Efficient Informed Proposals for Discrete Distributions via Newton's Series Approximation

Yue Xiang
Renmin University of China

Dongyao Zhu
Independent Researcher

Bowen Lei
Texas A&M University

Dongkuan Xu
North Carolina State University

Ruqi Zhang
Purdue University

Abstract

Gradients have been exploited in proposal distributions to accelerate the convergence of Markov chain Monte Carlo algorithms on discrete distributions. However, these methods require a natural differentiable extension of the target discrete distribution, which often does not exist or does not provide effective gradient guidance. In this paper, we develop a gradient-like proposal for any discrete distribution without this strong requirement. Built upon a locally-balanced proposal, our method efficiently approximates the discrete likelihood ratio via Newton's series expansion to enable a large and efficient exploration in discrete spaces. We show that our method can also be viewed as a multilinear extension, thus inheriting its desired properties. We prove that our method has a guaranteed convergence rate with or without the Metropolis-Hastings step. Furthermore, our method outperforms a number of popular alternatives in several different experiments, including the facility location problem, extractive text summarization, and image retrieval.

1 INTRODUCTION

Discrete structures are common in the real world, from discrete data such as text [Wang and Cho, 2019, Gu et al., 2017] and genomes [Wang et al., 2010], to discrete models such as low-precision neural networks [Courbariaux et al., 2016, Peters and Welling, 2018] and graphical models of molecules [Gilmer et al., 2017]. As data and models become complex and large-scale, it is desirable to develop efficient proposals in Markov chain Monte Carlo (MCMC) algorithms that allow us to sample from these complex high-dimensional discrete distributions [Zhang et al., 2022a].

Gradients have been widely utilized in proposal distributions to accelerate the convergence of MCMC, such as the Langevin algorithm [Roberts and Tweedie, 1996, Roberts and Stramer, 2002] and Hamiltonian Monte Carlo (HMC) [Duane et al., 1987, Neal et al., 2011]. These gradient-based methods are mainly designed for continuous distributions and require a natural neighborhood to define gradients. However, since there is no natural neighborhood in discrete distributions, it becomes challenging to incorporate gradients in the proposal to accelerate sampling.

Previous research has been devoted to making gradient-based proposals for efficient discrete sampling, however, they either require natural differentiable relaxation or sacrifice convergence speed. As shown in the "Natural continuous extension available" column in Table 1, Gibbs with gradient proposal [Grathwohl et al., 2021] and discrete Langevin proposal (DLP) [Zhang et al., 2022a] assume the existence of an underlying differentiable distribution in the discrete space and exploit gradient information to speed up sampling and inference. In the top right of Table 1, the locally-balanced proposal [Zanella, 2020] does not require this assumption, while it only conducts local moves in small windows, which leads to slow convergence, especially in high-dimensional tasks. Therefore, the question we need to answer is how to design a method that can sample efficiently in discrete spaces and has no strong requirement

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

Table 1: Proposals for discrete distributions.

	Natural differentiable extension available	Natural differentiable extension unavailable
Update one coordinate in a step	Gibbs with gradient proposal	Locally-balanced proposal
Update multiple coordinates in a step	Discrete Langevin proposal	**Newton proposal (Ours)**

of natural continuous expansion.

To answer this question, we present a gradient-like informed proposal to efficiently sample from any discrete distribution without the strong requirement of continuous relaxations. To find more informative directions during sampling, our method estimates the likelihood ratio of making discrete moves through Newton's series expansion, so we call our proposal *Newton proposal*. In addition, we design a coordinatewise factorization scheme in our method, so we can update multiple coordinates in a single move, which further improves the sampling efficiency. We summarize our contributions as follows:

- We propose a new informed proposal, Newton proposal, for discrete distributions. It allows multiple coordinates to be updated simultaneously while not requiring that the discrete distribution to be naturally extended to the continuous domain.
- We show that our Newton proposal can be obtained from Newton's series approximation to the target discrete distribution or from Taylor expansion to the multilinear extension of the discrete distribution, which justifies Newton proposal's desirable properties.
- We theoretically prove the convergence rate of our Newton scheme without and with the Metropolis-Hastings correction, demonstrating its efficient sampling in discrete distributions.
- We experimentally show that Newton proposal outperforms existing discrete proposals and some optimization-based methods when sampling from high-dimensional complex discrete distributions, including facility location, text summarization, and image retrieval.

2 RELATED WORK

Informed Proposal Various informed Metropolis-Hastings (MH) proposal distributions have been designed to avoid slow mixing and slow convergence brought by random walk MH proposals. Using symmetric proposal distributions, random walk MH schemes are easy to implement, but no information about the target distribution is utilized and the new state is proposed randomly. In the contrary, informed proposal distributions elaborate information about the target distribution, such as the gradient of the target to bias the proposal distribution towards high probability, resulting in substantial improvements of MCMC performances. However, most of these informed proposals are based on derivatives and it is nontrivial to extend such methods to discrete spaces. As a consequence, most MCMC proposals for discrete spaces often rely on symmetric and uninformed proposal distributions, which can induce slow convergence.

Continuous Relaxation-Based Method Gradient-based informed proposals can be applied to discrete distributions via continuous relaxations [Pakman and Paninski, 2013, Nishimura et al., 2020, Han et al., 2020, Zhou, 2020, Jaini et al., 2021, Zhang et al., 2022b]. They are usually implemented by transporting the problem into a continuous domain, performing updates under gradient-based proposals there, and transforming back after sampling. The efficiency of this kind of continuous relaxation highly depends on the properties of the relaxed continuous distributions which may be arbitrarily difficult to sample from, such as being highly multi-modal. To avoid these pitfalls, for discrete distributions which can be displayed as continuous, differentiable functions accepting real-valued inputs but are evaluated only on a discrete subset of their domain, Gibbs-with-gradient proposal [Grathwohl et al., 2021] and discrete Langevin proposal [Zhang et al., 2022a] use gradients to inform discrete updates directly for these discrete distributions rather than transport the discrete domain to a continuous one. However, most discrete distributions in the real world do not have a natural continuous extension, or the natural extension is still not differentiable. This is why we propose Newton proposal.

Locally-Balanced Proposal Based on local neighborhood information at the current location, the locally-balanced proposal [Zanella, 2020] is an informed framework that is applicable to both discrete and continuous spaces. When sampling from discrete distributions, it does not require natural differentiable extensions. Locally-balanced proposals have been extended to continuous-time Markov processes [Power and Goldman, 2019] and have been tuned via mutual information [Sansone, 2022]. It has also been used in Multiple-try Metropolis (MTM) algorithms to achieve fast convergence [Gagnon et al., 2022]. It is very expensive to construct locally-balanced proposals when the local neighborhood is large or the dimension is high, preventing them from making large moves in discrete spaces. The path auxiliary proposal [Sun et al., 2021] explores a larger neighborhood by making a sequence of small moves. An adaptive locally-balanced proposal (ALBP) [Sun et al., 2022] has been proposed to determine the update size automatically. However, it still only updates one coordinate per gradient computation and the update has to be done in sequence. On the contrary, our Newton proposal can update many coordinates in parallel.

3 PRELIMINARY

We consider sampling from a target distribution

$$\pi(\theta) = \frac{1}{Z}\exp(U(\theta)), \quad \forall\theta \in \Theta \tag{1}$$

where θ is a d-dimensional variable, Θ is a finite variable domain, the energy function U is a scalar-valued function, and Z is the normalizing constant for π to be a distribution. In this paper, we restrict Θ to a factorized domain, *i.e.*, $\Theta = \prod_{i=1}^{d}\Theta_i$, and mainly consider Θ to be $\{0,1\}^d$ or $\{0,1,\ldots,L-1\}^d$, which correspond to the binary variable and the categorical one, respectively.

As we state in related work, Locally-balanced proposal [Zanella, 2020] does not require a natural continuous distribution, so it is a flexible framework to build efficient and informed proposals for discrete distributions:

$$Q_{g,\sigma}(\theta, d\theta') \propto g\left(\frac{\pi(\theta')}{\pi(\theta)}\right) K_\sigma(\theta, d\theta'), \tag{2}$$

where g is a continuous function from $[0,\infty)$ to itself satisfying $g(t) = \text{tg}(1/t)$, $\forall t > 0$. $K_\sigma(\theta, d\theta')$ is a symmetric kernel and σ is a scale parameter.

When we set $g(t) = \sqrt{t}$, $K_\sigma(x,\cdot) = N\left(x,\sigma^2\right)$ and $\alpha = \sigma^2$ as the well-known Metropolis-Adjusted Langevin Algorithm (MALA) proposal [Roberts and Rosenthal, 1998], we get an informed proposal as

$$q_0\left(\theta' \mid \theta\right) \propto \exp\left(\frac{U(\theta') - U(\theta)}{2} - \frac{\|\theta' - \theta\|^2}{2\alpha}\right), \tag{3}$$

where the local difference $U(\theta') - U(\theta)$ shows the likelihood ratio between a given input θ and other discrete states θ'. In case where summing over the full space of θ' is so expensive that the normalizing constant becomes intractable, the locally-balanced proposal often restricts its domain to a small neighborhood. For example, the Gibbs-with-gradient proposal [Grathwohl et al., 2021] only considers local moves inside a Hamming ball with small window sizes.

Finite Difference Finite Difference is a mathematical expression of the form $f(x+b) - f(x+a)$, which is an approximation of derivatives. Specifically, a forward finite difference, denoted $\Delta_h[f]$, of a function f is defined as

$$\Delta_h[f](x) = f(x+h) - f(x).$$

When the window size $h = 1$, h can be omitted:

$$\Delta[f](x) = f(x+1) - f(x).$$

As for the finite difference with respect to a vector, let's consider $x \in \{0,1,\ldots,L-1\}^d$ as a d-dimensional vector and $f : \{0,1,\ldots,L-1\}^d \to \mathbb{R}$ as a scalar-valued function. The finite difference of f with respect to the vector x is defined as

$$\Delta[f](x) = \left(\Delta[f](x)_1, \ldots, \Delta[f](x)_d\right).$$

Specifically,

$$\Delta[f](x)_i = f\left(\neg_i x\right) - f(x),\ \forall i \in \{1,\ldots,d\},$$

where $\neg_i x$ changes the i-th coordinate from x_i to $x_i + 1$ and keeps the other $d-1$ coordinates the same as x.

Newton's Series Expansion As the discrete analog of the continuous Taylor expansion, Newton's series expansion is used to approximate discrete functions. In Newton's series expansion, we use finite differences instead of gradients to indicate neighborhood information.

Let's consider the scalar version first. The Newton series consists of the terms of the Newton forward difference equation:

$$f(x) = \sum_{k=0}^{\infty}\frac{\Delta^k[f](a)}{k!}(x-a)_k$$

where $(x)_k = x(x-1)(x-2)\cdots(x-k+1)$ and $\Delta^k[f](x)$ represents k-th order forward finite difference defined as

$$\Delta^k[f](x) = \sum_{i=0}^{k}\left(\begin{array}{c}k\\ i\end{array}\right)(-1)^{k-i}f(x+i).$$

Specifically, the first-order Newton's series expansion is

$$f(x) \approx f(a) + \Delta[f](a)(x-a). \tag{4}$$

When x and a are d-dimensional vectors, $\Delta[f](a)$ is also a d-dimensional vector and the corresponding first-order Newton's series expansion becomes

$$f(x) \approx f(a) + \Delta[f](a)^\top \cdot (x-a). \tag{5}$$

In this paper, we use the first-order Newton's series expansion under $h = 1$ to approximate the likelihood of discrete moves.

4 EFFICIENT INFORMED PROPOSALS VIA NEWTON'S SERIES APPROXIMATION

To sample from discrete distributions efficiently, we propose Newton proposal. We use Newton's series expansion to approximate the likelihood of making discrete updates, and factorize the discrete domain coordinatewise to reduce the computation cost significantly.

4.1 Informed Proposal via Newton's Series Approximation

Consider a common d-dimensional case $\Theta = \{0, 1, \cdots, L-1\}^d$. In each iteration, several coordinates are flipped and we update the current samples θ to θ'. We use a first-order forward Newton's series expansion with window size $h = 1$ to approximate $U(\theta') - U(\theta)$:

$$U(\theta') - U(\theta) \approx \Delta[U](\theta)^\top \cdot (\theta' - \theta). \tag{6}$$

We use the Newton's series expansion in (6) to approximate the local difference $U(\theta') - U(\theta)$ in (3):

$$\begin{aligned} & q\left(\theta' \mid \theta\right) = \widehat{q_0}\left(\theta' \mid \theta\right) \\ =& \frac{1}{Z_\Theta(\theta)} \exp\left(\frac{\Delta[U](\theta)^\top \cdot (\theta' - \theta)}{2} - \frac{\|\theta' - \theta\|^2}{2\alpha}\right) \\ \propto & \exp\left(\frac{1}{2\alpha}\left(-\left(\theta' - \theta\right)^2 + \alpha \Delta[U](\theta)^\top \cdot (\theta' - \theta)\right)\right) \\ & \cdot \exp\left(-\frac{\alpha}{8}\Delta[U](\theta)^2\right) \\ =& \exp\left(-\frac{1}{2\alpha}\|\theta' - \theta - \frac{\alpha}{2}\Delta[U](\theta)\|^2\right) \end{aligned} \tag{7}$$

We add a term in the fourth line to get the perfect square form because $\Delta[U](\theta)$ is independent of θ' and will not affect the normalized result (see the appendix for detailed proof).

In this way, we obtain the new proposal in a perfect square form by Newton's series expansion:

$$q\left(\theta' \mid \theta\right) = \frac{\exp\left(-\frac{1}{2\alpha}\left\|\theta' - \theta - \frac{\alpha}{2}\Delta[U](\theta)\right\|_2^2\right)}{Z_\Theta(\theta)} \tag{8}$$

where the normalizing constant is summed over Θ:

$$Z_\Theta(\theta) = \sum_{\theta' \in \Theta} \exp\left(-\frac{\left\|\theta' - \theta - \frac{\alpha}{2}\Delta[U](\theta)\right\|_2^2}{2\alpha}\right).$$

In a word, finite difference in Newton's series approximation serves as a guide when exploring the discrete space, similar to what gradients do in proposals for continuous distributions. It provides neighborhood information about the target distribution so that the sampler can propose new states more informatively rather than "blindly". Moreover, the perfect square form gives us possibility to accelerate the computation without restrict our domain in a small neighborhood, which we will discuss in detail later.

4.2 Efficient Newton Proposal via Coordinatewise Factorization

The computation cost of the informed proposal in (8) depends on the normalizing constant $Z_\Theta(\theta)$ in (9), which needs to sum up all states in the discrete space. Therefore, it is desirable to narrow down the space to make $Z_\Theta(\theta)$ tractable.

Unlike most locally-balanced proposals which restrict the domain to a small neighborhood, $e.g.$, a hamming ball [Zanella, 2020, Grathwohl et al., 2021], a key feature of the proposal (8) is that it is displayed as a Euclidean norm and can be factorized coordinatewise [Zhang et al., 2022a]. To see this, we write (8) as $q\left(\theta' \mid \theta\right) = \prod_{i=1}^d q_i\left(\theta'_i \mid \theta\right)$ where $q_i\left(\theta'_i \mid \theta\right)$ is a simple categorical distribution:

$$\operatorname{Cat}\left(\sigma\left(\frac{1}{2}\Delta[U](\theta)_i\left(\theta'_i - \theta_i\right) - \frac{\left(\theta'_i - \theta_i\right)^2}{2\alpha}\right)\right). \tag{9}$$

Here, Cat stands for categorical distribution, σ denotes SoftMax function and $\theta'_i \in \Theta_i$. Recall that $\Delta[U](\theta)_i = U\left(\neg_i\theta\right) - U(\theta)$, $\forall i \in \{1, \ldots, d\}$ where $\neg_i\theta$ changes the i-th coordinate from θ_i to $\theta_i + 1$ while keeping the other coordinates the same as θ.

Since both the domain Θ and the proposal over all coordinates in (8) can be factorized coordinatewisely, we can update each coordinate in parallel, thus greatly speeding up the computation. In this way, Newton proposal fills in the blank in bottom right of Table 1. It enables us to sample from high-dimensional complex discrete distributions with better mixing and faster convergence, no matter whether the discrete distribution has an natural differentiable extension.

When modeling the proposal distribution over all coordinates jointly, the overall cost of constructing the proposal in (8) is $\mathcal{O}\left(L^d\right)$ for $\{0, 1, \ldots, L-1\}^d$. Thanks to the coordinatewise factorization, the cost of Newton proposal is reduced to $\mathcal{O}(Ld)$. This allows the sampler to explore the full space with the neighborhood information without paying a prohibitive cost.

4.3 A Variant: With a MH Correction

It is optional to add a Metropolis-Hastings (MH) step [Metropolis et al., 1953, Zhang et al., 2022a], which is usually combined with proposals to make the Markov chain reversible. Specifically, after generating the next position θ' from a distribution $q(\cdot \mid \theta)$, the MH step accepts it with probability

$$\min\left(1, \exp\left(U\left(\theta'\right)-U(\theta)\right) \frac{q\left(\theta \mid \theta'\right)}{q\left(\theta' \mid \theta\right)}\right). \tag{10}$$

By rejecting some of the proposed states, the Markov chain is guaranteed to converge asymptotically to the target distribution. The sampler with our Newton proposal is outlined in Algorithm 1.

We call Newton proposal without the MH step as unadjusted Newton algorithm (**UNA**) and that with the MH step as Metropolis-adjusted Newton algorithm (**MANA**). Similar to MALA and ULA in continuous spaces [Grenander and Miller, 1994, Roberts and Stramer, 2002], MANA contains $2Ld$ (for finite difference computation) plus 2 (for the MH correction) function evaluations and is guaranteed to converge to the target distribution. Although UNA may have asymptotic bias, it only requires Ld function evaluations, which is valuable especially when performing the function evaluation is expensive such as in large-scale Bayesian inference [Welling and Teh, 2011, Durmus and Moulines, 2019]

Algorithm 1 Samplers with Newton Proposal.

```
given: Stepsize $\alpha$.
loop
    for $i = 1, \ldots, d$ do
        (Can be done in parallel)
        construct $q_i(\cdot \mid \theta)$ as in Equation (9)
        sample $\theta_i' \sim q_i(\cdot \mid \theta)$
    end for
    ▷ Optionally, do the MH step
    compute $q\left(\theta' \mid \theta\right)=\prod_i q_i\left(\theta_i' \mid \theta\right)$
        and $q\left(\theta \mid \theta'\right)=\prod_i q_i\left(\theta_i \mid \theta'\right)$
    set $\theta \leftarrow \theta'$ with probability in Equation (10)
end loop
output: samples $\{\theta_k\}$.
```

5 AN ALTERNATIVE VIEW OF NEWTON PROPOSAL

After getting the Newton proposal via Newton's series, we give an alternative way to derive Newton proposal via multilinear extension, which gives us more intuition about the efficient informed proposal. From the multilinear extension viewpoint, we find further connection with Discrete Langevin Proposal (DLP) [Zhang et al., 2022a].

5.1 An Equivalent Form via Multilinear Extension

In addition to approximating the likelihood ratio of flipping each dimension with Newton's series expansion, we find that our Newton proposal can also be obtained by conducting Taylor expansion on the multilinear extension of the discrete distribution. This connection gives us another interesting viewpoint of the Newton proposal. We briefly show the equivalence between these two viewpoints in the binary case here and put the categorical case and detailed proof in the appendix.

Let us consider the d-dimensional binary distribution. The coordinates can be denoted as a finite set $D = \{1, \cdots, d\}$. The sampling process corresponds to choosing which coordinate to flip, so the discrete distribution is a set function defined over the power set of D as $f: 2^D \rightarrow \mathbb{R}$. A discrete distribution may not have a natural continuous extension, but its multilinear extension $F: [0,1]^d \rightarrow \mathbb{R}$ can always be defined as :

$$F(\theta)=\sum_{S \subseteq D} f(S) \prod_{i \in S} \theta_i \prod_{i \in D \backslash S}\left(1-\theta_i\right). \tag{11}$$

As we can see, $F(\theta)$ is a continuous, differentiable function which accepts real-valued inputs from the interval $[0,1]^d$. This makes it possible to approximate the likelihood of discrete moves with Taylor series expansion. Besides, an inspiring fact is that $F(\theta)$ keeps the same value with f when they are evaluated on the discrete subset $\{0,1\}^d$. For $i \in D$, since F is linear in θ_i, we have the partial derivative of $F(\theta)$ as:

$$\begin{aligned} \frac{\partial F}{\partial \theta_i}(\theta) =& F\left(\theta_1, \ldots, \theta_{i-1}, 1, \theta_{i+1}, \ldots, \theta_d\right) \\ & -F\left(\theta_1, \ldots, \theta_{i-1}, 0, \theta_{i+1}, \ldots, \theta_d\right) \end{aligned}$$

We can see that the partial derivative measures the difference between the energy function of the original state and the flipped one, which is exactly the finite difference of the two states. Since F and f take the same value on the discrete domain, the Newton proposal can be equivalently obtained by conducting Taylor expansion on the multilinear extension of the target discrete distribution.

As for categorical distribution, we need to decide not only which coordinate to flip, but also which level to flip to. Fortunately, the differentiable function $F(\theta)$ can be obtained with a generalized multilinear extension [Sahin et al., 2020], on which the likelihood ratio of flipping each coordinate to any level can be defined. The detailed algorithm is in the appendix.

5.2 Comparison with Discrete Langevin Proposal

Our Newton proposal and the Discrete Langevin Proposal (DLP) [Zhang et al., 2022a] can be seen as two different approximations of the locally-balanced proposal [Zanella, 2020] when taking functions g and K like MALA, as shown in (3).

DLP is motivated by utilizing gradients to guide the sampling and inference. Thus it requires that the discrete distribution can be displayed as a differentiable function which is only evaluated on a discrete domain, $e.g.$, Ising models, so that the gradient can be defined. In this way, DLP can be viewed as a first-order Taylor series approximation to the local difference term inside $q_0(\theta' \mid \theta)$ in (3) with:

$$U(x) - U(\theta) \approx \nabla U(\theta)^\top (x - \theta),\ \forall x \in \Theta.$$

In contrast, our Newton proposal circumvents this shortcoming by using Newton's series expansion, a natural tool to approximate discrete functions. Specifically, Newton proposal approximates the local difference in $q_0(\theta' \mid \theta)$ with:

$$U(x) - U(\theta) \approx \Delta[U](\theta)^\top (x - \theta),\ \forall x \in \Theta.$$

The finite difference $\Delta[U](\theta)$ is defined on the grid-like discrete domain and can also guide to explore the discrete space like what gradients do in continuous relaxation-based methods.

In addition to the differences in methods and requirements, our Newton proposal has more general applications than DLP: (1) When the discrete distribution has a natural differential extension, such as Ising model, Restricted Boltzmann Machines (RBMs) and Potts model, DLP and the Newton proposal both work. In some special cases such as some Ising models, when the multilinear extension and the natural continuous extension of the original discrete distribution are the same, DLP and Newton proposal have the same results. (2) The strict requirement about natural differential relaxation limits DLP to be widely used in more complex scenarios. When the target discrete distribution lacks a differentiable extension, DLP does not work while the Newton proposal is still well-suited for these tasks, such as the facility location problem, text summarization, etc.

For the computation cost, DLP contains one gradient computation whose cost depends on d and L, while the Newton proposal contains Ld function evaluations. Both gradient and function computations can be done in parallel.

6 THEORETICAL ANALYSIS

In this section, we analyze the asymptotic convergence of UNA and the asymptotic efficiency of MANA. Specifically, we first prove in Section 6.1 that when the stepsize $\alpha \to 0$, the asymptotic bias of UNA is zero when the discrete distribution is a second-order modular function, whose gain reduction keeps the same for any set [Korula et al., 2018] (see the rigorous definition in the appendix). Later in Section 6.2, we derive the asymptotic efficiency of MANA.

6.1 Asymptotic Convergence of UNA for Second-order Modular Functions

Besides the property of the proposal itself, the effectiveness of a proposal also depends on how close its underlying stationary distribution is to the target distribution because if it is far, even if using the MH step to correct the bias, the acceptance probability will be very low. We consider a second-order modular distribution, which appears in common tasks such as Ising models. The following theorem summarizes UNA's asymptotic accuracy for such discrete distributions.

Theorem 1. When the discrete distribution is a second-order modular function [Korula et al., 2018], its multilinear extension $F(\theta)$ is quadratic which can be expressed as $F(\theta) = \theta^\top A\theta + b^\top \theta$. The Markov chain following transition $q(\cdot \mid \theta)$ in (9) (i.e. UNA) is reversible with respect to some distribution π_α and π_α converges weakly to π as $\alpha \to 0$. In particular, let λ_{min} be the smallest eigenvalue of A, then for any $\alpha > 0$,

$$\|\pi_\alpha - \pi\|_1 \leq Z \cdot \exp\left(-\frac{1}{2\alpha} - \frac{\lambda_{\text{min}}}{2}\right).$$

Theorem 1 shows that the asymptotic bias of UNA decreases at a $\mathcal{O}(\exp(-1/(2\alpha))$ rate which vanishes to zero as the stepsize $\alpha \to 0$. Besides, the asymptotic bias of UNA is related to the smallest eigenvalue of A.

Example. Let's consider the well-known model in thermodynamic systems, the Ising model. Note that its distribution is second-order modular:

$$f(D) = \sum_{u,v \in D} A(u,v) + \sum_{u \in D} b(u)$$

where $A(u,v)$ represents the interaction between any two adjacent sites $u,v \in D$, and a site $u \in D$ has an external magnetic field $b(u)$ interacting with it. We run UNA with varying stepsizes on a 2 by 2 Ising model, as shown in Figure 1. For each stepsize, we run the chain long enough to ensure its convergence. The results clearly show that the distance between the stationary distribution of UNA and the target distribution decreases as the stepsize decreases.

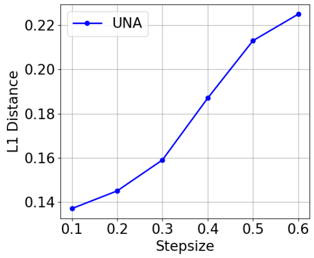


Figure 1: UNA with varying stepsizes on an Ising model.

6.2 Asymptotic Efficiency of MANA

To understand the asymptotic efficiency of MCMC transition kernels, we study the asymptotic variance and the spectral gap of the kernel. The asymptotic variance is defined as

$$\operatorname{var}_p(h, Q) = \lim_{T\to\infty} \frac{1}{T} \operatorname{var}\left(\sum_{t=1}^{T} h\left(x_t\right)\right)$$

where $h : \mathcal{X} \to R$ is a scalar-valued function, Q is a p-stationary Markov transition kernel. The asymptotic variances measures the additional variance incurred when using sequential samples from Q to estimate $E_p[h(x)]$. The spectral gap is defined as

$$\operatorname{Gap}(Q) = 1 - \lambda_2$$

where λ_2 is the second largest eigenvalue of the transition probability matrix of Q. For transition probability matrices with non-negative eigenvalues, the spectral gap is related to the mixing time, with larger values corresponding to faster mixing [Levin and Peres, 2017].

Since our method approximates $Q_{g,\sigma}(\theta, d\theta')$ in (2), we should expect some decrease in efficiency. We characterize this decrease in terms of the asymptotic variance and the spectral gap, under the Lipschitz-like assumption on $\Delta[U](\theta)$. In particular, we show that the decrease is a constant factor that depends on the Lipschitz constant of $\Delta[U](\theta)$ and the dimension of the target distribution.

Theorem 2 Let $Q\left(\theta', \theta\right)$ and $\tilde{Q}\left(\theta', \theta\right)$ be the Markov transition kernels given by the Metropolis-Hastings algorithm using the locally-balanced proposal $q_0\left(\theta' \mid \theta\right)$ and our approximation $q\left(\theta' \mid \theta\right)$. Let the finite difference $\Delta[U](\theta)$ has an Lipschitz-like property with constant L, and $\pi(\theta) = \frac{\exp(U(\theta))}{Z}$. Then it holds

1. $\operatorname{var}_\pi\left(h, \tilde{Q}\right) \leq \frac{\operatorname{var}_\pi(h,Q)}{c} + \frac{1-c}{c} \cdot \operatorname{var}_\pi(h)$.
2. $\operatorname{Gap}\left(\tilde{Q}\right) \geq c \cdot \operatorname{Gap}(Q)$

where $c = e^{-\frac{1}{2}LD^2}$ and $D = \sup_{\theta' \in \Theta} \|\theta' - \theta\|$.

Remark. We can see that the constant D is correlated with the dimension of the target discrete distribution. In high-dimensional scenarios, D will be a large constant, leading to loose bounds of the asymptotic variance and spectral gap. However, on one hand, there is a gap between theory and experiment [Kwisthout and Van Rooij, 2013]. It may be hard to achieve the bound in practice. On the other hand, we can add a slight restriction on the number of changed coordinates in a single step. In this way, D can be reduced to a small number as we expect, and we can get tighter bounds in theory.

7 EXPERIMENTS

We conduct a comprehensive empirical evaluation for Newton proposal on synthetic and real-world sampling tasks. The unadjusted and Metropolis-adjusted Newton proposals are denoted as UNA and MANA, respectively. We release the code at https://github.com/DongyaoZhu/Newton-Proposal-for-Discrete-Sampling. Baselines and evaluation tasks are described below.

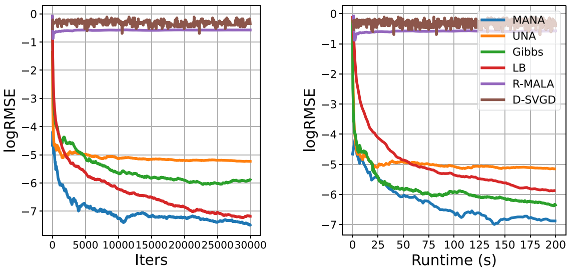
Baselines. We compare the performance of Newton proposals with widely-used sampling methods for discrete distributions, including (1) two Gibbs-based methods: Gibbs sampling, Gibbs with Gradient (GWG) [Grathwohl et al., 2021]; (2) two methods which perform sampling in a continuous space by gradient-based methods and then transforms the collected samples to the original discrete space: discrete Stein Variational Gradient Descent (D-SVGD) [Han et al., 2020] and relaxed MALA (R-MALA) [Grathwohl et al., 2021]; (3) one gradient-based method which requires the target discrete distribution to have a differential relaxation: Discrete Langevin Proposal (DLP) [Zhang et al., 2022a] and (4) the locally-balanced sampler (LB) [Zanella, 2020]. Specifically, we denote DULA and DMALA for unadjusted and Metropolis-adjusted DLPs, respectively. All methods are implemented in Pytorch and we use the official release of code from previous papers when possible.

Evaluation Tasks. (1) Discrete distributions without natural differentiable extensions. Since GWG and DLP require gradients and can not be applied, we mainly compare the Newton proposal with Gibbs and LB in these tasks. (2) Discrete distributions with natural differentiable extensions. We also apply Newton proposal to these distributions such as the Ising model which is binary and Potts model which is categorical, to show the broad applicability of our method. In this scenario, we also compare Newton proposal with continuous relaxation methods (GWG and DLP) and the results are included in the appendix.

7.1 Facility Location

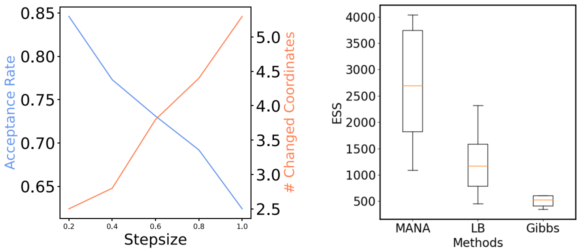
In the facility location task, we are given a set of facilities denoted $\mathcal{V}$ and a set of m customers to decide whether to open a facility or not [Krause et al., 2008], corresponding to sampling from a binary distribution. If the i-th facility $(i \in \mathcal{V})$ is opened, then it provides service of value $c_{i,j}$ to customer j $(j \in \{1, \cdots, m\})$. We suppose that each customer chooses the opened facility with highest value, then the total value provided to all customers is $\sum_{j=1}^{m} \max_{i\in\mathcal{V}} c_{i,j}$. Besides, we penalize the number of selected facilities to ensure the most total utility with a small number of facilities. Therefore, the distribution of the facility location model can be represented as $f(S) = \sum_{j=1}^{m} \max_{i\in\mathcal{V}} c_{i,j} - \lambda|\mathcal{V}|$, where λ is a hyperparameter, controlling the strength of the penalty term. To evaluate the Newton proposal on facility location task, we generate the utility matrix $\boldsymbol{C}$ from a gaussian mixture model with $m = 64$ and $|\mathcal{V}| = 15$. We run 30000 iterations and set the stepsizes of MANA and UNA as 1 and 0.2, respectively.

We first compare the root-mean-square error (RMSE) between the estimated mean and the true mean under $\lambda = 10$ in Figure 2. The blue lines of MANA are both below other lines, indicating that MANA is the fastest to converge in terms of both iterations and running time. This demonstrates (1) sampling in the original discrete space is important: D-SVGD and R-MALA get poor results because this task is complex and the relaxed distributions are hard to sample from; (2) the finite difference makes the exploration over the discrete space more informative rather than "blind" compared to Gibbs; (3) the MH step enables MANA to make larger and more effective moves with a larger step size without worrying about unconvergence compared to UNA; (4) changing many coordinates in one step accelerates the convergence compared to LB and Gibbs. We then show that Newton proposal can change multiple coordinates in one iteration while still maintaining a high acceptance rate in Figure 3(a). When the stepsize $\alpha = 1$, on average MANA can change **5.3** coordinates in one iteration with an acceptance rate 62.4% in the MH step, while the acceptance rate of LB proposal is only 35%. In Figure 3(b), we compare the effective sample size (ESS) for exact samplers (*i.e.*, having the target distribution as its stationary distribution). MANA outperforms other methods, indicating the correlation among its samples is low due to making significant updates in each step.



(a) log RMSE w.r.t. Iterations (b) log RMSE w.r.t. Runtime

Figure 2: Facility location model results. MANA outperforms the baselines in both number of iterations and running time.



(a) AccRate, #Changed Dims w.r.t. Stepsize (b) ESS of Proposals

Figure 3: Facility location sampling results. **Left:** Newton proposal keeps a higher acceptance rate. **Right:** MANA yields the largest effective sample size (ESS) among all the methods compared.

7.2 Extractive Text Summarization

Extractive text summaries are formed by selecting several sentences S from source documents that best fit certain quality measurements. [Lin and Bilmes, 2011] designed their metrics on S for both similarity and diversity, formally defined as $\mathcal{F}(S) = \mathcal{L}(S) + \lambda\mathcal{R}(S)$ subject to some cost constraint $C(S) \leq K$, where $\mathcal{L}(S)$ measures the coverage or "fidelity" of summary set S to the document, $\mathcal{R}(S)$ rewards diversity in S, and $\lambda > 0$ is a trade-off coefficient. The undifferentiable distribution of the summary $f(S) = \max \mathcal{F}(S)$ makes the task well-suited for our methods based on pseudo-gradients. Futhermore, due to the limited time constraint, the proposals with non-parallel updates cannot explore enough into the distribution, while our Newton proposal will be able to avoid this issue with multidimensional updates. We use an exponential decay schedule on step size to encourage faster convergence under limited time constraint. The final result is then given by a sample-wise majority vote algorithm [Wang et al., 2022] on the collection of samples we produce.

We report results of MANA, LB and Gibbs sampler on DUC 2002 dataset [Over and Liggett, 2002], which contains about 30 sentences per document. ROUGE scores

Table 2: Performance of sampling methods on Extractive Text Summarization on DUC-2002 dataset. We report $\mathcal{F}(S)$ at 500, 750 and 1000 steps, as well as ESS, runtime of 1000 steps (s), the ROUGE-2 recall (R), F-measure (F) and Precision (P) (%).

Method	$\mathcal{F}(500)$	$\mathcal{F}(750)$	$\mathcal{F}(1000)$	R	F	P	ESS	Runtime
MANA	**6.28**	**6.40**	**6.46**	**8.72**	**8.85**	**10.96**	**57.2**	6.7
LB	6.28	6.39	6.42	7.91	8.40	10.68	41.8	9.9
Gibbs	6.01	6.12	6.42	7.91	8.31	10.62	15.2	**1.0**

Table 3: Performance of sampling methods on Image Retrieval on Holidays dataset. $\mathcal{F}(S)$ and mean Average Precision (mAP) are reported at 500, 750, 1000 steps.

Method	$\mathcal{F}(500)$	$\mathcal{F}(750)$	$\mathcal{F}(1000)$	mAP
MANA	**11.36**	**11.37**	**11.38**	**0.84**
LB	11.04	11.05	11.05	0.55
Gibbs	11.01	11.01	11.03	0.53

[Lin, 2004] are widely used for text summarization evaluation, and we compare ROUGE-2 scores (precision P, recall R, and F-measure F) of different samplers. In addition, we show the average objective scores $\mathcal{F}(S)$ at 500th, 750th and 1000th iteration, respectively. As shown in Table 2, Newton proposal constantly outputs highest $\mathcal{F}(S)$ and ROUGE-2 scores under limited time constraints.

7.3 Image Retrieval

Given a database of images and a query image, we look for a subset from the database that best matches the query. We follow the same settings in the extractive text summarization experiments, and we use the discontinuous score function $\mathcal{F}(S)$ proposed by [Yang et al., 2014] to measure the matching of a particular collection of images to a query image. We empirically found that sample-wise majority vote algorithm [Wang et al., 2022] did not perform well, thus we also propose a dimension-wise majority vote algorithm: given a collection of N samples $X \in \{0,1\}^{N,D}$, each dimension d will have a count of $x_d = 1$, and the dimensions of top counts are selected (details in appendix).

We evaluate various methods on the INRIA Holidays Dataset [Jegou et al., 2008] which consists of 1491 images (dimensions) and 500 queries. Our results in $\mathcal{F}(S)$ values and mean Average Precision (mAP) are reported in Table 3. The $\mathcal{F}(S)$ shows that the sampler with Newton proposal quickly reaches and stably keeps a better solution to the maximization problem than other methods despite limited time constraint. Our high mean Average Precision demonstrates that our solution is a better approximation to the ground truth labels compared to other methods.

8 CONCLUSION

We propose a new gradient-like efficient informed proposal, the Newton proposal, for general discrete distributions. This proposal better explores discrete spaces under the guidance of the finite difference produced by Newton's series expansion, which does not require natural differentiable expansions. Additionally, the factorization on coordinates allows multiple coordinates to be updated simultaneously, leading to a faster convergence rate. To the best of our knowledge, Newton proposal makes the first attempt to utilize Newton's series expansion and multilinear extension in discrete sampling, which fills the gap of efficient sampling for complex discrete distributions when gradients are not available. For different application scenarios, we develop several variants with Newton proposal, including unadjusted and Metropolis-adjusted versions. We theoretically prove the convergence and efficiency of Newton proposal without and with the MH step. Empirical results on various problems demonstrate the superiority of our method over baselines in general settings.

Acknowledgments

We would like to thank Yingzhen Li for helpful discussions and the anonymous reviewers for their thoughtful comments on the manuscript.

References

Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

Jianrong Wang, Ahsan Huda, Victoria V Lunyak, and I King Jordan. A gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics*, 26(20):2501–2508, 2010.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Jorn WT Peters and Max Welling. Probabilistic binary neural networks. *arXiv preprint arXiv:1809.03368*, 2018.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Ruqi Zhang, Xingchao Liu, and Qiang Liu. A langevin-like sampler for discrete distributions. In *International Conference on Machine Learning*, pages 26375–26396. PMLR, 2022a.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning*, pages 3831–3841. PMLR, 2021.

Giacomo Zanella. Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

Ari Pakman and Liam Paninski. Auxiliary-variable exact hamiltonian monte carlo samplers for binary distributions. *Advances in neural information processing systems*, 26, 2013.

Akihiko Nishimura, David B Dunson, and Jianfeng Lu. Discontinuous hamiltonian monte carlo for discrete parameters and discontinuous likelihoods. *Biometrika*, 107(2):365–380, 2020.

Jun Han, Fan Ding, Xianglong Liu, Lorenzo Torresani, Jian Peng, and Qiang Liu. Stein variational inference for discrete distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 4563–4572. PMLR, 2020.

Guangyao Zhou. Mixed hamiltonian monte carlo for mixed discrete and continuous variables. *Advances in Neural Information Processing Systems*, 33:17094–17104, 2020.

Priyank Jaini, Didrik Nielsen, and Max Welling. Sampling in combinatorial spaces with survae flow augmented mcmc. In *International Conference on Artificial Intelligence and Statistics*, pages 3349–3357. PMLR, 2021.

Dinghuai Zhang, Nikolay Malkin, Zhen Liu, Alexandra Volokhova, Aaron Courville, and Yoshua Bengio. Generative flow networks for discrete probabilistic modeling. *arXiv preprint arXiv:2202.01361*, 2022b.

Samuel Power and Jacob Vorstrup Goldman. Accelerated sampling on discrete spaces with non-reversible markov processes. *arXiv preprint arXiv:1912.04681*, 2019.

Emanuele Sansone. Lsb: Local self-balancing mcmc in discrete spaces. In *International Conference on Machine Learning*, pages 19205–19220. PMLR, 2022.

Philippe Gagnon, Florian Maire, and Giacomo Zanella. Improving multiple-try metropolis with local balancing. *arXiv preprint arXiv:2211.11613*, 2022.

Haoran Sun, Hanjun Dai, Wei Xia, and Arun Ramamurthy. Path auxiliary proposal for mcmc in discrete space. In *International Conference on Learning Representations*, 2021.

Haoran Sun, Hanjun Dai, and Dale Schuurmans. Optimal scaling for locally balanced proposals in discrete spaces. *arXiv preprint arXiv:2209.08183*, 2022.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6): 1087–1092, 1953.

Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4): 549–581, 1994.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Aytunc Sahin, Yatao Bian, Joachim Buhmann, and Andreas Krause. From sets to multisets: provable variational inference for probabilistic integer submodular models. In *International Conference on Machine Learning*, pages 8388–8397. PMLR, 2020.

Nitish Korula, Vahab Mirrokni, and Morteza Zadimoghaddam. Online submodular welfare maximization: Greedy beats 1/2 in random order. *SIAM Journal on Computing*, 47(3):1056–1086, 2018.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Johan Kwisthout and Iris Van Rooij. Bridging the gap between theory and practice of approximate bayesian inference. *Cognitive Systems Research*, 24:2–8, 2013.

Andreas Krause, Jure Leskovec, Carlos Guestrin, Jeanne VanBriesen, and Christos Faloutsos. Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management*, 134(6):516–526, 2008.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 510–520, 2011.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Paul Over and Walter Liggett. Introduction to duc-2002: an intrinsic evaluation of generic news text. In *Document Understanding Conference*, 2002.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Fan Yang, Zhuolin Jiang, and Larry S Davis. Submodular reranking with multiple feature modalities for image retrieval. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2014.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer, 2008.

Efficient Informed Proposals for Discrete Distributions via Newton's Series Approximation: Supplementary Materials

1 Detailed Derivation of Newton Proposal

We give more details about the derivation of our Newton proposal.

We start from the MALA-like locally-balanced proposal [Zanella, 2020] we mentioned in Section 3:

$$q_0\left(\theta' \mid \theta\right) = \frac{1}{Z_\Theta(\theta)} \exp\left(\frac{U(\theta') - U(\theta)}{2} - \frac{\|\theta' - \theta\|^2}{2\alpha}\right) \tag{1}$$

where $Z_\Theta(\theta)$ is the normalizing constant. When we use Newton's series expansion to approximate the local difference $U(\theta') - U(\theta)$, we get

$$\begin{aligned} q\left(\theta' \mid \theta\right) = \widehat{q_0}\left(\theta' \mid \theta\right) &= \frac{1}{Z_\Theta(\theta)} \exp\left(\frac{\Delta_h[U](\theta)^\top \cdot (\theta' - \theta)}{2} - \frac{\|\theta' - \theta\|^2}{2\alpha}\right) \\ &\propto \exp\left(-\frac{1}{2\alpha}\left((\theta' - \theta)^2 - \alpha\Delta_h[U](\theta)^\top \cdot (\theta' - \theta) + \frac{\alpha^2}{4}\Delta_h[U](\theta)^2\right)\right) \\ &= \exp\left(-\frac{1}{2\alpha}\|\theta' - \theta - \frac{\alpha}{2}\Delta_h[U](\theta)\|^2\right) \end{aligned} \tag{2}$$

where the second line is because $\Delta_h[U](\theta)$ is independent of θ' and will not affect the normalized result.

Since (2) is actually an ℓ-2 norm, $q\left(\theta' \mid \theta\right)$ can be factorized coordinatewisely. Besides, we assume the domain can be factorized coordinatewisely. Therefore, we can factorize $q\left(\theta' \mid \theta\right)$ in (2) as $q\left(\theta' \mid \theta\right) = \prod_{i=1}^d q_i\left(\theta_i' \mid \theta\right)$ and

$$\begin{aligned} q_i\left(\theta_i' \mid \theta\right) &= \exp\left(-\frac{1}{2\alpha}(\theta_i' - \theta_i - \frac{\alpha}{2}\Delta_h[U](\theta))^2\right) \\ &\propto \exp\left(\frac{1}{2}\Delta_h[U](\theta_i)(\theta_i' - \theta_i) - \frac{(\theta_i' - \theta_i)^2}{2\alpha}\right) \\ &= \text{Softmax}\left(\frac{1}{2}\Delta_h[U](\theta_i)(\theta_i' - \theta_i) - \frac{(\theta_i' - \theta_i)^2}{2\alpha}\right). \end{aligned} \tag{3}$$

Then we get our Newton proposal which is easy to compute in parallel:

$$\text{Categorical}\left(\text{Softmax}\left(\frac{1}{2}\Delta_h[U](\theta_i)\left(\theta_i' - \theta_i\right) - \frac{\left(\theta_i' - \theta_i\right)^2}{2\alpha}\right)\right). \tag{4}$$

2 Algorithm for Binary Variables

When the variable domain Θ is binary $\{0, 1\}^d$, if we flip any coordinate θ_i to θ_i', $(\theta_i' - \theta_i)^2$ is always 1. Thanks to the coordinatewise factorization, the sample space only contains 2 states: flipping or remaining the original state, which makes the normalizing constant $Z_\Theta(\theta)$ tractable. In this way, we could simplify Algorithm 1 in the main body of our paper further and obtain the following algorithm, which clearly shows that our method can be cheaply computed in parallel on CPUs and GPUs. We give the pseudo code when sampling from binary distributions with Newton proposal as follows.

Algorithm 1 Samplers with Newton Proposal on Binary Domains.

```
Input: Stepsize $\alpha$.
loop
    Compute $P(\theta) = \frac{\exp\left(-\frac{1}{2}\Delta_h[U](\theta)\odot(2\theta-1)-\frac{1}{2\alpha}\right)}{\exp\left(-\frac{1}{2}\Delta_h[U](\theta)\odot(2\theta-1)-\frac{1}{2\alpha}\right)+1}$
    sample $\mu \sim \text{Unif}(0,1)^d$
    $I \leftarrow \dim(\mu \leq P(\theta))$
    $\theta' \leftarrow \text{flipdim}(I)$
    ▷ Optionally, do the MH step
    compute $q(\theta' \mid \theta) = \prod_i q_i(\theta'_i \mid \theta) = \prod_{i \in I} P(\theta)_i \cdot \prod_{i \notin I}(1 - P(\theta)_i)$
    compute $P(\theta') = \frac{\exp\left(-\frac{1}{2}\Delta_h[U](\theta')\odot(2\theta'-1)-\frac{1}{2\alpha}\right)}{\exp\left(-\frac{1}{2}\Delta_h[U](\theta')\odot(2\theta'-1)-\frac{1}{2\alpha}\right)+1}$
    compute $q(\theta \mid \theta') = \prod_i q_i(\theta_i \mid \theta') = \prod_{i \in I} P(\theta')_i \cdot \prod_{i \notin I}(1 - P(\theta')_i)$
    set $\theta \leftarrow \theta'$ with probability
        $$\min\left(1, \exp\left(U(\theta') - U(\theta)\right)\frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}\right)$$
end loop
Output: samples $\{\theta_k\}$.
```

3 Newton Proposal for Categorical Variables

3.1 Method 1: Newton's Series Approximation

When using one-hot vectors to represent categorical variables, our Newton proposal becomes

$$\text{Categorical}\left(\text{Softmax}\left(\frac{1}{2}\Delta_h[U](\theta_i)^\top(\theta'_i - \theta_i) - \frac{\|\theta'_i - \theta_i\|_2^2}{2\alpha}\right)\right), \tag{5}$$

where θ_i, θ'_i are one-hot vectors.

If the variables are ordinal with clear ordering information, we can also use integer representation $\theta \in \{0, 1, \ldots, L-1\}^d$ and compute the Newton proposal as in Equation (4).

3.2 Method 2: Multilinear Extension

As we stated in Section 5.1, for binary variables, our Newton proposal obtained from the first-order Newton's series approximation is equivalent to that obtained from the first-order Taylor series approximation to the multilinear extension of the original discrete target distribution. For categorical variables, we not only need to decide which coordinate to flip, but also need to decide the level. By introducing the concepts of 'multiset', we generalize the multilinear extension to categorical distributions and thus extend our Newton proposal to categorical variables.

3.2.1 Multiset

In classical sets, distinct elements can only occur once. A multiset is a natural generalization of a set, where elements can be contained repeatedly. The number of times an element occurs is called the multiplicity $\mu(i)$ of the element i. A multiset $\mathcal{M}_\mathcal{V}$ is defined as a pair $\langle\mathcal{V}, \mu\rangle$, where $\mathcal{V}$ is the support and $\mu : \mathcal{V} \to \mathbb{N}$ is a function defining multiplicity for each element [Sahin et al., 2020]. Given this definition, we can use the integer vector of the multiset's multiplicity to represent any multiset. Also, we can transfer several important notions from multisets to integer vectors, such as the notion of a subset, set intersection, set union and set difference.

Now consider sampling from a d-dimensional discrete distribution. The support can be represented as $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$, where $\mathcal{X}_i = \{0, 1, \ldots, L_i - 1\}$ and the discrete function f is an integer function defined as $f \colon \mathcal{X} \to \mathbb{N}$. For ease of notation, we assume that L_i does not depend on i and $\mathcal{X}_i$ is the same along each dimension, *i.e.*, $\mathcal{X}_i = \{0, 1, \ldots, L-1\}$,

$\forall i \in 1, \ldots, d$. The obtained sample will be an integer vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ ($x_i \in \mathcal{X}_i$, $\forall i \in 1, \ldots, n$) and can be equivalently represented as a multiset. Note that the integer x_i ($\forall i \in 1, \ldots, n$) can be also represented as a binary vector $\boldsymbol{x}_i$ of length $L-1$. Take the $k=3$ case as an example. $\boldsymbol{x}_i$ can be $(0,0)^\top$, $(1,0)^\top$ or $(0,1)^\top$, corresponding to the level of 0, 1, 2, respectively. For the simplicity in calculation, we will use the binary representation in the following parts.

3.2.2 Generalized Multilinear Extension

Given the discrete distribution $f(\theta)$, generalized multilinear extension will extend f to a continuous domain while keeping the values on original discrete domain.

Let $\boldsymbol{\rho}_i \in \mathbb{R}_+^{L-1}$ be the marginals of a d-dimensional categorical distribution and $\boldsymbol{\rho} := [\boldsymbol{\rho}_1; \ldots; \boldsymbol{\rho}_d] \in \{0,1\}^{(L-1)\times d}$ is the concatenation of all $\boldsymbol{\rho}_i$ vectors. Each $\boldsymbol{\rho}_i$ lives in the $L-1$ dimensional simplex Δ^{L-1}. The simplex Δ^{L-1} is defined as

$$\begin{aligned}\Delta^{L-1} := \big\{\rho_i \in \mathbb{R}^{L-1} : \rho_{i,1} + \ldots + \rho_{i,L-1} \leq 1 \\ \rho_{ij} \geq 0, j = 1, \ldots, L-1\big\}.\end{aligned}$$

We define the union of n simplexes as Δ_n^{L-1}, and naturally $\boldsymbol{\rho} \in \Delta_d^{L-1}$. Once we sample from $\boldsymbol{\rho}$, we get $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d) \in \{0,1\}^{(L-1)\times d}$. The generalized multilinear extension F is defined on the space of product of categorical distributions and can be written as:

$$F(\boldsymbol{\rho}) = \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_d}[f(\boldsymbol{\theta})]. \tag{6}$$

We need to compute the sum of L^d elements to compute the expectation in Equation (6). Note that when $L=2$, this extension corresponds to the multilinear extension of a set function. Here is an example with $d=2$ and $L=3$. In this case we have two categorical distributions which take three different values.

$$\begin{aligned}F(\boldsymbol{\rho}) =& F([\boldsymbol{\rho}_1; \boldsymbol{\rho}_2]) = F(\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22}) \\ =& f(\begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix})(1-\rho_{11}-\rho_{12})(1-\rho_{21}-\rho_{22}) + f(\begin{pmatrix}0 & 0\\ 1 & 1\end{pmatrix})\rho_{12}\rho_{22} + f(\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix})\rho_{11}(1-\rho_{21}-\rho_{22}) \\ &+ f(\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix})(1-\rho_{11}-\rho_{12})\rho_{21} + f(\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix})\rho_{12}(1-\rho_{21}-\rho_{22}) + f(\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix})(1-\rho_{11}-\rho_{12})\rho_{22} \\ &+ f(\begin{pmatrix}1 & 1\\ 0 & 0\end{pmatrix})\rho_{11}\rho_{21} + f(\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix})\rho_{11}\rho_{22} + f(\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix})\rho_{12}\rho_{21},\end{aligned}$$

where we have the following constraints

$$\rho_{11} + \rho_{12} \leq 1, \rho_{21} + \rho_{22} \leq 1, \rho_{ij} \geq 0, i, j = 1, 2.$$

3.2.3 Newton Proposal via Generalized Multilinear Extension

Similar to the multilinear extension for the binary domain, we can calculate the first-order partial derivative of the generalized multilinear extension of f, *i.e.*, $F(\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_d)$. Given $\boldsymbol{\rho} \in \Delta_d^{L-1}$, let $\mathcal{R}_\mathcal{V}$ be a random multiset where elements appear independently with probabilities $\boldsymbol{\rho}_i$. Since F is multilinear, the partial derivative can be written as a difference of two generalized multilinear extensions:

$$\begin{aligned}\frac{\partial F}{\partial \rho_{ij}} &= F(\boldsymbol{\rho}_1, \boldsymbol{\rho}_i = \boldsymbol{e}_j, \boldsymbol{\rho}_n) - F(\boldsymbol{\rho}_1, \boldsymbol{\rho}_i = \boldsymbol{0}, \boldsymbol{\rho}_n) \\ &= \mathbb{E}_{\mathcal{R}_\mathcal{V}\sim\rho}\left[f\left(\mathcal{R}_\mathcal{V} \cup \mathcal{E}_i^j\right)\right] - \mathbb{E}_{\mathcal{R}_\mathcal{V}\sim\rho}\left[f\left(\mathcal{R}_\mathcal{V} \setminus \mathcal{E}_i^j\right)\right]\end{aligned}$$

where $\cup$ and $\setminus$ corresponds to union and set difference between multisets, respectively. $\boldsymbol{e}_j \in \Delta_n^{k-1}$ is a unit vector with the j-th element 1. $\mathcal{E}_i^j$ is the multiset whose i-th element has multiplicity j. In this way, we can get $\nabla F \in \mathbb{R}^{(L-1)\times d}$ and calculate the proposal of each coordinate as below:

$$\text{Categorical}\left(\text{Softmax}\left(\frac{1}{2}\nabla F(\theta)_i^\top (\theta_i' - \theta_i) - \frac{\|\theta_i' - \theta_i\|_2^2}{2\alpha}\right)\right). \tag{7}$$

3.3 Experiments

We implement the Newton proposal on the dim$= 4 \times 4 \times 3 \times 3$ Potts model.

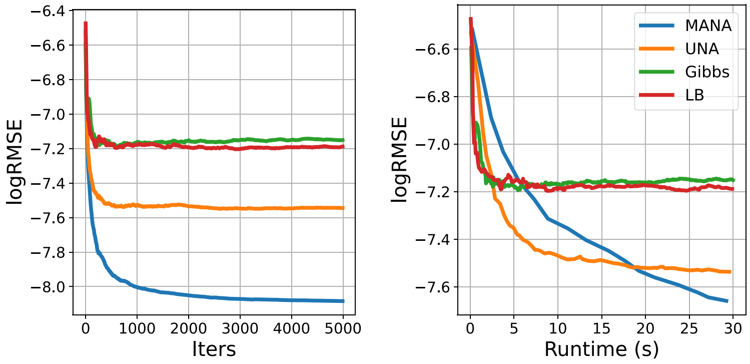


Figure 1: Potts model: lines of MANA run under others.

Figure 1 shows that MANA converges fast in both the number of iterations and running time. Our Newton proposal achieves promising performance for categorical variables.

4 Details and Proof of Theorem 1

As we discuss in Section 5, our Newton proposal can be equivalently obtained by conducting Taylor expansion on the multilinear extension of the target discrete distribution. In this section, we focus on second-order modular functions and study the asymptotic convergence of UNA.

4.1 Definitions of Submodular Function and Multilinear Extension

A **submodular function** is a set function whose value has the property that the difference in the incremental value of the function that a single element makes when added to an input set decreases as the size of the input set increases. In a word, submodular functions have a natural diminishing returns property. If f is submodular then its multilinear extension, $i.e.$, F, is concave along any line $d \geq 0$.

For a discrete distribution whose function is a set function $f : 2^D \to \mathbb{R}$, its **multilinear extension** $F : [0,1]^d \to \mathbb{R}$ is defined as:

$$F(\theta) = \sum_{S \subseteq D} f(S) \prod_{i \in S} \theta_i \prod_{i \in D \setminus S} (1 - \theta_i)\,. \tag{8}$$

It is possible to relate properties of f to properties of its multilinear extension F. In particular, we have:

Proposition 1. Let F be the multilinear extension of f, then:

1. If f is non-decreasing, then F is non-decreasing along any direction $d \geq 0$.
2. If f is submodular then F is concave along any line $d \geq 0$.

Both properties can be established by first looking at how F behaves along coordinates axes. We first calculate the first and second order derivative of $F(\theta)$.

1. Let $i \in D$, since F is linear in θ_i, we have:

$$\frac{\partial F}{\partial \theta_i}(\theta) = \quad F(\theta_1, \ldots, \theta_{i-1}, 1, \theta_{i+1}, \ldots, \theta_d) - F(\theta_1, \ldots, \theta_{i-1}, 0, \theta_{i+1}, \ldots, \theta_d)$$

Let R be the random subset of $D\backslash\{i\}$ where each element $j \in D\backslash\{i\}$ is included with probability θ_j, then we can rewrite:

$$\frac{\partial F}{\partial \theta_i}(\theta) = \mathbb{E}[f(R \cup \{i\})] - \mathbb{E}[f(R)]. \tag{9}$$

2. Similarly, let R_2 be the random subset of $D\backslash\{i,j\}$ and R_3 be the random subset of $D\backslash\{j\}$ where each element k is included with probability θ_k, we have:

$$\begin{aligned}\frac{\partial^2 F}{\partial \theta_i \partial \theta_j}(\theta) &= \mathbb{E}[f(R_2 \cup \{i,j\})] - \mathbb{E}[f(R_2 \cup \{i\})] - \mathbb{E}[f(R_2 \cup \{j\})] + \mathbb{E}[f(R_2)] \\ &= \mathbb{E}[\big(f(R_2 \cup \{i,j\}) - f(R_2 \cup \{i\})\big) - \big(f(R_2 \cup \{j\}) - f(R_2)\big)].\end{aligned} \tag{10}$$

By submodularity of f, the last quantity in (10) is non-positive, $i.e.$, $\frac{\partial^2 F}{\partial \theta_i \partial \theta_j}(\theta) \leq 0$.

Let $\theta \in [0,1]^n$ and $d \geq 0$. We define the function $F_{\theta,d}(\lambda) = F(\theta + \lambda d)$ of the real variable λ. We note that $F'_{\theta,d}(\lambda) = \langle d, \nabla F(\theta + \lambda d)\rangle$ and $F''_{\theta,d} = d^T H_f(\theta + \lambda d)d$.

1. If f is non-decreasing, then $\nabla F(\theta + \lambda d) \geq 0$ and $\langle d, \nabla F(\theta + \lambda d)\rangle \geq 0$. Hence $F_{\theta,d}$ is nondecreasing.
2. If f is submodular, then $H_f(\theta + \lambda d) \leq 0$ and $d^T H_f(\theta + \lambda d)d \leq 0$. Hence $F_{\theta,d}$ is concave.

4.2 Second-Order Modular

For a submodular function f, let $MG(A,e) = f(A \cup \{e\}) - f(A)$ denote the marginal gain from adding element e to set A. For sets A, S, we define $GR(A,S,e) = MG(A,e) - MG(A \cup S, e)$ as the amount by which S reduces the marginal gain from adding e to A. (Here, GR stands for Gain Reduction.) Note that by definition of submodularity, $GR(A,S,e)$ is always non-negative.

The function f is said to be **second-order modular** if, for all sets A, B, S such that $A \subseteq B$, and $S \cap B = \emptyset$, and all elements e, we have: $GR(A,S,e) = GR(B,S,e)$ [Korula et al., 2018].

Specifically, if for any set R_2, $\forall i, j$, $f(R_2 \cup \{i,j\}) - f(R_2 \cup \{i\}) = b_j$ and $f(R_2 \cup \{j\}) - f(R_2) = A_{ij} + b_j$ where A_{ij} and b_j are both constants, then $GR(R_2, \{i\}, j) = b_j$. At this time, f is second-order modular. That is to say, the second-order modular function is in the form of $f(D) = \sum_{u,v \in D} A(u,v) + \sum_{u \in D} b(u)$, whose multilinear function is

$$F(\theta) = \theta^\top A \theta + b\theta, \tag{11}$$

which is exactly in the same form with the energy function of Ising model.

4.3 Proof of Theorem 1

Proof. We finish the proof in the view of multiliear extension, $i.e.$, we see the multilinear extension of the original discrete distribution as the energy function. We first prove the weak convergence and then prove the convergence rate with respect to the stepsize α.

(1) Weak convergence. When $f(D)$ is second-order modular, the Hessian matrix of its multilinear extension F in Equation (11) will be a constant, $i.e.$, $\frac{\partial^2 F}{\partial \theta_i \partial \theta_j}(\theta) = A_{ij}$, $\forall \theta$. We have that $\nabla F(\theta) = 2A^\top \theta + b$, $\nabla^2 F(\theta) = 2A$. Since $\nabla^2 F(\theta)$ is a constant, we can rewrite the proposal distribution as the following

$$\begin{aligned} q_\alpha(\theta' \mid \theta) &= \frac{\exp\left(\frac{1}{2}\nabla F(\theta)^\top (\theta' - \theta) - \frac{1}{2\alpha}\left\|\theta' - \theta\right\|^2\right)}{\sum_x \exp\left(\frac{1}{2}\nabla F(\theta)^\top (x - \theta) - \frac{1}{2\alpha}\|x - \theta\|^2\right)} \\ &= \frac{\exp\left(\frac{1}{2}\nabla F(\theta)^\top (\theta' - \theta) + \frac{1}{2}(\theta' - \theta)^\top A (\theta' - \theta) - (\theta' - \theta)^\top \left(\frac{1}{2\alpha}I + \frac{1}{2}A\right)(\theta' - \theta)\right)}{\sum_x \exp\left(\frac{1}{2}\nabla F(\theta)^\top (x - \theta) + \frac{1}{2}(x - \theta)^\top A (x - \theta) - (x - \theta)^\top \left(\frac{1}{2\alpha}I + \frac{1}{2}A\right)(x - \theta)\right)} \\ &= \frac{\exp\left(\frac{1}{2}\left(F(\theta') - F(\theta)\right) - (\theta' - \theta)^\top \left(\frac{1}{2\alpha}I + \frac{1}{2}A\right)(\theta' - \theta)\right)}{\sum_x \exp\left(\frac{1}{2}(F(x) - F(\theta)) - (x - \theta)^\top \left(\frac{1}{2\alpha}I + \frac{1}{2}A\right)(x - \theta)\right)} \end{aligned}$$

where the last equation is because the Taylor expansion $F\left(\theta'\right)-F(\theta)=\nabla F(\theta)^{\top}\left(\theta'-\theta\right)+\frac{1}{2}\left(\theta'-\theta\right)^{\top} 2A\left(\theta'-\theta\right)$.

Let $Z_\alpha(\theta)=\sum_x \exp\left(\frac{1}{2}(F(x)-F(\theta))-(x-\theta)^{\top}\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-\theta)\right)$, and $\pi_\alpha=\frac{Z_\alpha(\theta)\pi(\theta)}{\sum_x Z_\alpha(x)\pi(x)}$, now we will show that q_α is reversible w.r.t. π_α. We have that

$$
\begin{aligned}
\pi_\alpha(\theta) q_\alpha\left(\theta' \mid \theta\right) &= \frac{Z_\alpha(\theta)\pi(\theta)}{\sum_x Z_\alpha(x)\pi(x)} \cdot \frac{\exp\left(\frac{1}{2}\left(F\left(\theta'\right)-F(\theta)\right)-\left(\theta'-\theta\right)^{\top}\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)\left(\theta'-\theta\right)\right)}{Z_\alpha(\theta)} \\
&= \frac{\exp\left(\frac{1}{2}\left(F\left(\theta'\right)+F(\theta)\right)-\left(\theta'-\theta\right)^{\top}\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)\left(\theta'-\theta\right)\right)}{Z \cdot \sum_x Z_\alpha(x)\pi(x)}.
\end{aligned}
\tag{12}
$$

We can see that the expression in (12) is symmetric in θ and θ'. Therefore q_α is reversible and the stationary distribution is π_α. Now we will prove that π_α converges weakly to π as $\alpha \to 0$. Notice that for any θ,

$$
\begin{aligned}
Z_\alpha(\theta) &= \sum_x \exp\left(\frac{1}{2}(F(x)-F(\theta))-(x-\theta)^{\top}\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-\theta)\right) \\
&\overset{\alpha\downarrow 0}{\Longrightarrow} \sum_x \exp\left(\frac{1}{2}(F(x)-F(\theta))\right)\delta_\theta(x) \\
&= 1,
\end{aligned}
$$

where $\delta_\theta(x)$ is a Dirac delta. It follows that π_α converges pointwisely to $\pi(\theta)$. By Scheffé's Lemma, we attain that π_α converges weakly to π.

(2) Convergence Rate w.r.t. Stepsize.

Let us consider the convergence rate in terms of L_1-norm

$$
\left\|\pi_\alpha-\pi\right\|_1=\sum_\theta\left|\frac{Z_\alpha(\theta)\pi(\theta)}{\sum_x Z_\alpha(x)\pi(x)}-\pi(\theta)\right|.
$$

We write out each absolute value term

$$
\begin{aligned}
\left|\frac{Z_\alpha(\theta)\pi(\theta)}{\sum_x Z_\alpha(x)\pi(x)}-\pi(\theta)\right| &= \pi(\theta)\left|\frac{Z_\alpha(\theta)}{\sum_x Z_\alpha(x)\pi(x)}-1\right| \\
&= \pi(\theta)\cdot \\
&\left|\frac{1+\sum_{x\neq\theta}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(\theta)-(x-\theta)^{\top}\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-\theta)\right)}{1+\sum_y \frac{1}{Z}\exp(F(y))\sum_{x\neq y}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(y)-(x-y)^{\top}\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-y)\right)}-1\right|.
\end{aligned}
$$

Since $\lambda_{\min}(A)\|x\|^2 \leq x^{\top}Ax, \forall x$, it follows that

$$
(x-\theta)^{\top}\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-\theta) \geq \frac{1+\alpha\lambda_{\min}}{2\alpha}\|x-\theta\|^2.
$$

We also notice that $\min_{x\neq\theta}\|x-\theta\|^2 = 1$, thus when $\frac{Z_\alpha(\theta)}{\sum_x Z_\alpha(x)\pi(x)} - 1 > 0$, we get

$$\begin{aligned}
\left|\frac{Z_\alpha(\theta)\pi(\theta)}{\sum_x Z_\alpha(x)\pi(x)} - \pi(\theta)\right| &= \pi(\theta)\cdot \\
&\quad \left(\frac{1+\sum_{x\neq\theta}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(\theta)-(x-\theta)^\top\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-\theta)\right)}{1+\sum_y \frac{1}{Z}\exp(F(y))\sum_{x\neq y}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(y)-(x-y)^\top\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-y)\right)} - 1\right) \\
&\leq \pi(\theta)\left(1+\sum_{x\neq\theta}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(\theta)-\frac{1+\alpha\lambda_{\min}}{2\alpha}\|x-\theta\|^2\right)-1\right) \\
&\leq \pi(\theta)\left(1+\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right)\sum_{x\neq\theta}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(\theta)\right)-1\right) \\
&= \pi(\theta)\left(\sum_{x\neq\theta}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(\theta)\right)\right)\cdot\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right) \\
&\leq \pi(\theta)\left(\sum_x \exp(F(x))\right)\cdot\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right) \\
&= \pi(\theta)Z\cdot\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right).
\end{aligned}$$

Similarly, when $\frac{Z_\alpha(\theta)}{\sum_x Z_\alpha(x)\pi(x)} - 1 < 0$, we have,

$$\begin{aligned}
\left|\frac{Z_\alpha(\theta)\pi(\theta)}{\sum_x Z_\alpha(x)\pi(x)} - \pi(\theta)\right| &= \pi(\theta)\cdot \\
&\quad \left(1-\frac{1+\sum_{x\neq\theta}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(\theta)-(x-\theta)^\top\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-\theta)\right)}{1+\sum_y \frac{1}{Z}\exp(F(y))\sum_{x\neq y}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(y)-(x-y)^\top\left(\frac{1}{2\alpha}I+\frac{1}{2}A\right)(x-y)\right)}\right) \\
&\leq \pi(\theta)\left(1-\frac{1}{1+\sum_y \frac{1}{Z}\exp(F(y))\sum_{x\neq y}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(y)-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right)}\right) \\
&= \pi(\theta)\left(\frac{\sum_y \frac{1}{Z}\exp(F(y))\sum_{x\neq y}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(y)-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right)}{1+\sum_y \frac{1}{Z}\exp(F(y))\sum_{x\neq y}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(y)-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right)}\right) \\
&\leq \pi(\theta)\left(\sum_y \frac{1}{Z}\exp(F(y))\sum_{x\neq y}\exp\left(\frac{1}{2}F(x)-\frac{1}{2}F(y)\right)\right)\cdot\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right) \\
&\leq \pi(\theta)\left(\sum_x \exp(F(x))\right)\cdot\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right) \\
&= \pi(\theta)Z\cdot\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right).
\end{aligned}$$

Therefore, the difference between π_α and π can be bounded as follows

$$\|\pi_\alpha - \pi\|_1 \leq \sum_\theta \pi(\theta)Z\cdot\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right) = Z\cdot\exp\left(-\frac{1+\alpha\lambda_{\min}}{2\alpha}\right).$$

□

5 Proof of Theorem 2

Proof. Our proof follows from Theorem 1 of [Grathwohl et al., 2021] and Theorem 2 of [Zanella, 2020], which state that for two p-reversible Markov transition kernels $Q_1(x', x)$ and $Q_2(x', x)$, if there exists $c > 0$ for all $x' \neq x$ such that $Q_1\left(x', x\right) > c \cdot Q_2\left(x', x\right)$ then

1. $\operatorname{var}_p\left(h, Q_1\right) \leq \frac{\operatorname{var}_p\left(h, Q_1\right)}{c} + \frac{1-c}{c} \cdot \operatorname{var}_p(h)$
2. $\operatorname{Gap}\left(Q_1\right) \geq c \cdot \operatorname{Gap}\left(Q_2\right)$

where $\operatorname{var}_p(h, Q)$ is the asymptotic variance and $\operatorname{Gap}(Q)$ is the spectral gap, which are both defined in the main body of this paper. $\operatorname{var}_p(h)$ is the standard variance $E_p\left[h(x)^2\right] - E_p[h(x)]^2$. Our proof proceeds by showing we can bound $Q^{\nabla}\left(x', x\right) \geq c \cdot Q\left(x', x\right)$, and the results of the theorem then follow directly from Theorem 2 of [Zanella, 2020].

5.1 Definitions

We begin by writing down the proposal distribution of interest and their corresponding Markov transition kernels. For ease of notion we define some values

$$\begin{aligned}
\Delta\left(\theta', \theta\right) &:= U\left(\theta'\right) - U(\theta); \\
\tilde{\Delta}\left(\theta', \theta\right) &:= \Delta_h[U](\theta)^{\top}\left(\theta' - \theta\right); \\
D &:= \sup_{\theta' \in \Theta}\left\|\theta' - \theta\right\|.
\end{aligned}$$

Then our original proposal, $i.e.$, the MALA-like locally-balanced proposal for θ' is

$$q_0\left(\theta' \mid \theta\right) = \frac{\exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right) - \frac{1}{2\alpha}\left(\theta' - \theta\right)^2\right)}{Z(\theta)} \tag{13}$$

where we have defined

$$Z(\theta) = \sum_{\theta' \in \Theta} \exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right) - \frac{1}{2\alpha}\left(\theta' - \theta\right)^2\right).$$

When we examine the acceptance rate of the proposal we find

$$\begin{aligned}
& \exp\left(U\left(\theta'\right) - U(\theta)\right) \frac{q_0\left(\theta \mid \theta'\right)}{q_0\left(\theta' \mid \theta\right)} \\
= & \exp\left(\Delta(\theta', \theta)\right) \frac{\exp\left(\frac{1}{2}\Delta(\theta, \theta') - \frac{1}{2\alpha}(\theta - \theta')^2\right) Z(\theta)}{\exp\left(\frac{1}{2}\Delta(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right) Z(\theta')} \\
= & \exp\left(\frac{1}{2}\Delta(\theta', \theta) + \frac{1}{2}\Delta(\theta, \theta')\right) \frac{Z(\theta)}{Z(\theta')} \\
= & \frac{Z(\theta)}{Z\left(\theta'\right)}
\end{aligned}$$

Then the acceptance rate of the target proposal in (13) can be simplified as

$$\min\left\{1, \exp\left(U\left(\theta'\right) - U(\theta)\right) \frac{q_0\left(\theta \mid \theta'\right)}{q_0\left(\theta' \mid \theta\right)}\right\} = \min\left\{1, \frac{Z(\theta)}{Z\left(\theta'\right)}\right\}.$$

This corresponding Markov transition kernel is

$$\begin{aligned}
Q\left(\theta', \theta\right) &= q_0\left(\theta' \mid \theta\right) \min\left\{1, \frac{Z(\theta)}{Z\left(\theta'\right)}\right\} \\
&= \min\left\{\frac{\exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right) - \frac{1}{2\alpha}\left(\theta' - \theta\right)^2\right)}{Z(\theta)}, \frac{\exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right) - \frac{1}{2\alpha}\left(\theta' - \theta\right)^2\right)}{Z\left(\theta'\right)}\right\}.
\end{aligned}$$

Our proposed Newton proposal is the first-order Newton's series approximation of the original target proposal (13) for $\theta' \in \Theta$:

$$q(\theta' \mid \theta) = \frac{\exp\left(\frac{1}{2}\tilde{\Delta}(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right)}{\tilde{Z}(\theta)}$$

where we have defined

$$\tilde{Z}(\theta) = \sum_{\theta' \in \Theta} \exp\left(\frac{1}{2}\tilde{\Delta}(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right).$$

For our Newton proposal, we simplify the term in the acceptance rate of the proposal as

$$\begin{aligned}
&\exp\left(U(\theta') - U(\theta)\right)\frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)} \\
&= \exp\left(\Delta(\theta', \theta)\right)\frac{\exp\left(\frac{1}{2}\tilde{\Delta}(\theta, \theta') - \frac{1}{2\alpha}(\theta - \theta')^2\right)\tilde{Z}(\theta)}{\exp\left(\frac{1}{2}\tilde{\Delta}(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right)\tilde{Z}(\theta')} \\
&= \exp\left(\Delta(\theta', \theta) + \frac{1}{2}\tilde{\Delta}(\theta, \theta') - \frac{1}{2}\tilde{\Delta}(\theta', \theta)\right)\frac{\tilde{Z}(\theta)}{\tilde{Z}(\theta')}
\end{aligned}$$

Then the Markov transition kernel $\tilde{Q}(\theta', \theta) =$

$$\begin{aligned}
&q(\theta' \mid \theta)\min\left\{1, \exp\left(\Delta(\theta', \theta) + \frac{1}{2}\tilde{\Delta}(\theta, \theta') - \frac{1}{2}\tilde{\Delta}(\theta', \theta)\right)\frac{\tilde{Z}(\theta)}{\tilde{Z}(\theta')}\right\} \\
&= \min\left\{\frac{\exp\left(\frac{1}{2}\tilde{\Delta}(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right)}{\tilde{Z}(\theta)}, \frac{\exp\left(\Delta(\theta', \theta) + \frac{1}{2}\tilde{\Delta}(\theta, \theta') - \frac{1}{2\alpha}(\theta' - \theta)^2\right)}{\tilde{Z}(\theta')}\right\}
\end{aligned}$$

5.2 Preliminaries

It can be seen that $\tilde{\Delta}(\theta', \theta)$ is a first-order Newton's series approximation to $\Delta_h(\theta', \theta)$. When the finite difference $\Delta[U](\theta)$ has an analog of Lipschitz continuity, *i.e.*, $\left|\tilde{\Delta}(\theta', \theta) - \Delta(\theta', \theta)\right| \leq \frac{L}{2}\|\theta' - \theta\|^2$, since $\|\theta' - \theta\|^2$ is bounded, we have

$$-\frac{L}{2}D^2 \leq \tilde{\Delta}(\theta', \theta) - \Delta(\theta', \theta) \leq \frac{L}{2}D^2$$

5.3 Normalizing Constant Bounds

We derive upper- and lower-bounds on $\tilde{Z}(\theta)$ in terms of $Z(\theta)$.

$$\begin{aligned}
\tilde{Z}(\theta) &= \sum_{\theta' \in \Theta} \exp\left(\frac{1}{2}\tilde{\Delta}(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right) \\
&= \sum_{\theta' \in \Theta} \exp\left(\frac{1}{2}\Delta(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right) \cdot \exp\left(\frac{1}{2}\tilde{\Delta}(\theta', \theta) - \frac{1}{2}\Delta(\theta', \theta)\right) \\
&\leq \sum_{\theta' \in \Theta} \exp\left(\frac{1}{2}\Delta(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right) \cdot \exp\left(\frac{LD^2}{4}\right) \\
&= \exp\left(\frac{LD^2}{4}\right)\sum_{\theta' \in \Theta} \exp\left(\frac{1}{2}\Delta(\theta', \theta) - \frac{1}{2\alpha}(\theta' - \theta)^2\right) \\
&= \exp\left(\frac{LD^2}{4}\right)Z(\theta)
\end{aligned}$$

Following the same argument we can show

$$\tilde{Z}(\theta) \geq \exp\left(\frac{-LD^2}{4}\right) Z(\theta).$$

In this way, we get bounds between the normalizing constants of original proposal and our approximated proposal.

5.4 Inequalities of Minimums

We show $\tilde{Q}\left(\theta', \theta\right) \geq c \cdot Q\left(\theta', \theta\right)$ for $c = \exp\left(\frac{-LD^2}{2}\right)$. Since both $Q\left(\theta', \theta\right) = \min\{a, b\}$ and $\tilde{Q}\left(\theta', \theta\right) = \min\left\{\tilde{a}, \tilde{b}\right\}$, it is sufficient to show $\tilde{a} \geq c \cdot a$ and $\tilde{b} \geq c \cdot b$ to prove the desired result. We begin with the a terms

$$\begin{aligned}
\frac{\tilde{a}}{a} &= \frac{\exp\left(\frac{1}{2}\tilde{\Delta}\left(\theta', \theta\right) - \frac{1}{2\alpha}(\theta' - \theta)^2\right)}{\tilde{Z}(\theta)} \frac{Z(\theta)}{\exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right) - \frac{1}{2\alpha}(\theta' - \theta)^2\right)} \\
&= \frac{Z(\theta)}{\tilde{Z}(\theta)} \exp\left(\frac{1}{2}\tilde{\Delta}\left(\theta', \theta\right) - \frac{1}{2}\Delta\left(\theta', \theta\right)\right) \\
&\geq \exp\left(\frac{-LD^2}{4}\right) \exp\left(\frac{1}{2}\tilde{\Delta}\left(\theta', \theta\right) - \frac{1}{2}\Delta\left(\theta', \theta\right)\right) \\
&\geq \exp\left(\frac{-LD^2}{4}\right) \exp\left(\frac{-LD^2}{4}\right) \\
&= \exp\left(\frac{-LD^2}{2}\right)
\end{aligned}$$

Now the b terms

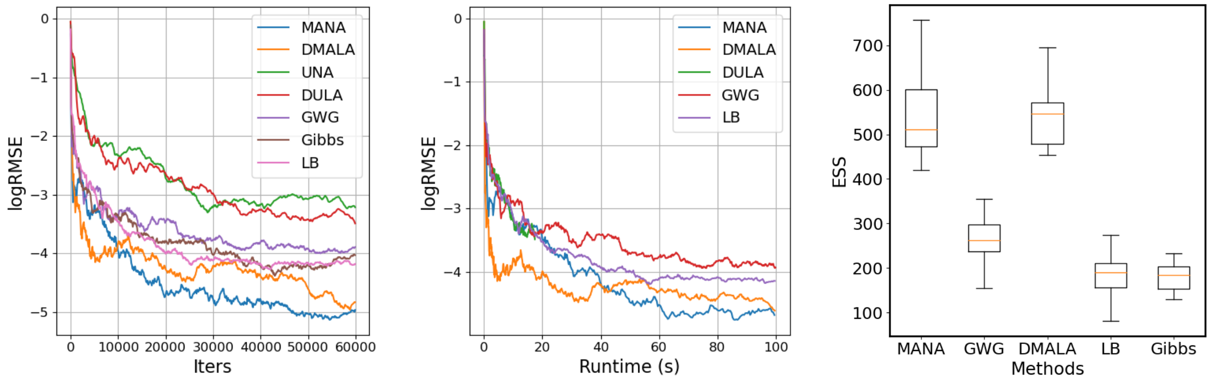
$$\begin{aligned}
\frac{\tilde{b}}{b} &= \frac{\exp\left(\Delta(\theta', \theta) + \frac{1}{2}\tilde{\Delta}(\theta, \theta') - \frac{1}{2\alpha}(\theta' - \theta)^2\right)}{\tilde{Z}(\theta)'} \frac{Z\left(\theta'\right)}{\exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right) - \frac{1}{2\alpha}(\theta' - \theta)^2\right)} \\
&= \frac{Z\left(\theta'\right)}{\tilde{Z}\left(\theta'\right)} \frac{\exp\left(\Delta\left(\theta', \theta\right) + \frac{1}{2}\tilde{\Delta}\left(\theta, \theta'\right)\right)}{\exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right)\right)} \\
&= \frac{Z\left(x'\right)}{\tilde{Z}\left(x'\right)} \exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right) + \frac{1}{2}\tilde{\Delta}\left(\theta, \theta'\right)\right) \\
&\geq \exp\left(\frac{-LD^2}{4}\right) \exp\left(\frac{1}{2}\Delta\left(\theta', \theta\right) + \frac{1}{2}\tilde{\Delta}\left(\theta, \theta'\right)\right) \\
&= \exp\left(\frac{-LD^2}{4}\right) \exp\left(\frac{1}{2}\tilde{\Delta}\left(\theta, \theta'\right) - \frac{1}{2}\Delta\left(\theta, \theta'\right)\right) \\
&\geq \exp\left(\frac{-LD^2}{4}\right) \exp\left(\frac{-LD^2}{4}\right) \\
&= \exp\left(\frac{-LD^2}{2}\right)
\end{aligned}$$

5.5 Conclusions

We have shown that $\tilde{a} \geq \exp\left(\frac{-LD^2}{2}\right) a$ and $\tilde{b} \geq \exp\left(\frac{-LD^2}{2}\right) b$ and therefore it holds that

$$\tilde{Q}\left(\theta', \theta\right) \geq \exp\left(\frac{-LD^2}{2}\right) Q\left(\theta', \theta\right)$$

From this, the main result follows directly from Theorem 2 of [Zanella, 2020]. □



(a) log RMSE w.r.t. Iterations

(b) log RMSE w.r.t. Runtime

(c) ESS of Proposals

Figure 2: Ising model sampling results. **(a)** MANA converges faster than the baselines in number of iterations. **(b)** MANA converges faster than the baselines in the same running time. **(c)** MANA and DMALA yield the largest effective sample size (ESS) among all the methods compared.

6 Experiments on Ising Model

In the main body of the paper, we have shown that for distributions without natural differentiable extension, our Newton proposal outperforms popular baselines, including Gibbs sampler and locally-balanced sampler (LB) [Zanella, 2020], while continuous relaxation-based proposals become valid. We also apply Newton proposal to discrete distributions with natural differentiable distributions, such as the Ising model, to show the broad applicability of our method. In this case, we compare our Newton proposal with proposals which do not rely on gradients (Gibbs sampler and LB) and continuous relaxation methods (GWG[Grathwohl et al., 2021] and DLP [Zhang et al., 2022]).

6.1 Experiment Settings

We have shown in Section 5 that Newton proposal is equivalent to DLP when the discrete distribution has a natural differential extension. Therefore, if Newton proposal and DLP are applied to sample from the same discrete distribution, they will have the same results. However, we also demonstrate in Theorem 1 that the smallest eigenvalue of A is related to the asymptotic convergence. Consider the Ising model whose distribution is

$$f(D) = \sum_{u,v \in D} A(u,v) + \sum_{u \in D} b(u), \tag{14}$$

where A is a binary adjacency matrix, a is the connectivity strength and b is the bias. We can see the diagonal of matrix A in Equation (14) will not affect the distribution in the discrete domain because u and v are different elements in the set D. Meanwhile, DLP is applied to a discrete distribution in a 0-diagonal quadratic form, which can be seen as the multilinear extension of the distribution of Newton proposal when A is 0-diagonal. We are interested in the performance of Newton proposal and DLP when λ_{min} in Newton proposal is larger than that in DLP.

6.2 Ising Model Sampling Results

We consider a 4 by 4 lattice Ising model with random variable $\theta \in \{-1,1\}^d$, and $d = 4 \times 4 = 16$. The distribution is

$$f(D) = \sum_{u,v \in D} A(u,v) + \sum_{u \in D} b(u)$$

where A is a binary adjacency matrix, $a = 0.1$ is the connectivity strength and $b = 0.2$ is the bias.

We run 60000 iterations with all samplers. To make the comparison of convergence speed fair, we tune the stepsizes so that MANA and DMALA change almost the same number of coordinates in a single update. The stepsizes of MANA, DMALA, UNA and DULA as 0.5, 0.8, 0.1 and 0.1, respectively.

We first compare the root-mean-square error (RMSE) between the estimated mean and the true mean in Figure 2. MANA is the fastest to converge in terms of both iterations and runtime. This demonstrates (1) changing many coordinates in one step accelerates the convergence compared to LB and GWG; (2) the finite difference works like the gradient in cases with natural differential extensions to explore the discrete space. In fact, the Newton proposal is equivalent to DLP when the discrete distribution has a natural differential extension, as shown in Section 5. That's why DMALA and MANA achieve similar convergence when they change the same number of coordinates in a single update. MANA and DMALA converge obviously faster than UNA and DULA because the MH correction accelerates the convergence for this task. In Figure 2(c), we compare the effective sample size (ESS) of different samplers. MANA and DMALA significantly outperform other methods, indicating the correlation among its samples is low due to making significant updates in each step.

7 ROUGE Score in Text Summarization

We used the official implementation of Rouge score evaluation toolkit (https://pypi.org/project/rouge-score/). We follow the same ROUGE score configuration as the settings in [Lin, 2004]: ROUGE version 1.5.5 with options: -a -c 95 -b 665 -m -n 4 -w 1.2.

8 Dimension-wise Majority Vote in Image Retrieval

For the image retrieval task, the pseudo code for the dimension-wise majority vote algorithm mentioned in the main body of our paper is as follows:

Algorithm 2 Dimension-wise Majority Vote

```
$X \in \{0,1\}^{K,D}$
ans ← zeros([D])
indices ← argsort(mean($X$))
▷ select top dimensions under cost constraint $C$
ans[indices[: $C$]] = 1
return ans
```

References

Giacomo Zanella. Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

Aytunc Sahin, Yatao Bian, Joachim Buhmann, and Andreas Krause. From sets to multisets: provable variational inference for probabilistic integer submodular models. In *International Conference on Machine Learning*, pages 8388–8397. PMLR, 2020.

Nitish Korula, Vahab Mirrokni, and Morteza Zadimoghaddam. Online submodular welfare maximization: Greedy beats 1/2 in random order. *SIAM Journal on Computing*, 47(3):1056–1086, 2018.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning*, pages 3831–3841. PMLR, 2021.

Ruqi Zhang, Xingchao Liu, and Qiang Liu. A langevin-like sampler for discrete distributions. In *International Conference on Machine Learning*, pages 26375–26396. PMLR, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.