Permute-and-Flip: An optimally stable and watermarkable decoder for LLMs

Xuandong Zhao[1], Lei Li[2], Yu-Xiang Wang[3],
[1]UC Berkeley [2]Carnegie Mellon University [3]UC San Diego

Existing watermarks work with the standard decoder:
softmax(logits)

$$\textbf{Softmax sampling: } y_t \sim p(y) = \frac{e^{u(y|x,y_{1:t-1})/T}}{\sum_{\tilde{y}} e^{u(\tilde{y}|x,y_{1:t-1})/T}}$$

Is softmax(logits) the optimal choice?

Can we benefit from co-designing the decoder and watermarking scheme?

PF Decoding

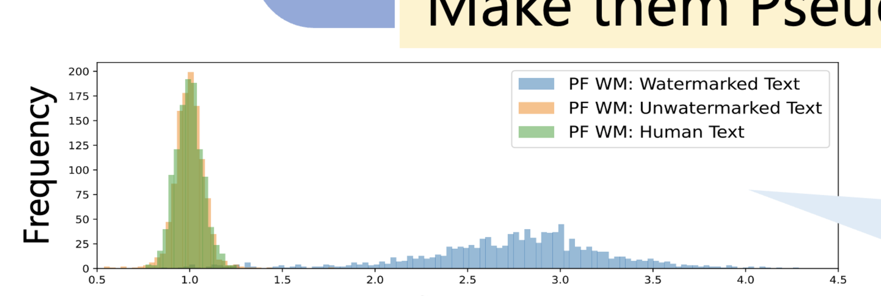
Permute-and-Flip Decoding

Sampling	PF Decoding
$y_t = \arg\max_{y\in\mathcal{V}} \frac{u_t(y)}{T} + G_t(y)$	$y_t = \arg\max_{y\in\mathcal{V}} \frac{u_t(y)}{T} + E_t(y)$
$G_t(y) \sim Gumbel(0,1)\ i.i.d$	$E_t(y) \sim Exponential(1)\ i.i.d$

PF Watermark

Gumbel Watermark [Aaronson, 2023]	**PF Watermark** [Ours]
$y_t = \arg\max_{y\in\mathcal{V}} \frac{u_t(y)}{T} + G_t(y)$	$y_t = \arg\max_{y\in\mathcal{V}} \frac{u_t(y)}{T} + E_t(y)$
$\boxed{G_t(y) \sim Gumbel(0,1)\ i.i.d}$	$\boxed{E_t(y) \sim Exponential(1)\ i.i.d}$

Make them Pseudo-random!



A Clear Gap
↓
Watermark Detectability

Stability

Definition 2.1 (Stability). We say a decoding algorithm $\mathcal{A}$ is L-stable if for any prompt x, prefix $y_{\le t}$, and for any perturbed $\tilde{u}$ such that $\|\tilde{u} - u\|_\infty \le \delta$, the log-probability ratio satisfies

$$\left|\log\left\{\frac{p_{\mathcal{A}(\tilde{u}(\cdot|x,y_{\le t}))}(y)}{p_{\mathcal{A}(u(\cdot|x,y_{\le t}))}(y)}\right\}\right| \le L\delta \quad \forall y \in \mathcal{V}.$$

- **Avoid catastrophic failure**
- **Implies diversity**

Guarantees

Theorem 3.1. *Let the logits function be u and $u^* = \max_{y\in\mathcal{V}} u(y)$. Let $\mathrm{PF}(u)$ be the distribution of PF-sampling, and $\mathrm{Softmax}(u)$ be the distribution in equation 1, both with temperature parameter T. The following statements are true.*

1. *(**Same stability**) PF-Sampling is $(2/T)$-stable.*
2. *(**Nearly greedy**) PF-sampling obeys,*
$$\mathbb{E}_{y\sim\mathrm{PF}(u)}[u(y)] \ge u^* - T\log|\mathcal{V}|.$$
3. *(**"Never worse"**) For the same T, PF-sampling is never worse than Softmax-sampling.*
$$\mathbb{E}_{y\sim\mathrm{PF}(u)}[u(y)] \ge \mathbb{E}_{y\sim\mathrm{Softmax}(u)}[u(y)]$$
4. *(**"Up to 2x better"**) There exists logits u such that PF-sampling is 2x smaller in terms of suboptimality,*
$$\mathbb{E}_{y\sim\mathrm{PF}(u)}[u^* - u(y)] \le \frac{1}{2}\mathbb{E}_{y\sim\mathrm{Softmax}(u)}[u^* - u(y)].$$
5. *(**Optimal stability-perplexity tradeoff**) For any decoder P that is $2/T$-stable, if there exists u such that*
$$\mathbb{E}_{y\sim P(u)}[u(y)] > \mathbb{E}_{y\sim\mathrm{PF}(u)}[u(y)]$$
then there must be another $\tilde{u}$ such that
$$\mathbb{E}_{y\sim P(\tilde{u})}[\tilde{u}(y)] < \mathbb{E}_{y\sim\mathrm{PF}(\tilde{u})}[\tilde{u}(y)].$$

$$\mathrm{TestScore}_{\mathrm{Gumbel}}(y_{1:n}) = \sum_{t=m+1}^{n} -\log(1 - r_t(y_t))$$

No Watermark

$$\mathbb{E}[\mathrm{TestScore}(y_{1:n})] = n - m$$

Watermarked

$$\mathbb{E}[\mathrm{TestScore}(y_{1:n})] = \sum_{t=m+1}^{n} \mathbb{E}\left[\sum_{y\in\mathcal{V}} p_t(y) H_{\frac{1}{p_t(y)}}\right]$$
$$\ge (n-m) + \left(\frac{\pi^2}{6} - 1\right)\sum_{t=m+1}^{n} \mathbb{E}\left[\mathrm{Entropy}[p_t(\cdot)]\right]$$

$$\mathrm{TestScore}_{\mathrm{PF}}(y_{1:n}) = \sum_{t=m+1}^{n} -\log(r_t(y_t))$$

No Watermark

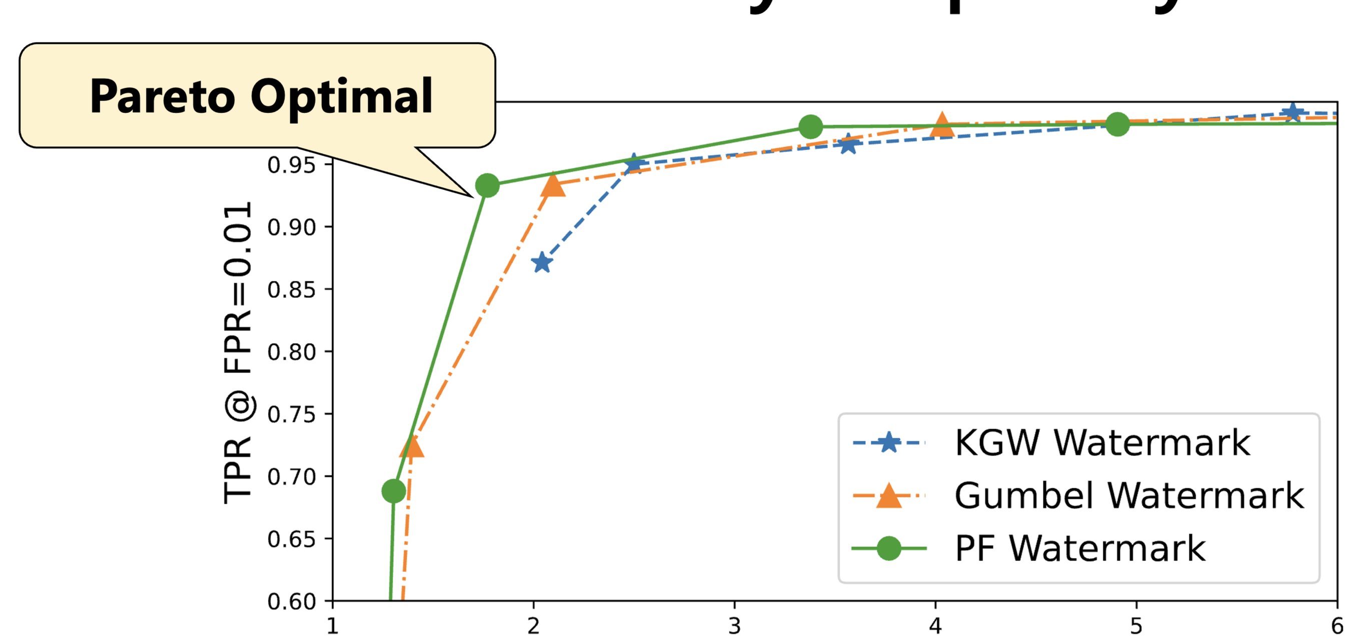
$$\mathbb{E}\left[\mathrm{TestScore}_{\mathrm{PF}}(y_{1:n}) | y_{1:n}\right] = n - m$$

Watermarked

$$\mathbb{E}\left[\mathrm{TestScore}_{\mathrm{PF}}(y_{1:n})\right]$$
$$= \sum_{t=m+1}^{n} \mathbb{E}\left[\sum_{y\in\mathcal{V}} \int_0^{e^{u_t(y)-u_t^*}} \left(-\log r \cdot \prod_{y'\in\mathcal{V}, y'\neq y} \left(1 - r\cdot e^{u_t(y')-u_t(y)}\right)\right) dr\right]$$

Results

Better Detectability-Perplexity Tradeoff



Method	AUC↑	TPR↑	PPL1↓	PPL2↓	Seq-rep-5↓	MAUVE↑
C4, T=1.0, Llama2-7B						
Greedy	-	-	$1.14_{0.01}$	$1.24_{0.03}$	0.56	0.05
Sampling	-	-	$12.47_{0.32}$	$15.31_{0.41}$	0.02	0.98
PF	-	-	$8.94_{0.20}$	$10.75_{0.25}$	0.03	0.90
KGW WM	0.989	0.991	$16.62_{0.38}$	$20.62_{0.49}$	0.01	1.00
Gumbel WM	0.997	0.988	$11.41_{0.27}$	$14.12_{0.36}$	0.04	0.93
PF WM	0.995	0.984	$8.33_{0.20}$	$10.28_{0.29}$	0.05	0.99