6. Appendix

6.1. Proofs of Theorems 3.3 and 3.4

Theorem 3.3. *For an approximate MIPS method with additive error $c > 0$, let the inverse partition function estimate using Gumbels and exact MIPS be $\hat{Z}^{-1}$ and the estimate with approximate MIPS be $\widetilde{Z}^{-1}$. Then,*

$$\hat{Z}^{-1} \leq \widetilde{Z}^{-1} \leq e^{c}\hat{Z}^{-1}$$

Proof. Let H be defined as before and let $\hat{H}$ be the corresponding value with the approximate method. Since the approximate method has additive error c,

$$H - c \leq \hat{H} \leq H$$

$$e^{c}e^{-H} \geq e^{-\hat{H}} \geq e^{-H}$$

Using the notation from the theorem statement, $\hat{Z}^{-1} = e^{-H}$ and $\widetilde{Z}^{-1} = e^{-\hat{H}}$. Thus,

$$e^{c}\hat{Z}^{-1} \geq \widetilde{Z}^{-1} \geq \hat{Z}^{-1}$$

□

Theorem 3.4. *For an approximate MIPS method with additive error $c > 0$, let the sample with the approximate MIPS be $\widetilde{i}$. Then,*

$$e^{-c}P_\theta(x_k) \leq P(\widetilde{i} = k) \leq e^{c}P_\theta(x_k)$$

Proof. Let $\widetilde{i}$ be the maximum from the approximate MIPS. For convenience, let $a_i = \theta \cdot \phi(x_i)$. Let G_i be the corresponding Gumbel variables. Note that,

$$P_\theta(x_k) = P(k = \operatorname*{argmax}_j a_j + G_j)$$
$$P_\theta(x_k) = P(a_k + G_k > \max_{j \neq k} a_j + G_j)$$

Let us examine the probability of sampling a particular k with the approximate MIPS method.

$$P(\widetilde{i} = k) \leq P(\max_{j \neq k} a_j + G_j - c < a_k + G_k)$$
$$P(\widetilde{i} = k) \geq P(\max_{j \neq k} a_j + G_j < a_k + G_k - c)$$

The maximum has a distribution of $\log(Z - e^{a_k}) + G'$ where G' has a Gumbel distribution.

$$P(\widetilde{i} = k) \leq P(\log(Z - e^{a_k}) + G' - c < a_k + G_k)$$

$$P(\widetilde{i} = k) \geq P(\log(Z - e^{a_k}) + G' < a_k + G_k - c)$$

Rearranging terms,

$$P(\widetilde{i} = k) \leq P(G' - G_k < a_k - \log(Z - e^{a_k}) + c)$$

$$P(\widetilde{i} = k) \geq P(G' - G_k < a_k - \log(Z - e^{a_k}) - c)$$

Since the difference of two Gumbel distributions is a logistic distribution,

$$P(\widetilde{i} = k) \leq \frac{1}{1 + e^{-(a_k - \log(Z - e^{a_k}) + c)}}$$

$$P(\widetilde{i} = k) \geq \frac{1}{1 + e^{-(a_k - \log(Z - e^{a_k}) - c)}}$$

And thus,

$$e^{-c}\frac{e^{a_k}}{Z} \leq P(\widetilde{i} = k) \leq e^{c}\frac{e^{a_k}}{Z}$$

$$e^{-c}P_\theta(x_k) \leq P(\widetilde{i} = k) \leq e^{c}P_\theta(x_k)$$

□

6.2. Tables

Here we present some tables from the experimental results section. The tables with the probabilities without model averaging for the states that were presented in Section 4.2 are shown as Table 3 and 4. The probabilities of the words generally make sense given that the model is a bag-of-words-like model. The "point_b" may appear odd at first, but is often used in conjunction with flights as in the context of the phrase "from point A to point B".

The words used to train the basketball category in Section 4.3 can be seen in Table 5. The top 50 words for the generative models learned by the four different methods mentioned in Section 4.3 can be seen in Table 6 and 7. The training words appear in black black, the other basketball-related words appear in blue, and the non-basketball-related words appear in red. The non-basketball-related words are counted as mistakes.

For the HSKM method, we used $p = 100$ as a beam size and $b = 10$ as a branching factor. We also used $k = 1$ and $t = 100$.

Order	Probability
1	0.5122
2	0.05799
3	0.02643
10	0.01056
100	6.632e-4
1000	1.6336e-5

Table 3. Probabilities (without using model averaging) used for synthetic model averaging inference

Order	Word	Probability
1	point_b	2.8917e-5
2	stresses	2.0855e-5
3	turbulence	2.0143e-5
4	ordeal	1.8597e-5
5	stress	1.7599e-5
10	important_thing	1.5307e-5
100	formal_training	1.0142e-5
1000	gabby's_seat	6.7475e-6
10000	garamba	4.2853e-6
100000	joe_hart_vincent_kompany	2.3600e-6

Table 4. Words and probabilities (without using model averaging) for word2vec model averaging inference

Basketball-related Words
hoop
shoot
basket
dribble
pass
three_pointer
key
sideline
bench
court
coach
player
center
guard

Table 5. The basketball-related words used for training the basketball category for the gradient descent experiment

Rank	Exact Method	MRG: Exact MIPS
1	bench	sideline
2	sideline	dribble
3	dribble	bench
4	guard	hoop
5	hoop	ball
6	three_pointer	three_pointer
7	ball	guard
8	basket	basket
9	layup	loose_ball
10	defender	layup
11	loose_ball	timeout
12	teammate	defender
13	shot_clock	shot_clock
14	jump_shot	pointer
15	pointer	jump_shot
16	coach	an_ncaa_college
17	timeout	teammate
18	dribbling	coach
19	with_seconds_left	free_throw
20	player	dribbling
21	free_throw	scrimmage
22	scrimmage	player
23	an_ncaa_college	shoot
24	shoot	with_seconds_left
25	teammates	buzzer
26	receiver	teammates
27	shooting_guard	receiver
28	pass	pass
29	fast_break	dribbled
30	point_guard	shooting_guard
31	buzzer	fast_break
32	yard_line	yard_line
33	dribbled	midcourt
34	defenders	point_guard
35	midfield	referee
36	free_throws	an_nba_basketball
37	an_nba_basketball	defenders
38	assistant_coach	quickness
39	referee	free_throws
40	midcourt	perimeter
41	locker_room	griner
42	perimeter	midfield
43	quickness	locker_room
44	griner	inbounds_pass
45	scoring	assistant_coach
46	playmaker	jump_shots
47	court	basketball
48	playmaking	court
49	jump_shots	inbounded
50	inbounds_pass	forward

Table 6. Top 50 words according to models learned by gradient descent with exact method and with our Gumbel reduction using exact MIPS. The training words are black, basketball-related words are blue, and non-basketball-related words are red.

Rank	MRG: HSKM MIPS	Mean Heuristic
1	layup	layup
2	three_pointer	pointer
3	pointer	three_pointer
4	ball	ball
5	loose_ball	loose_ball
6	free_throws	free_throws
7	free_throw	end_zone
8	jump_shot	with_seconds_left
9	two_free_throws	yard_touchdown
10	fast_break	free_throw
11	with_seconds_left	two_free_throws
12	shot_clock	an_ncaa_college
13	timeout	puck
14	puck	jump_shot
15	dribble	dunk
16	dunk	timeout
17	basket	yard_line
18	dunks	field_goal
19	layups	yard_field_goal
20	end_zone	fast_break
21	penalty_area	an_alley_oop
22	an_alley_oop	bench
23	three_pointers	sideline
24	yard_touchdown	yard_touchdown_pass
25	jumper	three_pointers
26	driving_layup	basket
27	midfield	th_minute
28	bench	penalty_area
29	field_goal	dunks
30	backboard	midfield
31	reverse_layup	scrimmage
32	yard_line	dribble
33	technical_foul	shot_clock
34	sideline	jumper
35	free_throw_line	touchdown
36	dribbled	technical_foul
37	buzzer	yard_gain
38	scrimmage	yard_pass
39	an_ncaa_college	layups
40	midcourt	driving_layup
41	foul	final_seconds
42	ump_shots	with_seconds_remaining
43	an_offensive_rebound	reverse_layup
44	th_minute	teammate
45	with_seconds_remaining	free_throw_line
46	short_jumper	dribbled
47	point_guard	an_nba_basketball
48	fouls	midcourt
49	foul_line	yard_run
50	teammates	teammates

Table 7. Top 50 words according to models learned by gradient descent with our Gumbel reduction using HKSM and with the mean heuristic. The training words are black, basketball-related words are blue, and non-basketball-related words are red.