Sentiment Controlled LLM Decoding via Hamiltonian Monte Carlo

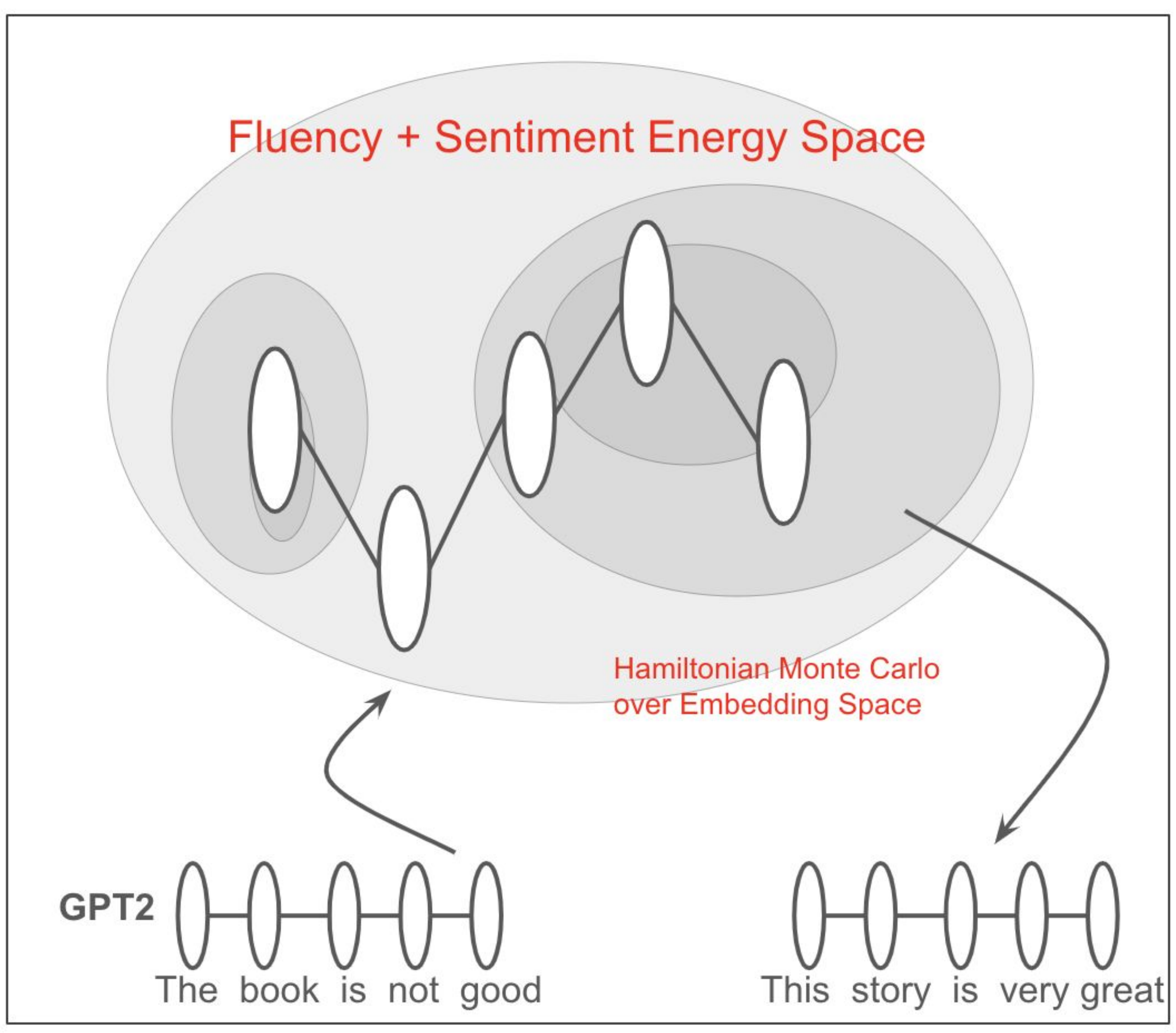
Hojin Sohn
sohn31@purdue.edu

PURDUE UNIVERSITY

Goal: HMC sampling directly within the continuous space of token embeddings, achieve more contextually coherent and controlled sentiment text generation compared to other decoding methods.

Motivation

- Coherent and sentiment-controlled text generation from minimal prompts is a key capability driving the next frontier in NLP.
- Recent studies on Langevin sampling over logit space / embedding space and discrete auto-regressive bias sampling
- Leveraging HMC sampling for broader exploration, avoiding local minima and ensuring diversity.



Approach

Idea: HMC over embedding space with an energy function

- This approach explores the continuous token embedding space, enabling more coherent and sentiment-controlled text generation. The process is initialized with text generated by a standard language model.
- The energy function is designed to incorporate both fluency and sentiment objectives, encouraging outputs that are both grammatically fluent and sentiment-aligned.
- Sequences are updated via Hamiltonian dynamics, which leverage momentum and gradient-based exploration to enable broader, more structured sampling behavior.
- The accept/reject step in HMC ensures samples follow the desired posterior distribution.

Solution depiction

Algorithm 1 Hamilton Monte Carlo sampling steps

Require: Input sequence s, Output length L, base LM, Sentiment score function f and threshold ϵ, step size α, Leapfrog step size δ, Leapfrog steps L, momentum std dev σ

Ensure: Accepted samples

Initialize $\boldsymbol{Y}_0$ from LM embeddings prompted with s
for $t = 0$ to $t = 500$ **do**
  $\boldsymbol{\Phi}_0^{(i)} \sim \mathcal{N}(0, \sigma)$ for each token i
  $\boldsymbol{X}_0 = \boldsymbol{Y}_t$
  Compute energy $E(\boldsymbol{X}_0)$
  for $l = 0$ to $L - 1$ **do**
    $\boldsymbol{\Phi}_{(l+\frac{1}{2})\delta} = \boldsymbol{\Phi}_{l\delta} - \frac{\delta}{2} \frac{\partial E}{\partial \boldsymbol{Y}}\Big|_{\boldsymbol{Y}=\boldsymbol{X}_{l\delta}}$
    $\boldsymbol{X}_{(l+1)\delta} = \boldsymbol{X}_{l\delta} + \delta \boldsymbol{R}^{-1} \boldsymbol{\Phi}_{(l+\frac{1}{2})\delta}$
    $\boldsymbol{\Phi}_{(l+1)\delta} = \boldsymbol{\Phi}_{(l+\frac{1}{2})\delta} - \frac{\delta}{2} \frac{\partial E}{\partial \boldsymbol{Y}}\Big|_{\boldsymbol{Y}=\boldsymbol{X}_{(l+1)\delta}}$
  end for
  $\alpha = \min\left(1, \exp\left(-H(\boldsymbol{X}_{L\delta}, \boldsymbol{\Phi}_{L\delta}) + H(\boldsymbol{X}_0, \boldsymbol{\Phi}_0)\right)\right)$
  if Uniform$(0,1) \le \alpha$ **then**
    $\boldsymbol{Y}_{t+1} = \boldsymbol{X}_{L\delta}$; save to samples
  else
    $\boldsymbol{Y}_{t+1} = \boldsymbol{Y}_t$
  end if
  if t % 10 == 0 **then**
    $\lambda_i^t = \max\left(0, \lambda_i^{t-1} + \alpha \nabla_{\lambda_i} E(\boldsymbol{Y})\right)$
  end if
end for

$$E(Y) = -\log P_{LM}(\text{project}(Y)|x) - \lambda(\epsilon - f(Y))$$

$f(Y) =$ a sentiment classifier trained by adding a linear layer on top of GPT2LMHead's output representations.

Dataset: https://huggingface.co/datasets/stanfordnlp/sst2

Challenges

- Energy and Gradient Computation: Non-differentiable token mapping for energy (fluency + sentiment) complicated gradient computation w.r.t. Token Embeddings.
  - Used a Straight-Through Estimator to enable backpropagation through discrete token assignments
- Kinetic Energy in HMC: Improper mass matrix scaling caused low acceptance rates, leading to inefficient sampling.
  - Tuned the mass matrix to balance kinetic energy, improving acceptance
- Token Mapping: Mapping continuous embeddings to discrete tokens.
  - Applied distance-based mapping (nearest embeddings)
- Stuck in Local Minima / Unsatisfied Sentiment
  - Dynamically adjust sentiment weight based on the energy gradient every 10th iteration
  - Dynamically increase leapfrog step size if same samples are sampled

Experiments

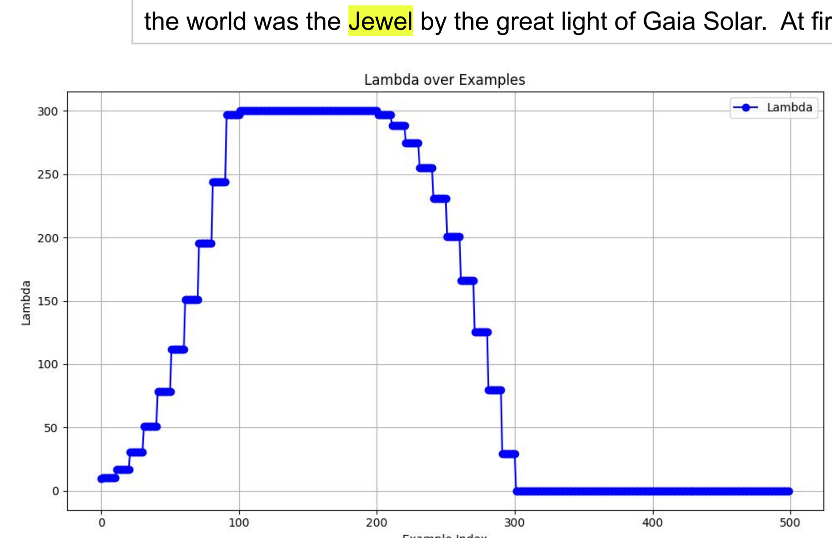
Prompt: "The book is "

	textattack/roberta-base-SST-2	DistillGPT2
Text	Sentiment Score	Perplexity
about the time he was told by his parents that it was "a bad idea" for him to start writing	0.003014187096	17.26315498
stars the time he was told by his parents that it was "a bad idea" for him to start inhab	0.002595550148	28.28341484
about the time he was told by his parents that it was "a bad idea" for him to start resusc	0.003512033727	22.61892128
about the time he was told by his parents that it was "a good idea" for him to start resusc	0.8878903985	21.5558815

Text	Sentiment Score	Perplexity
an adaptation of a book by John Steinbeck, and we're sure this is the first time the book has	0.09196235985	18.63439369
an appreciation of a book by John Steinbeckania and we are sure the is a first time way booknat	0.9962477088	537.3682251
an appreciation of a book by John Steinbeckania and we are sure the is a first time the booknat	0.9955881834	354.7834778
an appreciation of a book by John Steinbeckania and we are sure this is a first time the booknat	0.9942049384	232.3453979

Prompt: "Once upon a day, "

Text	Sentiment Score	Perplexity
the universe had been shaped by a great wave of random mutations. At first, the universe was simply a flat patch of time. But	0.4706687033	32.1572876
the universe had the influenced by a great sweep of intelligent luc, At first, the heavens was just a flat patch of time, but	0.9777960181	161.2251434
the Spirit was the blessing by the great sweep of powerful Solar's At first, the cosmos was just a flat patch of time, but	0.9977024198	205.9025726
the earth was the blessing by the great light of powerful Solar, At first, the flowering was just a flat patch of time, but	0.9943186641	134.0469971
the earth was the blessing by the great light of luc Solar. At first, the Celestial was just a lonely patch of time, but	0.9937207699	146.9455719
the earth was the Jewel by the great light of Gaia Solar. At first, the cosmos was just a small fragment of time, but	0.9952920675	77.05979156
the universe was the Jewel by the great light of Gaia Solar. At first, the cosmos was just a small fragment of time, but	0.9946040511	76.57580566
the world was the Jewel by the great light of Gaia Solar. At first, the cosmos was just a small fragment of time, but	0.9967356324	78.70362854



Lambda increases to satisfy sentiment / decrease after sentiment constraint gets satisfied (over threshold)