EXPOSING THE IMPLICIT ENERGY NETWORKS BEHIND MASKED LANGUAGE MODELS VIA METROPOLIS–HASTINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

While recent work has shown that scores from models trained by the ubiquitous masked language modeling (MLM) objective effectively discriminate probable and improbable sequences, it is still an open question if these MLMs specify a principled probability distribution over the space of possible sequences. In this paper, we interpret MLMs as energy-based sequence models and propose two energy parametrizations derivable from the trained MLMs. In order to draw samples correctly from these models, we develop a tractable *sampling* scheme based on the Metropolis–Hastings Monte Carlo algorithm. In our approach, samples are proposed from the same masked conditionals used for training the masked language models, and they are accepted or rejected based on their energy values according to the target distribution. We validate the effectiveness of the proposed parametrizations by exploring the quality of samples drawn from these energy-based models for both open-ended unconditional generation and a conditional generation task of machine translation. We theoretically and empirically justify our sampling algorithm by showing that the masked conditionals on their own do not yield a Markov chain whose stationary distribution is that of our target distribution, and our approach generates higher quality samples than other recently proposed undirected generation approaches (Wang and Cho, 2019; Ghazvininejad et al., 2019).

1 INTRODUCTION

Masked language modeling (MLM) objectives (Devlin et al., 2018; Yang et al., 2019; Gu et al., 2017) for sequences, although recent, have become ubiquitous for many Natural Language Processing (NLP) applications (Liu et al., 2019; Zhang et al., 2019; Rogers et al., 2021) because they are easy to optimize and enable learning of highly expressive and flexible representations by the virtue of conditioning on bidirectional context (Peters et al., 2018; Devlin et al., 2018). However despite their popularity, they lack a principled probabilistic interpretation and hence sampling from MLMs, or characterizing uncertainty about the predictions made with them has remained elusive. This drawback is reflected in the observation that recently proposed non-probabilistic approaches for *generating* high-scoring sequences from these MLMs (Ghazvininejad et al., 2019) still trail probabilistic autoregressive models (Brown et al., 2020; Sutskever et al., 2014) despite having access to greater bidirectional context while generating.

In this work, we posit that optimizing MLM objectives results in training of *implicit energy-based sequence models* that correspond to probability distributions over natural language sequences by assigning a score to each possible sequence in the large but finite sequence space. To explore the veracity of this claim, we *develop and experiment with two energy parametrizations* (or scoring schemes) that can be easily derived from the representations learned by the trained MLMs' transformer networks. These parametrizations have been inspired by the success of recent work on using MLMs for sentence-level judgements for discriminating between probable and improbable sequences (Salazar et al.; Zhang et al., 2019). Although, it is easy to compute the energy/score of a sequence with these MLM-based parametrizations, the bidirectional nature of MLMs precludes efficient sampling algorithms like ancestral sampling. Therefore, a *primary contribution* of our work is to *develop Metropolis-Hastings (MH) based sampling algorithms* for these energy networks. While it is tempting to formulate a Gibbs sampling scheme (Gelfand and Smith, 1990) based on the positional masked

conditional distributions used for training the MLMs (Wang and Cho, 2019), we theoretically and empirically demonstrate that these masked conditional distributions do not necessarily correspond to any joint distribution or energy network and hence result in *invalid* Gibbs samplers. Instead, we propose to use these masked conditionals as proposal distributions for transitioning to a new state (sequence) in the Markov chain of an MCMC sampler based on the Metropolis-Hastings algorithm (Hastings, 1970). *Another contribution* of our work is to *design a block-replacement proposal distribution* for improve mixing of the Markov chain in our proposed MH sampling framework, which results in faster generation and better samples.

We empirically investigate the effectiveness of the two proposed energy parametrizations by examining the quality of samples drawn from these energy-models in two diverse settings: 1) conditional generation task of Machine Translation (MT), and 2) Open-ended unconditional generation. We observe that high BLEU scores for MT, and high fluency scores are correlated with low energy values which indicates that these parametrizations are reasonable proxies for the desired implicit bidirectional energy network trained via the MLM objective. We study the behavior of our sampling approach extensively with different proposal distributions. We also verify the soundness of our approach by sampling from regions around the mode by annealing the target distribution and finding our *samples* to be competitive with a prominent undirected (and non-probablistic) generation approach (Ghazvininejad et al., 2019) on MT performance. Moreover, human evaluation of the open ended generation samples further corroborates the effectiveness of our approach.

We find our proposed sampler generates high-quality sequences under the proposed energy parametrizations which suggests that the optimization of MLM objective is implicitly equivalent to training global energy network that induces probability distribution over the space of sequences. While in this work we primarily focus on sampling from the energy network underlying MLMs, our findings promote the development of more direct, stable and simple training procedures for energy-based sequence models inspired from the MLM objectives and our proposed sampling scheme.

Related work: Gradient based training of energy networks (LeCun et al., 2006; Zhao et al., 2016; Du and Mordatch, 2019) has been successful at training models for inducing distributions over continuous domains but are not suitable for training discrete sequence models for text. To overcome this problem, recent work has proposed continuous relaxations to the discrete domain (Belanger and McCallum, 2016; Grathwohl et al., 2021), but the unordered nature of discrete symbols in text leads to brittle and unsuccessful training. Direct training of energy networks for text tends to be expensive and unstable as well (Goyal et al., 2019; Deng et al., 2020; Tu et al., 2020; Zhang et al., 2017). While MLM objectives(Devlin et al., 2018; Clark et al., 2020a;b) in contrast, are easy to train and learn good representations of textual data, they do not have a probabilistic interprertation. In this work, we interpret MLMs as implicit energy networks and develop approaches to sample from them. While there have been attempts to generate sequences from MLMs in a non(pseudo)-probabilistic manner (Wang and Cho, 2019; Ghazvininejad et al., 2019; Gu et al., 2017; Mansimov et al., 2019), our techniques sample correctly from the energy networks underlying MLMs.

2 Masked Language Models and Energy Networks

We can only directly obtain the conditional distributions of the `[MASK]` tokens, conditioned on the rest of the tokens in the sequence from an MLM. In this section, we discuss potential parametrizations of energy functions that could correspond to the implicit networks trained via MLM objectives and describe how to obtain these energy values from the trained MLMs. Let $\mathcal{X}$ be the space of all finite sequences up to a maximum length, and $p(X;\theta)$ be the probability of the sequence $X \in \mathcal{X}$ under the target distribution defined by the energy function $E(X;\theta)$ parametrized by θ, defined as follows:

$$p(X;\theta) = \frac{e^{-E(X;\theta)}}{\sum_{X' \in \mathcal{X}} e^{-E(X';\theta)}} = \frac{\phi(X;\theta)}{Z(\theta)}$$

where ϕ represents the unnormalized score of the sequence X and $Z(\theta)$ is the intractable normalization constant computed by summing over all the sequences. We propose *two* parametrizations for the energy functions: 1) *Raw* scoring, and 2) *Locally normalized* scoring.

2.1 Raw Scoring

For each position t in the sequence X of length T, we associate a random variable $X_t \in \mathbb{V}$ with the t-th token, where $\mathbb{V}$ is the vocabulary. MLMs learn a representation $h(X_{\setminus t})$, for X_t

that is sensitive to the bidirectional surroundinng context $X_{\setminus t}$. For notational convenience, we use $X_{i=w,\setminus i}$ to denote a sequence X with the i-th variable taking on the value w. We use such bidirectional neural parametrizations to define an energy E_{raw} for X that corresponds to fully connected MRFs (Gibbs random fields, more precisely) as the sum of the local positional scores: $\mathbf{E_{raw}}(\mathbf{X};\theta) = -\sum_{t=1}^{T} \log \phi_t(X;\theta)$, where $\log \phi_t(X;\theta) = f(X_t, h(X_{\setminus t}));\theta)$. In our experiments, the positional log-potentials $f(X_t, h(X_{\setminus t}));\theta)$ are computed by masking the position t, then running a forward pass on the MLM's transformer and using the raw logits at the masked position.

Conditional distribution under E_{raw}: Performing Gibbs sampling from the MRF defined by E_{raw} requires computation of this conditional distribution of a token given the surrounding context:

$$p(X_i|X_{\setminus i};\theta) = \frac{\prod_t \phi_t(X;\theta)}{\sum_{w\in\mathbb{V}} \prod_t \phi_t(((X_{i=w,\setminus i});\theta))}$$

Computing this conditional is very expensive and would require running $|\mathbb{V}|$ passes of the MLM decoder just for computing the positional potential (ϕ_t) at a single position because these potentials form fully connected cliques. Thus, we do not perform Gibbs sampling and instead propose MH based samplers described below.

Relationship with the masked conditionals of MLMs: Wang and Cho (2019)'s prior attempt to interpret a MLM (like BERT) as an MRF incorrectly[1] assumes that the positional potentials are independent of each other and hence are not defined on a fully connected clique, i.e. $\phi_t(X;\theta) = \phi_t(X_t;\theta)$. This faulty assumption about the factorization of the positional potentials $\phi_t(X;\theta)$ leads to the following formulation of conditional distribution:

$$p_{mlm}(X_i|X_{\setminus i};\theta) = \frac{\prod_t \phi_t(X_t;\theta)}{\sum\limits_{w\in\mathbb{V}} \prod_t \phi_t(((X_{i=w,\setminus i})_t;\theta))} = \frac{\phi_i(X_i;\theta)}{\sum\limits_{X'_i\in\mathbb{V}} \phi_i(X'_i;\theta)} \prod_{t\neq i} \frac{\phi_t(X_t;\theta)}{\phi_t(X_t;\theta)} = \text{softmax}(\log \phi_i)$$

This *deficient* conditional distribution for the MRF corresponds to the free conditional distribution $p_{mlm}(X_t \mid X_{\setminus t})$ that is obtained by performing a softmax operation over `[MASK]` scores ($\in \mathbb{R}^{\mathbb{V}}$) used in the MLM training objective. These MLM free conditionals do not correspond to the MRF defined by E_{raw} i.e. $p_{mlm}(X_i \mid X_{\setminus i}) \neq p(X_i|X_{\setminus i};\theta(E_{raw}))$. In fact, these conditionals need not correspond to *any* consistent MRF over the sequences. As an example, consider a sequence of length 2 with the random variables X_1, X_2 that have a categorical distribution over a vocabulary $\mathbb{V} = \{a, b\}$. The following free conditionals are inconsistent (see Appendix) because they do not correspond to any valid joint distribution over $\{X_1, X_2\}$: $p(X_1 \mid X_2) = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}, p(X_2 \mid X_1) = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$. It should be noted that prior work on dependency networks (Heckerman et al., 2000) proposed a similar scheme of training the conditionals independently with separate regressions over the latent variables and the inconsistency of such conditionals is well documented (Gelman and Raghunathan, 2001; Dobra et al., 2004; Lowd, 2012).

Wang and Cho (2019) used the masked conditionals to define a pseudolikelihood ($\prod_{t=1}^{T} p_{mlm}(X_t \mid X_{\setminus t};\theta)$) maximization objective and argued that MLM training can be interpreted as stochastic maximization of this pseudolikelihood corresponding to the energy function E_{raw}. However, this is incorrect because the conditionals used to define the pseudolikelihood under E_{raw} are deficient and likely inconsistent. Despite the incongruity between MLM training and minimization of E_{raw}, we propose to use E_{raw} as one of the parametrizations of the energy function.

2.2 Locally Normalized Scoring

Recent work (Zhang et al., 2019) has shown that MLMs like BERT can be used to reliably score a set of sequences. Salazar et al. and Clark et al. (2020a) developed a scoring scheme to rescore hypotheses proposed by the beam search algorithm and showed downstream improvements over automatic speech recognition (ASR) and machine translation (MT) datasets. The scoring scheme corresponded to masking tokens one-by-one in a left-to-right manner and summing the log-probability of the token at each masked position in the sequence: $\mathbf{E_{local}}(\mathbf{X};\theta) = -\sum_{t=1}^{T} \log p_{mlm}(X_i|X_{\setminus i};\theta)$. This scoring scheme is also implicitly used while performing beam search with the non-autoregressive NMT

[1] This was addressed in their subsequently published erratum: `https://bit.ly/2TXS2KM`.

models proposed in Ghazvininejad et al. (2019). These positive results in prior work suggest that E_{local} is positively correlated with the true sentence scores according to the probabilistic models underlying the trained MLMs.

3 BACKGROUND: METROPOLIS HASTINGS

Metropolis Hastings (Hastings, 1970) is an MCMC algorithm that provides a recipe for sampling from the distribution p via a proposal distribution $q(X'; X, \gamma)$ parametrized by γ, which defines transition from sequence X to the sequence X' in the Markov chain. It assumes the ability to compute the unnormalized score $\phi(X)$ for every sequence X. At each sampling step we first draw a proposal X' from the proposal distribution. Then, we either transition to this new state with the acceptance probability $a(X'; X)$, or repeat the sequence X in the Markov chain. The acceptance probability for the step that ensures that the MCMC sampler satisfies *detailed balance* is: $a(X'; X) = \min\left(1, \frac{\phi(X')\ q(X;X')}{\phi(X)\ q(X';X)}\right)$. Additionally, since it is highly unlikely that the neurally parametrized models like MLMs will assign any sequence a probability 0, it is safe to assume *ergodicity* of the Markov chains with this sampler, which guarantees convergence to the desired target energy network distribution p. In our experiments, the unnormalized score $\phi(X)$ is computed by using the transformer parametrization of the MLM of interest. Both our energy formulations involve computing positional potentials which are obtained by iteratively masking the token at each position and running the forward pass of the MLM transformer.

3.1 MASKED CONDITIONALS AS PROPOSAL DISTRIBUTION FOR THE MH SAMPLER

As we discuss in Section 2.1, the masked conditionals used to train MLMs do not correspond to the two energy formulations we experiment with and are not appropriate for performing Gibbs sampling. In fact, our experiments demonstrate that performing Gibbs sampling using these masked conditionals leads to low-quality samples. However, these conditionals have been shown to be useful for scoring individual sequences and non-autoregressive generation. Therefore, we propose to define the proposal distribution $q(X', X)$ for the Metropolis-Hastings sampler by these masked conditionals. More concretely, to transition from the sequence X, we first mask the token in X at position i, i.e., $X_i = \texttt{[MASK]}$. Next, we do a Transformer decoder pass and get the masked conditionals p_{mlm} at position i. Then, the probability of the transition to sequence X' is the masked probability of the token at the i-th position in X', i.e.: $q(X', X) = p_{mlm}(X'_i | X_{\setminus i}; \theta)$, where $X_{\setminus i} = X'_{\setminus i}$ and $q(X, X') = p_{mlm}(X_i | X_{\setminus i}; \theta)$. For both Gibbs sampling and MH sampling schemes, we sweep over all the positions in a random order while generating sequences of a certain length. We denote one complete sweep over all the positions in a sequence of length T by the term *epoch*. We summarize our general approach in Alg. 1.

Algorithm 1 Metropolis Hastings algorithm for MLMs

1: **Input:** MLM transformer σ, Energy function f_E, MLM conditional proposal f_{mlm} ,sequence length T, number of epochs E
2: **Initialize:** $X \leftarrow \texttt{[MASK]}^T$
3: $X \leftarrow$ greedy-decode(MLM(X)) ▷ Warm-start with a random sequence
4: **for** e=0 to E **do**
5: **for** t=0 to T **do** ▷ left-to-right or random position selection
6: $\mathbf{E}_{\text{old}} \leftarrow f_E(\sigma(X))$ ▷ Energy of sequence X, $\mathcal{O}(T)$ op.
7: $X' \leftarrow X$, $w_o \leftarrow X_t$, $X_t \leftarrow \texttt{[MASK]}$ ▷ Store the t-th token in X as w_o and mask it.
8: $w_n \sim f_{mlm}(\sigma(X), t)$, $X'_t \leftarrow w_n$ ▷ Sample w_n from MLM conditional to propose X'.
9: $\mathbf{q}(\mathbf{X}', \mathbf{X}) = f_{mlm}(\sigma(X), t)[w_n]$, $\mathbf{q}(\mathbf{X}, \mathbf{X}') = f_{mlm}(\sigma(X), t)[w_o]$
10: $\mathbf{E}_{\text{new}} \leftarrow f_E(\sigma(X'))$ ▷ Energy of proposed sequence X', $\mathcal{O}(T)$ op.
11: $\mathbf{a}(\mathbf{X}'; \mathbf{X}) \leftarrow \min\left(1, \frac{\mathbf{e}^{-\mathbf{E}_{\text{new}}}\ \mathbf{q}(\mathbf{X}, \mathbf{X}')}{\mathbf{e}^{-\mathbf{E}_{\text{old}}}\ \mathbf{q}(\mathbf{X}', \mathbf{X})}\right)$ ▷ Acceptance probability of the MC transition.
12: **if** $u \sim \mathcal{U}(0, 1)$, $u \leq a$ **then** $X \leftarrow X'$
13: **Output:** sampled sequence X

Computational complexity: Amortizing the encoder cost and the cost of performing a softmax operation, if we denote the cost of doing one Transformer decoder pass over a masked sequence by C, then the computational complexity of evaluating MLM conditional is $\mathcal{O}(C)$. For E epochs and a sequence of length T, the cost of running a Gibbs sampler is $\mathcal{O}(TEC)$. For the MH sampler,

we additionally need to compute the unnormalized scores $\phi(X)$ which, for both the proposed parametrizations of energy, require masking of each position sequentially and running a Transformer decoder pass for each masked sequence. Hence the MH sampler is more computationally expensive with the complexity $\mathcal{O}(T^2EC)$.

3.2 VARIANTS OF PROPOSAL DISTRIBUTION

We studied our sampler with multiple proposal distributions. While all the variants of proposal distribution rely heavily on the masked conditionals from the pretrained MLM, they have different properties and as shown in the results, they exhibit very different behaviors.

Varying temperature: We experiment by changing the entropy of the masked conditionals via a temperature hyperparameter T: $q(X', X; T) = p_{mlm}(X'_i|X_{\setminus i}; \theta, T) = \text{softmax}(\frac{\log \phi_i}{T})$.

Variation based on Nucleus Sampling: We experiment with another method of changing the entropy of the masked conditional distribution that is inspired by Nucleus Sampling (Holtzman et al., 2019). It involves defining a nucleus boundary b, which prunes out the long tail of the vocabulary that falls outside of the cumulative probability b followed by renormalization over the pruned vocabulary $\mathbb{V}_b$ which is the smallest set such that $\sum_{w \in \mathbb{V}_b} p_{mlm}(X'_i = w|X_{\setminus i}; \theta) \geq b$.

Block MH sampling: Block sampling methods like block Gibbs sampling (Gelfand, 2000) result in better mixing of the Markov chain because they allow for perturbations to multiple variables. In our approach, we mask out multiple tokens in a sequence X in order to propose a new sequence X'. Let $\mathcal{I}$ be the set of positions by which X and X' differ. Then, the proposal distribution for the MH sampler is: $q(X', X) = \prod_{i \in \mathcal{I}} p_{mlm}(X'_i|X_{\setminus \mathcal{I}}; \theta)$. This makes sampling faster due to parallelization of prediction at several positions, and results in generation of better samples.

4 IMPLEMENTATION DETAILS

Pretrained Masked Language Model: We empirically study the proposed Metropolis Hastings scheme on the conditional generation task of neural machine translation (NMT) and the task of uncondtional generation. For *unconditional generation* we used HuggingFace's pytorch implementation[2] of uncased BERT-base and BERT-large. For *NMT*, to perform fair comparison we use the pretrained models[3] optimized by a prominent non-autoregressive algorithm–MASK-PREDICT (Ghazvininejad et al., 2019). This non-probabilistic algorithm uses a bidirectional Transformer (Vaswani et al., 2017) to encode the source-side sequence and trains the target-side bidirectional transformer-based decoder via the MLM objective while performing iterative refinement for decoding.

MCMC details for NMT: For all the sampling baselines, after a burn-in period of 7 epochs, we ran the Markov chain for at least 26 epochs over the dataset. Therefore, for a target sentence of length T, we made at least $T \times 33$ proposals in each Markov chain. For all of our sampling results described, we ran at least 5 Markov chains for each configuration described in the subsequent sections and report *averaged statistics* over these runs.

MCMC details for unconditional generation: For the reported experimental settings, we ran 500 chains for 100 epochs to produce 500 sequences of diverse lengths varying from $15 - 45$. For each of the Markov chains, we randomly select a length and start with a sequence consisting entirely of `[MASK]` tokens. We accept all the proposals until all the masked tokens are filled out in order to start the chain from a random sequence.

Data for NMT: We performed experiments via translating the validation and test sets of the WMT-14 German-English (De-En), and the WMT-16 Romanian-English (Ro-En) datasets and perform the same tokenization and pre/post-processing as Ghazvininejad et al. (2019).

Length prediction for NMT: We follow Ghazvininejad et al. (2019), and use a special `[LENGTH]` token along with the encoder's source side representation to predict the target length. We sample target sentences of different length via batching and padding as necessary.

Evaluating quality of samples: Aside from considering the energy values of the samples under our parametrization and other measures of qualitative evaluation, we report the following automatic

[2] `https://huggingface.co/transformers/pretrained_models.html`
[3] `https://github.com/facebookresearch/Mask-Predict`

metrics for NMT and unconditional generation respectively: 1) BLEU scores on the reference corpus give an idea about the practical quality of samples for conditional generation in low-entropy settings like NMT, 2) GPT2-xl (Radford et al., 2019) sentence perplexity (not token normalized) of random unconditionally generated samples provides a reliable idea of the fluency and fitness of the generated sequence. Compared to the BERT models, GPT-2 is an autoregressive model that has been trained on a larger amount of internet data than the BERT models and is a good language model.

5 METROPOLIS HASTINGS AND *Degenerate* GIBBS SAMPLING FOR MLMS

In this section, we empirically compare our proposed MH Sampling approach with both the energy formulations described in Section 2 (raw and norm) to the alternative proposed by Wang and Cho (2019) of performing Gibbs sampling with the masked free conditionals which we refer as degenerate Gibbs sampling (deg).

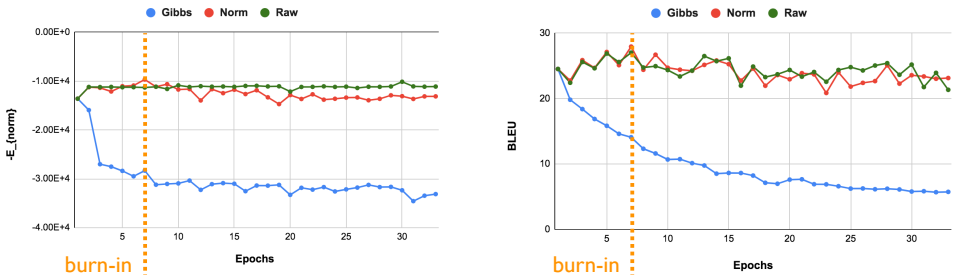

Figure 1: $-E_{\text{norm}}$ (left) and BLEU scores (right) on De-En (20) for NMT as a function of epochs for the two MH schemes (raw and norm) and the degenerate Gibbs sampling scheme (deg). We compute and report $-E_{\text{norm}}$ even for the samplers with E_{raw} parametrization.

In Figure 1, we notice that for NMT, although all the samplers start with the same random sequence, the proposed MH samplers generate high quality samples with low energy values and consistently good BLEU scores across all the epochs. The degenerate Gibbs sampler however, keeps on degrading to generating sequences with very low BLEU scores and high energy values. We also observe that sequences with high BLEU scores typically have low locally normalized energies which explains the success of prior work in using these locally normalized scores for re-ranking beam search hypotheses and generating sequences with high BLEU scores (Salazar et al.; Ghazvininejad et al., 2019). For *open-ended generation* we observed a similar pattern, but the chain with E_{norm} was consistently slightly worse than E_{raw}. This is explored more in subsequent results.

Next, we examine the *acceptance ratio* of the MH samplers. We focus on the average proportion of *novel MC transition rate*–the ratio of proposals that were distinct from the previous state *and* accepted–which indicates the entropy of the MCMC transition distribution. For NMT, the degenerate Gibbs sampler has acceptance probability of $\mathbf{1.0}$ by design and a novel transition ratio of $\mathbf{0.36}$, which indicates that the MLM conditionals are fairly peaked. Both the MH samplers have high acceptance rates ($\mathbf{0.9}$ and $\mathbf{0.91}$) but much lower novel transition ratio–$\mathbf{0.11}$ for RAW and $\mathbf{0.13}$ for NORM. This indicates slow mixing of the MH Markov chain. For unconditional generation the novel transition ratio of the dengenerate sampler is higher $\mathbf{0.58}$, but it is slightly lower for the MH samplers–$\mathbf{0.10}$ for RAW and $\mathbf{0.08}$ for NORM. This suggests, that the BERT conditionals yield proposals that are rejected at a very high rate under our parametrization schemes for open ended generation.

6 RESULTS WITH VARIANTS OF PROPOSAL DISTRIBUTIONS

6.1 EFFECT OF TEMPERATURE

In this section, we explore the effect of temperature on the proposal distributions parametrized by the MLM conditionals as described in Section 4.1, varying the proposal distributions from high entropy to low entropy. In Tables 1 and 2, we see that for MT, the MH sampler performs similarly across all the temperature with the performance improving slightly for lower temperature values, however unconditional generation is significantly more sensitive to the temperature changes, with

worse performance at the higher temperature. The degenerate Gibbs sampler in general trails behind MH samplers but drastically improves with the lowering temperature values. At low temperatures, it yields decent BLEU scores and more fluent sentences but it is noteworthy that the energy values are worse than the MH sampler. Most interestingly, the *novel* transition rates reflect the effect

Table 1: Average $E_{norm} \times 10^{-3}$ energy, novel MC transition rate, and BLEU scores for NMT across interleaved epochs for the degenerate Gibbs sampling (deg) and the locally normalized energy MH scheme (Norm) on De-En (20) under MLM proposal distributions with varying temperatures.

Temp	2.0		1.5		1.0		0.8		0.5	
	norm	deg	norm	deg	norm	deg	norm	deg	norm	deg
$E_{norm} \downarrow$	12.87	32.46	10.21	29.57	11.13	31.12	9.95	21.12	7.85	17.65
Novel $\leftrightarrow$	0.03	1.0	0.05	0.97	0.11	0.36	0.06	0.08	0.03	0.04
BLEU	25.91	14.53	24.78	10.12	24.74	9.03	25.84	24.77	27.23	26.12

Table 2: Average E_{raw} energy, novel MC transition rate, and average GPT-2 sentence perplexity for unconditional generation across generated sequences for the degenerate Gibbs sampling (deg) and the raw energy MH scheme (raw) under MLM proposal distributions with varying temperatures.

Temp	1.2		1.0		0.8		0.5	
	raw	deg	raw	deg	raw	deg	raw	deg
$E_{raw} \downarrow$	25.95	201.24	19.82	83.23	13.54	23.24	9.91	10.05
Novel $\leftrightarrow$	0.09	0.82	0.09	0.81	0.08	0.25	0.06	0.09
GPT-2 $\downarrow$	223.87	2238.6	108.12	314.74	82.23	88.15	77.44	77.98

of temperature very clearly. At high temperatures, the degenerate Gibbs sampler never proposes very few repeating transitions while in stark contrast, the novel transition rate of the MH sampler is extremely low. This is because of high rejection rates under the unsuitable high-entropy proposal distribution. While the BLEU/energy results for low-temperature settings seem to suggest that the degenerate Gibbs samplers are practically useful samplers, examining novel transition rates dispels this suggestion. At low temperatures, the novel transition rate is extremely small for the degenerate sampler indicating low-entropy of the MLM based transition distribution which in turn reduces the novel transition rates of the MH sampler as well. Hence, the impressive low-temperature results only corroborate the results of recently proposed non-probabilistic MLM-based generation models like MASK-PREDICT (Ghazvininejad et al., 2019) that do not explore the sequence space at all.

6.2 Effect of Nucleus Sampling

Adjusting the nucleus boundary can only decrease the entropy of the MLM proposal distribution. In Table 3, we observe effects of low-entropy proposal distribution that are similar to effects of lowering the temperature–decrease in novel transition rate with the samplers fixating around decent samples.

Table 3: Average $E_{norm} \times 10^{-3}$ energy, novel MC transition rate, and BLEU scores energy for the degenerate Gibbs sampling (deg) and the locally normalized energy MH scheme on De-En (20) under MLM proposal distributions with varying nucleus.

Nucleus	1.0		0.99		0.95		0.90		0.80	
	norm	deg	norm	deg	norm	deg	norm	deg	norm	deg
$E_{norm} \downarrow$	11.13	31.12	10.65	30.12	10.21	28.75	9.95	18.57	9.85	18.23
Novel $\leftrightarrow$	0.11	0.36	0.12	0.33	0.10	0.22	0.07	0.10	0.05	0.06
BLEU	24.74	9.03	24.95	14.03	26.15	18.04	27.35	23.25	27.23	23.55

These patterns of sensitivity to the proposal distribution's entropy (Tables 1,2, 3) strongly suggest that while the MLM objectives results in conditionals whose mode corresponds to high quality sequences, these conditionals are poorly calibrated and are not suitable for exploring the distribution over sequences via direct *sampling*. Our proposed technique exhibits robustness and good performance because it uses the MLM conditionals only to define energy scores, not directly sample with.

6.3 EFFECT OF BLOCK MH SAMPLING

In the results so far, we have observed that while the MH samplers yield good samples, their novel transition rate (**0.11–0.13**) is fairly low which results in slow mixing of the Markov chain. To improve the mixing rate we experiment with the proposal distribution for block MH sampling as describe in section 4.1. Because perturbations in a large number of positions also increase the chance of rejection of the new MH proposal, we balance exploration with tolerable rejection by annealing the number of masked positions with epochs. At the start of the Markov chain, we make large changes, but gradually make smaller changes as the chain progresses (details in Appendix). We also, experiment with a *block Gibbs* sampling variant of our degenerate Gibbs sampler. This block Gibbs sampler is incorrect as well, however, it is interesting to study because with temperature $T = 0.0$, it yields the MASK-PREDICT (Ghazvininejad et al., 2019) algorithm. We specify the results while keeping the other settings like temperature and nucleus boundary at their default value of 1.0.

Table 4: **Left:** Average BLEU scores, $E_{norm} \times 10^{-3}$, and novel transition rates, **Right:** Average GPT-2 sentence perplexity, E_{raw}, and novel transition rates for the two Block variants of the MH schemes (raw and norm) and the degenerate block Gibbs sampling scheme (deg).

	Deg	Raw	Norm
$E_{norm} \downarrow$	31.18	8.08	8.17
Novel $\leftrightarrow$	0.77	0.40	0.41
BLEU	9.03	27.12	26.78

	Deg	Raw	Norm
$E_{raw} \downarrow$	81.87	18.43	17.67
Novel $\leftrightarrow$	0.80	0.21	0.20
GPT-2 $\downarrow$	166.29	43.21	92.23

In Table 4, we notice that degenerate block Gibbs sampler performs poorly, while both the MH samplers show improvements in terms of BLEU, energy values, and GPT-2 scores over previous non-block MH sampling settings under default conditions. For unconditional generation, we see a clear difference in performance between the energy parametrizations with E_{raw} being superior to E_{norm}. This indicates that for high-entropy, complex target distributions the normalization constraint in E_{norm} might be hindering the learning of expressive energy functions. Moreover we notice that while our block-sampling scheme drastically increases the novel transition rate ($\approx$ **0.12** $\rightarrow$ **0.41**) for MT, the increase is less impressive for unconditional generation. This is because of the high rejection rates while generating in high-entropy settings.

7 ANNEALING THE ENERGY FUNCTION: SAMPLING AROUND THE MODES

In this section, we analyze the effectiveness of our proposed MH samplers by evaluating the samples drawn from regions around the mode of the energy functions and evaluating them against references for the task of MT. To achieve this, we propose to perform MH sampling from target distributions whose energy values are scaled by low temperatures i.e. $p(X; \theta, T) \propto e^{\frac{-E(X;\theta)}{T}}$. However, such low-entropy target distributions lead to increased rejection rates for the MH samplers. Therefore, we anneal the temperature as a linear function of epochs to gradually decrease the entropy of the target distribution. In Figure 2 (left, green), we observe that annealing results in dramatic improvements in locally normalized energy scores E_{norm}, leading to very low energy values. When comparing

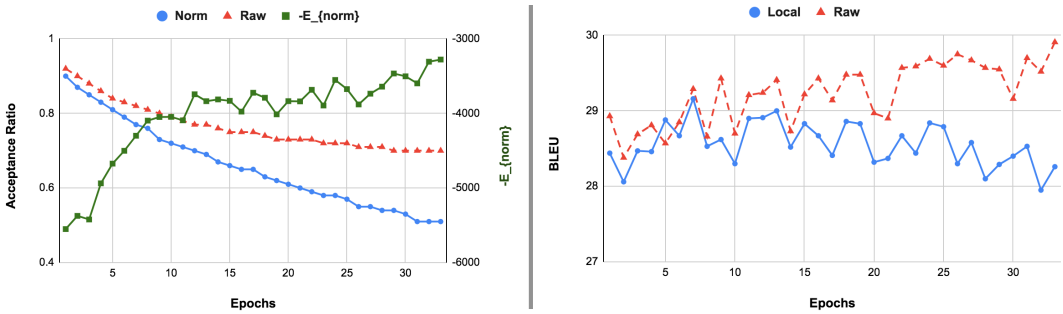

Figure 2: Comparison as a function of epochs for the two Energy parametrizations (red, blue) for Metropolis Hastings approach with annealing toward modes of target energy functions: E_{raw} (raw) and E_{local} on De-En (20). **Left:** acceptance rates, locally normalized energy (green) as a function of epochs. **Right:** MT performance.

the acceptance rates, we see that raw and the locally normalized energy parametrizations behave

differently as the target distribution temperature is annealed, with the MH samplers under the raw scores target distribution admitting larger acceptance rates across the epochs. This difference in acceptance rates also manifests itself in the performance in terms of BLEU scores of the samples under two energy parametrizations, with raw energy parametrization yielding higher BLEU scores.

Table 5: Performance of annealing based approach for sampling around the energy-based distributions' modes. BLEU scores reported on the full De-En and Ro-En test sets.

Baseline	De-En	Ro-En
Warm-start	20.27	24.38
Degenerate Gibbs (T=0.8)	27.88	29.79
Mask-predict (beam=1, It=10)	29.27	29.95

MH samplers	De-En	Ro-En
Local Energy	**29.74**	**31.13**
Raw Score Energy	**30.12**	**30.86**

In Table 5, we compare the performance of our annealing-based mode-finding approach on the task of machine translation with other related algorithms (details in Appendix). `Warm-start` refers to the greedy replacement of all the mask tokens with the pretrained MLM which is used as the starting sequence for our Markov chains. While it performs reasonably well, all the other approaches outperform it. We mainly compare our approach (Local and Raw score Energy) to the MASK-PREDICT algorithm (Ghazvininejad et al., 2019)–a prominent non-probabilistic and non-autoregressive generation technique, which also provides the pretrained conditional MLM for our MH samplers. We outperform both, the degenerate Gibbs sampling with temperature of 0.8 and MASK-PREDICT with beam-size 1, and 10 epochs. [4]. Although, the autoregressive approach is superior (30.18 (de-en) and 31.53 (ro-en)) to the MASK-PREDICT baseline, we perform competitively with it. As better MLM based NAT models are proposed for translation, our approach provides a way to interpret them probabilistically and draw samples from them.

8 EVALUATION OF UNCONDITIONAL GENERATION

To validate our findings on unconditional generation, we conduct human evaluation of our most basic setup–default parameters with non-block MH sampling. 3 humans familiar with the generation capabilities of language models were presented with 120 sentences generated by BERT-base and BERT-large with our proposed samplers, and were asked to provide 4-point likert ratings along two axes: *coherence* and *fluency*. The results in Table 6 indicate that GPT-2 scores are a reliable

Table 6: Coherence, Fluency (averaged across examples and humans), average GPT-2 sentence perplexity for sentences generated unconditioned by the degenerate Gibbs sampler (deg), and proposed MH samplers (norm an raw) with 2 different MLMs: BERT-base (base) and BERT-large (large).

	base-deg	base-norm	base-raw	large-deg	large-norm	large-raw
coherence	1.23	1.6	2.05	1.15	1.7	2.0
fluency	1.2	1.82	2.45	1.15	1.9	2.25
GPT-2 $\downarrow$	296.45	184.21	125.42	310.22	175.43	132.76

measure for comparing the systems we are interested in. Also, we notice that our samplers clearly outperform degenerate Gibbs sampling scheme and E_{raw} is better suited for unconditional generation than E_{norm}. Our samplers generally are more fluent than coherent. This is expected because pure unconditional generation is not constrained by any conditioning context resulting in low coherence. Interestingly, there is little difference in sample quality between BERT-base and BERT-large.

9 CONCLUSION

Our proposed Metropolis-Hastings based sampler enables us to draw high-quality samples from non-probabilistic masked language models. The empirical analysis and success of our approach with the two proposed energy parametrizations strongly suggests that the optimization of MLM objective results in training of an implicit global energy network that induces probability distribution over the space of sequences and its possible to sample from it using our method. While we primarily focus on sampling and generation, our findings open up avenues for devising more direct, stable and simple training (Deng et al., 2020) procedures for energy-based sequence models inspired from the MLM objectives and our proposed MH sampling scheme.

[4]Our MH samples are also competitive with the best performance using MASK-PREDICT with beam-size 5, 10 epochs–30.52 De-En, and 31.57 Ro-En.

REFERENCES

D. Belanger and A. McCallum. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992. PMLR, 2016.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

K. Clark, M.-T. Luong, Q. Le, and C. D. Manning. Pre-training transformers as energy-based cloze models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 285–294, Online, Nov. 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.20. URL https://www.aclweb.org/anthology/2020.emnlp-main.20.

K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020b.

Y. Deng, A. Bakhtin, M. Ott, A. Szlam, and M. Ranzato. Residual energy-based models for text generation. *arXiv preprint arXiv:2004.11714*, 2020.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004.

Y. Du and I. Mordatch. Implicit generation and modeling with energy based models. 2019.

A. E. Gelfand. Gibbs sampling. *Journal of the American statistical Association*, 95(452):1300–1304, 2000.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

A. Gelman and T. E. Raghunathan. Using conditional distributions for missing-data imputation. *Statistical Science*, 15:268–69, 2001.

M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

K. Goyal, C. Dyer, and T. Berg-Kirkpatrick. An empirical investigation of global and local normalization for recurrent neural sequence models using a continuous relaxation to beam search. In *Proceedings of 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

W. Grathwohl, K. Swersky, M. Hashemi, D. Duvenaud, and C. J. Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. *arXiv preprint arXiv:2102.04509*, 2021.

J. Gu, J. Bradbury, C. Xiong, V. O. Li, and R. Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1 (Oct):49–75, 2000.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

D. Lowd. Closed-form learning of markov networks from dependency networks. *arXiv preprint arXiv:1210.4896*, 2012.

E. Mansimov, A. Wang, S. Welleck, and K. Cho. A generalized framework of sequence generation with application to undirected sequence models. *arXiv preprint arXiv:1905.12790*, 2019.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021.

J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff. Masked language model scoring.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

L. Tu, R. Y. Pang, S. Wiseman, and K. Gimpel. Engine: Energy-based inference networks for non-autoregressive machine translation. *arXiv preprint arXiv:2005.00850*, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

A. Wang and K. Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, and L. Carin. Adversarial feature matching for text generation. In *International Conference on Machine Learning*, pages 4006–4015. PMLR, 2017.

J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

A APPENDIX

B ADDITIONAL DETAILS ABOUT THE GRAPHS

All the graphs in the main paper and Appendix have `epochs` on the x-axis. These epochs correspond to the **post burn-in** epochs. Since the Markov chains under comparison are warm-started with the same sequence, a burn-in period of 7+ epochs is sufficient to decouple the chains so that their behavior can be studied and compared reliably.

As described in Figure 1 of the main paper, we plotted the computed locally normalized energy $-E_{norm}$ for all the three sampling schemes i.e.: 1) degenerate Gibbs sampling, 2) MH sampling with energy parametrized via locally normalized MLM scores (Section 3.2), and 3) MH sampling with energy parametrized via raw scores (Section 3.1).

In Figure 3 of the Appendix, we show the raw energy scores $-E_{raw}$ and the locally normalized energy scores $-E_{norm}$ of the samples for a Markov chain corresponding to MH sampling with energy parametrized via raw scoring (Section 3.1). We observe that the trend followed by $-E_{raw}$ is similar as $-E_{norm}$, which suggests that both parametrizations of energy networks are correlated with each other.

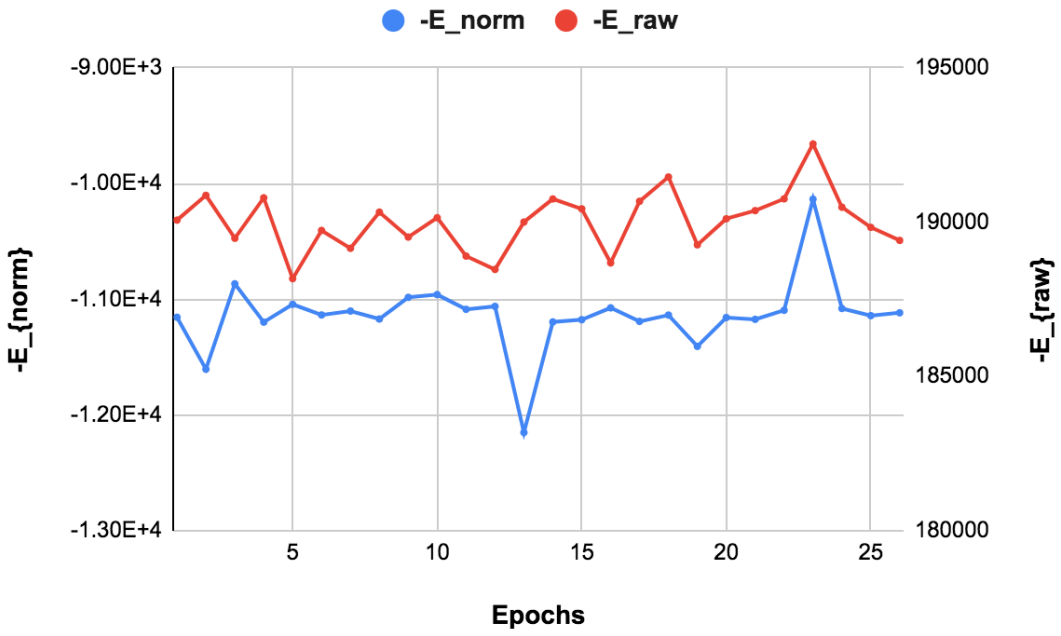

Figure 3: $-E_{norm}$ (left axis, blue) and $-E_{raw}$ (right axis, red) on De-En (20) as a function of epochs (post burn-in) for a run of proposed MH sampling scheme with raw energy scores defining the target distribution.

C DETAILS ON THE COUNTER-EXAMPLE IN SEC-2.1

We demonstrate via the following counter-example, that the learned free conditionals from an MLM need not correspond to *any* consistent MRF over the sequences. Consider a sequence of length 2 with the random variables X_1, X_2 that have a categorical distribution over a vocabulary $\mathbb{V} = \{a, b\}$. We provide an example of an *inconsistent* set of conditional distributions over X_1 and X_2 that a model like MLM, which estimates the conditionals independently, is capable of learning. The following free conditionals are inconsistent because they do not correspond to any valid joint distribution over $\{X_1, X_2\}$.

$$p(X_1 \mid X_2) = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}, p(X_2 \mid X_1) = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

These conditional distributions do not correspond to a valid joint distribution Intuitively, because X_2 is extremely predictive of X_1 but X_1 does not predict X_2 at all from the above conditionals, they cannot characterize a consistent joint distribution over $\{X_1, X_2\}$. Furthermore, these conditionals can be shown to violate the Bayes' rule. If use the provided conditionals $p(X_2|X_1)$ to compute the marginals: $p(X_2 = a) = \sum_{w \in a,b} p(X_2 = a|X_1 = w)$ and similarly compute $p(X_2 = b)$ as well then this inequality clearly follows:

$$\frac{p(X_1 = a, X_2 = a)}{p(X_1 = a, X_2 = b)} \propto \frac{p(X_1 = a \mid X_2 = a)}{p(X_1 = a \mid X_2 = b)} = 99$$
$$\neq$$
$$\frac{p(X_1 = a, X_2 = a)}{p(X_1 = a, X_2 = b)} \propto \frac{p(X_2 = a \mid X_1 = a)}{p(X_2 = b \mid X_1 = a)} = 1.$$

D EFFECT OF ANNEALING BLOCK SIZE

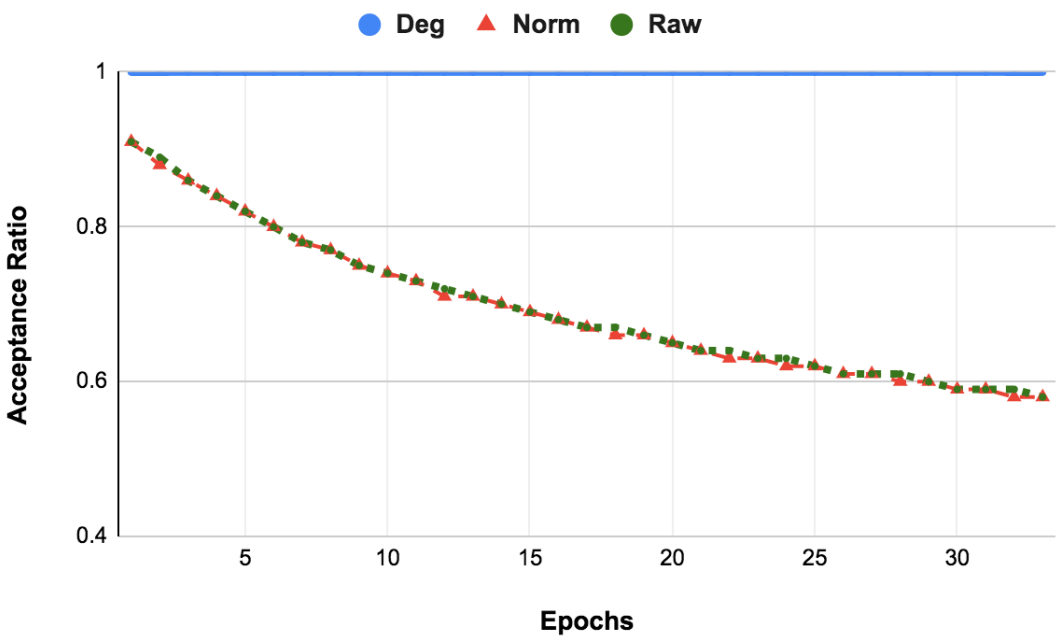

Figure 4: Acceptance ratio on De-En (20) as a function of epochs for the two Block variants of the Metropolis Hastings schemes (raw and locally normalized energy parametrizations) and the degenerate block Gibbs sampling scheme (deg).

We anneal the block sizes as a function of iteration i.e. larger block sizes at the beginning for fast mixing and smaller block sizes (eventually reducing to blocksize of 1). The effect of annealing the size of the blocks becomes clearer in Figure 4. The initial high acceptance rates indicate fast mixing and iterative refinement of the random initial sequence. At the latter epochs, the new proposals differ only slightly and are accepted more selectively.

E POTENTIAL IMPACT AND USAGE OF DATA

Our contribution is mostly algorithmic in that we attempt to interpret masked language models as energy-based sequence models and devise a strategy to generate samples from the probability

distribution over the sequences implicitly induced by the MMLM training objective. While there is no direct negative societal impact of our approach, we can envision an indirect impact of performing natural language generation with masked language models. Because MLMs like BERT are trained on vast amounts of data on the internet, they capture unsavory biases and stereotyopes in the text used to train them. So farm they have primarily been used to encode natural language sentences and fine-tuned for few-shot learning for specialized tasks. Our approach focuses on drawing samples from these models in a principled manner which would enable these models to be used for natural language generation as well.

We used the WMT datasets (2014 De-En, 2016 Ro-En) for our experiments which are public datasets that serve as standard datasets for training and evaluating the machine translation systems. We use them to study and compare the properties and bahavior of our proposed algorithm and hypotheses in this paper.

F EXPERIMENTAL SETUP FOR TABLE 5

Since this experiment focused on obtaining samples from around the mode of the energy parametrizations and comparing them against existing non-autoregressive baselines, to reduce the variance across runs, we used a temperature of proposal distribution as 0.8. As is evident from Table 1 (Novel transitions), this does reduce the diversity in generated samples, but the samples generated have high BLEU scores. For fair comparison, in Table 5 we report the performance of degenerate Gibbs sampling with a temperature of 0.8 as well. As can be noticed from the experiments throughout the paper, non-probabilistic greedy mask-based generation approaches tend to perform well when the only desriable attribute of the generated sequence is high BLEU score but fail spectacularly when the MLM conditionals are used for diverse samples characterizing a distribution over the sequences. Another trick to reduce vairance was to artificially set the acceptance rate to 1.0 for the first two epochs (after warm-starting) such that all the Markov chains have similar trajectory. We revert back to regular MH sampling acceptance rates after this initialization. As a result, we obtain similar results for Table 5 across all the Markov chain runs and the standard deviation in BLEU scores is very low at $\mathbf{0.13}$.

G ERROR BARS

For the results discussed in Figure 1, we report standard deviation of the difference between the quantities in the first Markov chain and all the other Markov chains run in our experiments. For our MH based approach, we observed standard deviation of 1.1 in BLEU scores and 0.38×10^3 for E_{norm}.

H EFFECT OF ANNEALING STRENGTH

Table 7: Comparison of different linear annealing strengths via best BLEU and E_{norm}, and average acceptance ratio of the two variants of the MH schemes (raw and local) on De-En (20).

Anneal	0.02			0.04			0.06		
	$E_{norm} \downarrow$	accept	BLEU	$E_{norm} \downarrow$	accept	BLEU	$E_{norm} \downarrow$	accept	BLEU
Norm	3277.6	0.52	29.16	3245.65	0.45	27.58	3187.80	0.41	26.95
Raw	3146.4	0.71	29.91	3088.65	0.62	28.32	2146.34	0.58	27.21

In Table 7, we show the effect of different annealing schedules. At each epoch, we subtract either $0.02, 0.04$, or 0.06 from the temperature of the target energy function. We see that 0.02 annealing schedule yields the best BLEU scores but interestingly, more aggressive schedules result in better (lower) energy values with lower acceptance rates.