decreases the uncertainty associated with the acceptance
24 probabilities.

25 **Assumption A.9.** The observed probability in Definition A.8 constitutes an unbiased estimator of
26 $P(T_i)$, as long as the observed distributions of S , A , and P coincide with those encountered at
27 inference time.

28 **Assumption A.10.** Let $Latency(\gamma)$ be the end-to-end wall-clock latency required to process γ
 29 tokens. and its finite difference is

$$\Delta Latency(\gamma) = Latency(\gamma + 1) - Latency(\gamma), \quad \gamma \geq 1.$$

30 We assume the following properties.

31 1. There exists a threshold γ_0 such that $\Delta Latency(\gamma)$ is a constant value for all $\gamma \geq \gamma_0$:

$$\Delta Latency(\gamma) = \Delta Latency(\gamma_0) \quad \text{for all } \gamma \geq \gamma_0.$$

32 2. For γ smaller than the threshold, $\Delta Latency(\gamma)$ is an increasing function:

$$\Delta Latency(\gamma + 1) > \Delta Latency(\gamma) \quad \text{for all } 1 \leq \gamma < \gamma_0.$$

33 From the above two properties, $\Delta Latency(\gamma)$ is a non-decreasing function for all $\gamma > 0$:

$$\Delta Latency(\gamma + 1) \geq \Delta Latency(\gamma) \quad \text{for all } \gamma > 0.$$

34 **Lemma A.11.** The expected number of accepted tokens $E(N|\gamma)$ is increasing and concave with
 35 respect to γ :

$$\frac{d}{d\gamma} E(N|\gamma) \geq 0 \tag{8}$$

$$\frac{d^2}{d\gamma^2} E(N|\gamma) \leq 0 \tag{9}$$

36 *Proof.* The first and second derivatives are computed as discrete differences, represented as Δ :

$$\frac{d}{d\gamma} E(N|\gamma) = \Delta^1 E(N|\gamma) = E(N|\gamma + 1) - E(N|\gamma) \tag{10}$$

$$\frac{d^2}{d\gamma^2} E(N|\gamma) = \Delta^2 E(N|\gamma) \tag{11}$$

$$= \Delta^1 E(N|\gamma + 1) - \Delta^1 E(N|\gamma) \tag{12}$$

$$= E(N|\gamma + 2) - 2E(N|\gamma + 1) + E(N|\gamma) \tag{13}$$

37 **Part 1: Proof of $\Delta^1 E(N|\gamma) \geq 0$ (monotonicity).**

38 Increasing the verification length γ can only increase the number of accepted tokens, and thus
 39 $\Delta^1 E(N|\gamma) \geq 0$.

40 **Part 2: Proof of $\Delta^2 E(N|\gamma) \leq 0$ (concavity).**

$$\Delta^1 E(N|\gamma) = E(N|\gamma + 1) - E(N|\gamma) \tag{14}$$

$$= \sum_{k=1}^{\gamma+1} k \cdot P_{\gamma+1}(k) - \sum_{k=1}^{\gamma} k \cdot P_{\gamma}(k) \tag{15}$$

$$= \underbrace{\sum_{k=1}^{\gamma-1} [k \cdot (P_{\gamma+1}(k) - P_{\gamma}(k))]}_{\alpha} + \underbrace{\gamma \cdot P_{\gamma+1}(\gamma) + (\gamma + 1) \cdot P_{\gamma+1}(\gamma + 1) - \gamma \cdot P_{\gamma}(\gamma)}_{\beta} \tag{16}$$

$$\Delta^2 E(N|\gamma) = \Delta^1 E(\gamma + 1) - \Delta^1 E(N|\gamma) \tag{17}$$

$$= E(N|\gamma + 2) - 2 \cdot E(N|\gamma + 1) + E(N|\gamma) \tag{18}$$

$$\begin{aligned} &= \sum_{k=1}^{\gamma-1} k \cdot \underbrace{[P_{\gamma+2}(k) - 2 \cdot P_{\gamma+1}(k) + P_{\gamma}(k)]}_{(a), \text{ derived from } \alpha} \\ &\quad + \underbrace{\gamma \cdot [P_{\gamma+2}(\gamma) - 2 \cdot P_{\gamma+1}(\gamma) + P_{\gamma}(\gamma)]}_{(b_1), \text{ derived from } \beta} \\ &\quad + \underbrace{(\gamma + 1) \cdot P_{\gamma+2}(\gamma + 1) - 2(\gamma + 1) \cdot P_{\gamma+1}(\gamma + 1) + (\gamma + 2) \cdot P_{\gamma+2}(\gamma + 2)}_{(b_2), \text{ derived from } \beta} \end{aligned} \tag{19}$$

41 $\Delta^2 E(N|\gamma)$ is smaller than zero because the term (a) is zero and $(b_1) + (b_2)$ is smaller than or equal
 42 to zero. We provide the proof below, but we briefly discuss why (a) is zero and $(b_1) + (b_2)$ is smaller
 43 than or equal to zero.

44 • **Term (a) = 0.** The probability of accepting k tokens remains unchanged when the
 45 verification length is increased, provided it is already greater than k .

46 • **Terms $(b_1) + (b_2) \leq 0$.** These terms are computed from the increases in expected accepted
 47 token length, excluding the unaffected parts. Because the increase from $\gamma + 1$ to $\gamma + 2$ (i.e.,
 48 $\Delta E(\gamma + 1)$) is smaller than the increase from γ to $\gamma + 1$ (i.e., $\Delta E(\gamma)$), these terms are
 49 negative.

50 *Proof of term (a) = 0:* Increasing γ (or $\gamma + 1$) does not affect the probability of accepting k tokens if
 51 k is already smaller than γ . Hence, $P_{\gamma+1}(k) - P_\gamma(k)$ (or $P_{\gamma+2}(k) - P_{\gamma+1}(k)$) equals zero.

52 For all k that satisfies $1 \leq k < \gamma$,

$$P_{\gamma+2}(k) - 2P_{\gamma+1}(k) + P_\gamma(k) = [P_{\gamma+2}(k) - P_{\gamma+1}(k)] - [P_{\gamma+1}(k) - P_\gamma(k)] \quad (20)$$

$$= 0 - 0 = 0 \quad (21)$$

53 *Proof of terms (b) + (c) ≤ 0 :*

$$(b) + (c) = \gamma \cdot [P_{\gamma+2}(\gamma) - 2 \cdot P_{\gamma+1}(\gamma) + P_\gamma(\gamma)]$$

$$+ (\gamma + 1) \cdot P_{\gamma+2}(\gamma + 1) - 2(\gamma + 1) \cdot P_{\gamma+1}(\gamma + 1)$$

$$+ (\gamma + 2) \cdot P_{\gamma+2}(\gamma + 2) \quad (22)$$

54 We can express all other probabilities in terms of $P_\gamma(\gamma)$ by applying A.4,

$$P_{\gamma+1}(\gamma) = P_\gamma(\gamma) \cdot P(T_{\gamma+1} \neq t_{\gamma+1}) \quad (23)$$

$$P_{\gamma+1}(\gamma + 1) = P_\gamma(\gamma) \cdot P(T_{\gamma+1} = t_{\gamma+1}) \quad (24)$$

$$P_{\gamma+2}(\gamma) = P_\gamma(\gamma) \cdot P(T_{\gamma+1} \neq t_{\gamma+1}) \quad (25)$$

$$P_{\gamma+2}(\gamma + 1) = P_\gamma(\gamma) \cdot P(T_{\gamma+1} = t_{\gamma+1}) \cdot P(T_{\gamma+2} \neq t_{\gamma+2}) \quad (26)$$

$$P_{\gamma+2}(\gamma + 2) = P_\gamma(\gamma) \cdot P(T_{\gamma+1} = t_{\gamma+1}) \cdot P(T_{\gamma+2} = t_{\gamma+2}) \quad (27)$$

55 then,

$$(b_0) + (b_1) = -P_\gamma(\gamma) \cdot [\gamma P(T_{\gamma+1} \neq t_{\gamma+1}) + P(T_{\gamma+1} = t_{\gamma+1}) \cdot P(T_{\gamma+2} \neq t_{\gamma+2})] \leq 0 \quad (28)$$

56 The last inequality comes from the fact that probability values are larger than or equal to zero.

57 Since both terms (a) = 0 and $(b_0) + (b_1) \leq 0$, we have:

$$\Delta^2 E(N|\gamma) \leq 0 \quad (29)$$

58 \square

59 **Lemma A.12.** The Goodput function is concave with respect to γ :

$$\Delta^2 \text{Goodput}(\gamma) \leq 0 \text{ for all } \gamma \quad (30)$$

$$(31)$$

Proof.

$$\Delta \text{goodput}(\gamma) = \text{goodput}(\gamma + 1) - \text{goodput}(\gamma) \quad (32)$$

$$= \frac{E(N|\gamma + 1)}{L(\gamma + 1)} - \frac{E(N|\gamma)}{L(\gamma)} \quad (33)$$

$$= \frac{L(\gamma) E(N|\gamma + 1) - E(N|\gamma) L(\gamma + 1)}{L(\gamma + 1) L(\gamma)} \quad (34)$$

$$= \frac{L(\gamma) [E(N|\gamma) + \Delta E(N|\gamma)] - E(N|\gamma) [L(\gamma) + \Delta L(\gamma)]}{L(\gamma + 1) L(\gamma)} \quad (35)$$

$$= \frac{L(\gamma) \Delta E(N|\gamma) - E(N|\gamma) \Delta L(\gamma)}{L(\gamma + 1) L(\gamma)} \quad (36)$$

Let $\Psi(\gamma) = L(\gamma) \Delta E(N|\gamma) - E(N|\gamma) \Delta L(\gamma)$, then (37)

$$\Delta \Psi(\gamma) = \Psi(\gamma + 1) - \Psi(\gamma) \quad (38)$$

$$= L(\gamma + 1) \Delta E(N|\gamma + 1) - E(N|\gamma + 1) \Delta L(\gamma + 1) \\ - L(\gamma) \Delta E(N|\gamma) + E(N|\gamma) \Delta L(\gamma) \quad (39)$$

$$= \underbrace{L(\gamma) \Delta^2 E(N|\gamma)}_{(a) \leq 0} + \underbrace{(-E(N|\gamma) \Delta^2 L(\gamma))}_{(b) \leq 0} \\ + \underbrace{\Delta L(\gamma) \Delta E(N|\gamma + 1) - \Delta L(\gamma + 1) \Delta E(N|\gamma)}_{(c) \leq 0} \quad (40)$$

60 (a) ≤ 0 : $\Delta^2 E(N|\gamma) \leq 0$ (from A.11)

61 (b) ≤ 0 : $E(N|\gamma) > 0$, $\Delta^2 L(\gamma) \geq 0$ (from A.10)

62 (c) ≤ 0 : where $\Delta L(\gamma) \leq \Delta L(\gamma + 1)$ and $\Delta E(N|\gamma + 1) < \Delta E(N|\gamma)$ (from A.10, A.11)

63 Because $\Delta \Psi(\gamma) < 0$, $\Psi(\gamma)$, i.e., $\Delta \text{goodput}(\gamma)$, is a decreasing function. That is, $\text{Goodput}(\gamma)$ is a
64 concave function.

65 Optionally, with conditions below,

there exist $\gamma_0 < \gamma_1$ such that (41)

$$\Delta \text{Goodput}(\gamma_0) > 0 \quad (42)$$

$$\Delta \text{Goodput}(\gamma_1) < 0 \quad (43)$$

66 we can find optimal value γ^* such that $\gamma_0 < \gamma^* < \gamma_1$ and $\text{Goodput}(\gamma^*)$ is the maximum goodput.

67 Our evaluations indicate that these conditions generally hold true. First, if we define $\text{Goodput}(0) = 0$
68 then $\Delta \text{Goodput}(0)$ is positive. Second, for draft lengths greater than or equal to 5, we consistently
69 observed a value of γ_1 such that $\Delta \text{Goodput}(\gamma_1) < 0$. \square