Semi-Empirical Objective Functions for MCMC Proposal Optimization

Chris Cannella and Vahid Tarokh
Department of ECE
Duke University
Durham, NC, USA

Abstract—Current objective functions used for training neural MCMC proposal distributions implicitly rely on architectural restrictions to yield sensible optimization results, which hampers the development of highly expressive neural MCMC proposal architectures. In this work, we introduce and demonstrate a semi-empirical procedure for determining approximate objective functions suitable for optimizing arbitrarily parameterized proposal distributions in MCMC methods. Our proposed Ab Initio objective functions are utilized to optimize these deep learning based proposal distribution model classes. Ideally, we would like to develop a practical means of optimizing MCMC performance over arbitrarily parameterized classes of proposal distributions.

We describe a set of first principles properties that may be reasonably assumed of the ground truth objective function underlying our notion of MCMC efficiency, this could be accomplished by straightforward optimization over a set of proposal distributions. Past computational limitations encouraged the use of architecturally limited MCMC schemes that yield useful results for a wide range of target distributions, such as Random Walk Metropolis (RWM) [1], Metropolis Adjusted Langevin Diffusion (MALA) [2], and Hamiltonian Monte Carlo (HMC) [3]. Although much research has been devoted to adaptive methods to improve MCMC performance, traditional adaptive methods [4]–[7] focus on optimization across highly restricted proposal distribution model classes. Ideally, we would like to develop a practical means of optimizing MCMC performance over arbitrarily parameterized classes of proposal distributions.

Given their demonstrated success in parameterizing expressive distributions, deep generative models are naturally suited to the problem of MCMC proposal optimization. Recent research regarding applications of deep learning to proposal optimization has yielded new MCMC schemes [8]–[10] and extensions of existing schemes [11]–[14]. A variety of objective functions are utilized to optimize these deep learning based approaches, with functional forms generally dependent on the type of MCMC scheme being optimized. As demonstrated by the experiments of this work, the objective functions currently used for MCMC proposal optimization rely on model class restrictions imposed on the proposal architecture and are not suitable for optimizing proposal distributions more expressive than traditional MCMC schemes. In this work, we introduce and demonstrate Ab Initio objective functions for MCMC proposal optimization intended to remain compatible with proposal distributions defined with deep generative models.

Presumably, there exists some “ground truth” objective function underlying our notion of MCMC sampling performance. However, even after the decades of research regarding MCMC methods, there appears to be no universal definition for what we mean by MCMC efficiency. Metrics like effective sample size (ESS) [15], [16] generally coincide with our notion of sampling performance, but no such metric serves as the canonical definition of MCMC efficiency. Theoretical analysis regarding optimal acceptance rates for particular MCMC schemes [17]–[21] considers restricted proposal schemes and targets within a continuous diffusional limit wherein our common performance metrics converge in their definition of optimality [17]. Within this diffusional limit, useful properties regarding MCMC optimality (e.g., the rules of thumb to seek an acceptance rate of 0.234 when using RWM and of 0.574 when using MALA) may be derived without specifically defining sampling performance. In light of the lack of an exactly specified objective function for MCMC efficiency, our Ab Initio objective functions seek to approximate the ground truth definition of sampling performance by adhering to certain reasonable first principles properties and fitting to reproduce the “mathematical observations” provided by existing theoretical analysis of optimal acceptance rates.

Our contributions are as follows:

- We describe a set of first principles properties that may be reasonably assumed of the ground truth objective function underlying our notion of MCMC efficiency.
- We illustrate the construction of an example Ab Initio objective function via the combination of simpler objective functions with coefficients fit to reproduce analytically known optimal behavior on certain reference problems.
- We verify the generality of the resulting Ab Initio objective function through its ability to reproduce analytically known optimal results for a range of optimization tasks beyond the reference problem used in its construction.
• Through a series of illustrative experiments, we demonstrate the advantages of Ab Initio objective functions for optimizing neural MCMC proposal distributions.

II. RELATION TO PRIOR WORK

Ab Initio techniques [22]–[24] are used in the physical sciences to simulate systems that would otherwise be unobservable in a laboratory setting. These Ab Initio methods are founded on principled approximations of fundamental physical laws with parameters chosen to reproduce the properties of reference systems that can be observed. We take inspiration from these methods in the physical sciences for our methodology.

The purpose of this work is to introduce a procedure for selecting objective functions for MCMC proposal optimization that are suited to the optimization of proposal model classes arbitrarily parameterized by deep generative models. Research into the applications of deep learning for MCMC proposal optimization [8]–[14] has yielded a number of candidates for this objective function. We therefore compare our Ab Initio objective functions to these candidates on the basis of their suitability for optimization of arbitrary proposal distributions.

Pure KL-divergence based objectives [11], [12], [25], [26] have found some success, particularly when optimizing re-sampling style schemes. These objectives are unable to properly optimize proposals within a diffusionary limit [27]. We therefore omit pure KL-divergence based objectives from the comparisons within this work.

Adversarial objectives have been used to optimize proposal distributions, notably with A-NICE-MC [8], [9]. We view adversarial training as approximately optimizing an existing performance measure (e.g. KL-divergence), rather than defining a fundamentally new performance measure. We therefore also omit adversarial objectives from our comparisons.

Mean squared jump distance [28] (MSJD) remains a popular objective for optimizing proposals. Notably, L2HMC [13] is optimized using a modification of MSJD with a regularization intended to encourage mixing of the resulting Markov Chain. As shown in our experiments, both MSJD and L2HMC’s modification can produce arbitrarily non-ergodic Markov Chains when optimizing proposal distributions with position dependence. Our Ab Initio objective functions avoid this undesirable optimization behavior by maintaining i.i.d. resampling from the target as their unique global minima. Another potential advantage of the Ab Initio objective of Equation (2) over MSJD based objective functions is that it does not rely on the notion of a metric of the underlying space of the proposal distribution, and so may be more suited to optimizing proposals in settings where the space is equipped with only a measure and not a metric (e.g. MCMC settings involving discrete or graph based distributions).

Recently, the Generalized Speed Measure (GSM) was introduced [27] as an objective function for MCMC proposal optimization. The GSM amounts to maximizing proposal entropy subject to the constraint of achieving a user specified acceptance rate. Although it is theoretically well motivated, the GSM relies on knowledge of optimal acceptance rates for a given optimization problem, which will generally be unknown when using very general neural architectures. The objective function of Scout MCMC [29] is inspired by the GSM, using KL-Divergence in place of proposal entropy, but similarly requires user knowledge of optimal acceptance rates to obtain an optimal result. Our example Ab Initio objective function of Equation (1) utilizes components found within the GSM and Scout MCMC, while also ensuring the functional limitations argued for in Section IV. A key advantage of our Ab Initio objective functions over the GSM is that they do not require prior knowledge to recover near optimal MCMC behavior (e.g. our Ab Initio objective function, whose construction only involved the knowledge of the optimal acceptance rate for RWM, recovers optimized MALA proposals with acceptance rates around 0.5). An additional advantage of our Ab Initio objective functions over the GSM is that they may be used to compare the efficiencies of proposals with differing optimal acceptance rates outside of the context of parameter optimization.

III. MCMC PROPOSAL OPTIMIZATION

In this work, we consider the task of optimizing a proposal density \( g_t(\mathbf{x}'|\mathbf{x}) \) to sample from a target density \( \pi(\mathbf{x}) \) for a MCMC task. Proposals are accepted with rate \( \alpha_{g_t}(\mathbf{x}'|\mathbf{x}) \) that ensures the resulting Markov Chain converges towards \( \pi(\mathbf{x}) \). For this work, we restrict ourselves to Metropolis-Hastings type schemes, wherein \( \alpha_{g_t}(\mathbf{x}'|\mathbf{x}) = \min\{1, \frac{\pi(\mathbf{x}'}{\pi(\mathbf{x})} g_t(\mathbf{x}|\mathbf{x}')} \}. \) For optimization, we utilize some measure of sampling performance to define an objective function \( \mathcal{L}[g; \pi] \) that imposes an ordering over proposal distributions by defining \( g_1 < g_2 \) exactly when \( \mathcal{L}[g_1; \pi] \geq \mathcal{L}[g_2; \pi] \). We do not seek to provide an argument for the application of deep learning methods to MCMC proposal optimization, so we will simply assume that all objective functions of interest are well-behaved for optimization via deep learning techniques (i.e. they are continuous and almost surely differentiable). We will also assume that all probability densities considered are positive and non-singular.

Let \( G \) be the set of all proposed proposals to consider during optimization and let \( D \) be the group of diffeomorphisms over the space of our data (\( \mathbb{R}^n \)). To perform optimization, we first select some \( T \in D \) that provides us with the coordinate system we will use for optimization. Defining:

\[
T \circ f(\mathbf{z}) = f(T^{-1}(\mathbf{z})) \left| \frac{\partial T^{-1}(\mathbf{z})}{\partial \mathbf{z}} \right|
\]

We finally optimize to find \( g_{opt} = \arg\min_{g \in G} \mathcal{L}[T \circ g; T \circ \pi] \).

The focus of this work is to illustrate the construction of objective functions such that, when the model class \( G \) is very expressive (e.g. having been parameterized by a deep generative model), the optimized proposal \( g_{opt} \) aligns with our notion of an efficient MCMC proposal.

IV. PROPERTIES OF THE GROUND TRUTH OBJECTIVE FUNCTION \( \mathcal{L}^* \)

For us to sensibly pursue the task of MCMC proposal optimization, we must assume the existence of some ground truth objective, \( \mathcal{L}^* \), that relies on our notion of sampling

4759
performance. As previously stated, we currently do not know a universal definition for $L^*$. We may, however, assume that $L^*$ is proper: Define an objective function to be proper if it attains a unique global minimum at $g(\vec{x}|\vec{x}) = \pi(\vec{x})$ and $\alpha_{g,\pi}(\vec{x}|\vec{x}) = 1$ for all $\vec{x}, \vec{x}$. The overall goal of our MCMC methodology is to approximate perfect i.i.d. sampling from the target, $\pi$. We therefore find it uncontroversial to assume that $L^*$ is proper.

b) $L^*$ is Representation Independent: Define an objective function, $L$, to be representation independent if, for all $T \in T$, and for all proposal distributions $g_1, g_2$, $L[g_1; \pi] > L[g_2; \pi]$ if and only if $L[T \circ g_1; T \circ \pi] > L[T \circ g_2; T \circ \pi]$. Similarly, we will say $L$ is representation invariant over $T$ if $L[T \circ g; T \circ \pi] = L[g; \pi]$ for all $g$. If $L^*$ were not representation independent over $D$, then our definition of sampling performance would depend on which $T$ is used when computing the optimization. To fully justify an ordering, we would need to justify our selection of a particular member from $D$, which we should expect to be exceptionally burdensome. We usually have little prior justification for selecting a particular $T$ and instead often use the coordinate system in which data was originally collected, perhaps applying some rescaling and recentering. Thus, unless contradicted by future experimental or theoretical results, we should assume that $L^*$ is at least representation independent.

c) $L^*$ Yields Established Optimal Results: Prior theoretical analysis has established certain properties regarding optimal proposal distributions for a number of MCMC schemes under diffusory limits. If $L^*$ is to correspond to the same notion of sampling performance, it must yield the same results. Thus, we should expect that optimization of $L^*$ will recover the properties established within these theoretical works when applied in the same diffusory limits.

V. CONSTRUCTING AB INITIO OBJECTIVE FUNCTIONS

Without knowing the exact form of $L^*$, we must resort to finding a useful approximation. Although the assumptions of being proper and representation independent limit the functional class to which $L^*$ belongs, this functional class remains quite expansive. This situation is greatly simplified by restricting our consideration to representation invariant objective functions, because the positive weighted combination of proper and representation invariant objective functions remains a proper and representation invariant objective function.

This allows us to construct potential approximations of $L^*$ by the combination of simpler objective functions as weighted by hyperparameter coefficients. These hyperparameter coefficients can then be fit so as to recover optimal properties established within existing theoretical works. We call the resulting objective function an Ab Initio objective function.

For this work, we consider a functional class of the form:

$$ L[g; \pi] = E_{g\sim\pi} [ L_{KL}(g(\vec{x}|\vec{x})||\pi(\vec{x}')) ] - A d E_{g\sim\pi}[\log \alpha(\vec{x}|\vec{x})] $$  \hspace{1cm} (1)

Where $A$ is a hyperparameter coefficient for fitting and $d$ is the dimensionality of the target. In relation to the GSM [27], this functional class differs primarily by the introduction of a cross entropy term that serves to ensure that, as proven in an extended version of this work [30], the functional class is proper and representation invariant. In relation to Scourt MCMC [29], the functional class differs by the introduction of the dimension dependent term, $d$, which prevents the objective function from yielding trivial optimizations in an asymptotic limit as $d \to \infty$. The choice of $d$ was made after a limited search, as it yields good agreement with known optimal results for a range of dimensionalities of practical interest, though future asymptotic analysis 3 can further refine this choice.

VI. FITTING AND VERIFYING AB INITIO OBJECTIVE FUNCTIONS

There are many possible approaches to fitting the coefficients of an Ab Initio objective function to recover optimal results over reference problems. General and principled methodologies for this fitting are provided by procedures like stochastic Levenberg-Marquardt optimization [31].

For simplicity, we determined $A$ by a manual approximate Newton-Rhapson procedure to match the theoretical result that RWM has an optimal acceptance rate of $0.234$ when targeting a multivariate gaussian of zero mean and identity covariance. Of course, alternative reference problems and their combinations may be used to fit Ab Initio objective functions, the effects of which are explored within the extended version of this work [30]. As the theoretical results are set within the diffusory limit of $d \to \infty$, we must match the theoretical result in a problem with sufficiently large dimension. Here, we perform this match to theoretical reference on a problem with 1000 dimensions. We recovered the desired acceptance rate with $A = 0.18125$ and the Ab Initio objective function considered through the remainder of this work is therefore:

$$ L[g; \pi] = E_{g\sim\pi} [ L_{KL}(g(\vec{x}|\vec{x})||\pi(\vec{x}')) ] - 0.18125 d E_{g\sim\pi}[\log \alpha(\vec{x}|\vec{x})] $$  \hspace{1cm} (2)

By fitting the parameters of an Ab Initio objective function, we hope to approximate a general notion of MCMC efficiency beyond the particularities of the reference problem(s) used in its construction. To verify this robustness, we utilized the Ab Initio function listed in Equation (2) to perform MCMC optimization 3 in a number of tasks where the optimal acceptance rate is analytically known, with varying target distribution (independent gaussian, laplace, cauchy, and uniform), dimensionality (100-10,000 dimensions), and MCMC...
scheme (RWM with gaussian proposals and MALA). These tasks are all within the diffusionary limit and involve a severely limited model class for proposal distributions (optimizing only proposal step size). Under these limitations, acceptance rate adequately summarizes differences between optimized model parameters and MSJD adequately summarizes the efficiency of the resulting Markov Chains. These measures are estimated for each optimized model based on 25,000 proposals from starting points independently sampled from the target distribution. To gather statistics regarding the mean and standard error of reported variables, each verification optimization is replicated a total 5 times. These results are listed in Table I.

We find that the Ab Initio objective of Equation (2) exhibits good agreement with known analytical results beyond the reference problem used in its construction. These results demonstrate that the Ab Initio objective remains a useful approximation of our notion of MCMC efficiency across a wide range of optimization problems within the diffusionary limit for which we have analytically known solutions. At the same time, defining the Ab Initio objective to be proper ensures, in principle, valid optimization behaviour in the opposite asymptotic limit of model complexity wherein i.i.d. resampling from the target lies within the proposal distribution’s model class. The Ab Initio objective is a sufficiently robust approximation to enable confident optimization of highly expressive neural MCMC proposal distributions.

| Reference Problem | Ab Initio | Analytic |
|-------------------|----------|----------|
| RWM               | 0.246(2) | 0.236(1) |
| MSJD              | 1.32(1) | 1.32(4)  |
| Scheme Robustness | Gaussian | Acc. Rate | 0.546(2) |
| MALA              | 38.4(2) | 38.4(2)  |
| Target Robustness | Uniform  | Acc. Rate | 0.135(2) |
| RWM               | 8.2(1e-3)| 8.2(1e-3)|
| Laplace           | 0.234(3) | 0.234(2) |
| Cauchy            | 0.237(3) | 0.235(2) |
| RWM               | 2.72(3) | 2.72(3)  |

**Dimensional Robustness**

VII. EXPERIMENTAL RESULTS

To demonstrate the advantages of Ab Initio objective functions, we present a series of illustrative experiments. In these experiments, we compare objective functions on the basis of their capabilities for optimizing very general parameterizations of proposal distributions that should be of interest for MCMC proposal optimization. Throughout these experiments, we use effective sample size (ESS) as an estimate of sampling performance. For reporting ESS performance, we follow the convention of [33] by calculating the one-dimensional ESS relating to uncertainty estimating first and second moments of the target distribution’s marginals and reporting the minimum such performance obtained across all dimensions. For comparisons, we consider MSJD optimization (MSJD Opt.) [28], L2HMC’s objective (L2HMC Obj.) [13], and the Generalized Speed Measure [27] targeting acceptance rates of 0.9, 0.6, and 0.3 (GSM-90, -60, and -30). As we are comparing the obtained optima of these objective functions, we estimate sampling from the target distribution by either direct sampling or by sampling form a long equilibrated HMC chain prepared before optimization, as opposed to the online, persistent single chain updates common used in many adaptive MCMC algorithms. Additional experimental and theoretical results are provided within the extended version of this work [30].

A. Optimizing Multi-Scheme Proposals

We now introduce a normalizing flow [34–36] based proposal. With \( T_\phi \) denoting a normalizing flow’s parameterized transformation, the functions \( \mu_L, \theta \), \( \mu_D, \theta \), \( \Sigma_L, \theta \), and \( \Sigma_D, \theta \) specified by neural networks, and \( \odot \) denoting element wise multiplication, our proposal distribution is defined via:

\[
\mathbf{n} \sim N(\mathbf{0} ; I) \\
\mathbf{x} | \mathbf{n} = \mu_L, \theta (T_\theta^{-1}(\mathbf{x})) + \Sigma_L, \theta (T_\theta^{-1}(\mathbf{x})) \odot \mathbf{n} \\
\mathbf{x} = \mu_D, \theta (\mathbf{x}) + \Sigma_D, \theta (\mathbf{x}) \odot T_\theta(\mathbf{x})
\]

This implementation is fundamentally a conditional normalizing flow [37] and is similar to NeuTra-HMC [10], which employs HMC within the latent space of a pre-fit normalizing flow to improve MCMC efficiency. This multi-scheme distribution differs from NeuTra-HMC by allowing both the flow’s transformation and the latent distribution to be optimized to improve MCMC efficiency, rather than being fixed from the start. This parameterization is, in an asymptotic limit of model complexity, able to approximate various existing MCMC schemes, including RWM, MALA, and i.i.d. resampling.

The target distribution is an equal mixture of 4 standard gaussians positioned in a cross formation with a maximal distance of 8 between component centers. We use the NICE architecture [38] to define the normalizing flow and vary the number of coupling layers (using either 8 or 3 layers), which influences how accurately the flow is able to approximate the target distribution. For each optimization\(^4\), we determine efficiency measures on the basis of 5 Markov Chains of 1000

\(^4\) For the multi-scheme proposal optimization experiment, we have \( N = 100, M = 8, \) and perform optimization for 20,000 gradient steps with a learning rate of 3e-4. Throughout this work, \( \mu_L, \theta, \mu_D, \theta, \Sigma_L, \theta \) and \( \Sigma_D, \theta \) are specified by standard ReLU networks. We specify these networks using the notation (input dimension \( I \), hidden dimension \( H \), hidden layers \( D \)), output dimension \( O \)). Here, \( \mu_L, \theta \), etc. have \( (I = 2, H = 16, D = 4, O = 2) \). Throughout this work, \( T_\theta \) follows the NICE architecture [38], with additive coupling layers specified by standard ReLU networks, specified using the notation (input dimension \( I \), hidden dimension \( H \), hidden layers \( D \)), output dimension \( O \)). For this experiment, the coupling layers follow \( (I = 1, H = 6, D = 3, O = 1) \).
Density plots of multi-scheme proposal distributions optimized using an MSJD objective and our Ab Initio objective are provided in Figure 1. Position dependence within the proposal distribution, leading to poor ESS performance.

This experiment demonstrates a fundamental difficulty in applying the GSM objective to multi-scheme proposals. Here, the flow’s depth determines it’s maximal accuracy in approximating independent resampling of the target distribution, creating a maximum acceptance rate that can be obtained while still approximating global resampling. Attaining greater acceptance rates is possible, but forces inefficient RWM-like behavior. In this case, 8 coupling layers within the flow permits a maximal resampling acceptance rate near 0.70, while 3 coupling layers lowers this maximal rate below 0.60. A user with 8 coupling layers who uses the GSM targeting an acceptance rate of 60% to optimize their distribution would fortuitously arrive at an efficient proposal distribution. However, if they had used a model with 3 coupling layers instead and targeted the same 60% acceptance rate, the result would be a proposal distribution far more inefficient than the actual capabilities of their proposal model class.

In both cases, the Ab Initio objective is able to optimize towards efficient approximations of i.i.d. resampling. As this experiment utilizes a complicated proposal distribution, it is reasonable to question whether the problems exhibited here arise from some peculiarities of the model class of Equation (3). Further optimization experiments involving the same objective functions and target distributions using instead position dependent mixtures of gaussians as proposal distributions yielded fundamentally similar results. We therefore conclude that the effects demonstrated within Table II may be traced back to the choice of objective function used for optimization.

### B. Optimizing Augmented Multi-Scheme Proposals

Augmentation with auxiliary variables (e.g. the momenta of HMC) is a common feature of MCMC schemes [12], [13] and, in this experiment, we augment the original distributions with a number of independent auxiliary variables equal to the distribution’s original dimensionality. Such augmentation further expands the model class of Equation (3) to enable the (asymptotic) approximation of MCMC schemes like HMC. To demonstrate the application of Ab Initio objectives to MCMC optimization tasks of a more practical nature, we consider the optimization of the multi-scheme proposals defined in Equation (3) for the sampling of regression weights in bayesian logistic regression for various UCI datasets [39]. For each optimization, we determine efficiency measures on the basis of 5 Markov Chains of 20000 total proposals starting from samples drawn from a long, equilibrated Markov Chain obtained using tuned HMC. Statistics regarding the mean and standard error of these measures are collected from 5 replications for each considered objective function.

The results of these optimizations, including comparisons of the computational costs for optimizing each objective, are given in Table II.

| Table II | Comparison of sampling performance obtained optimizing the multi-scheme proposals of Equation (3) using various objective functions. Results are reported for two choices for the number of coupling layers used in the proposal’s normalizing flow ($L = 8$ and $L = 3$). Value means are reported to at most the first significant digit of standard error (reported in parentheses). |
|----------|--------------------------------------------------------------------------------------------------|
| $L=8$   | Acc. Rate | MSJD | ESS per Prop | Grad per Sec | Acc. Rate | MSJD | ESS per Prop | Grad per Sec |
| Ab Initio, Eq. (2) | 0.67(3) | 22(1) | 0.33(5) | 7.6(1) | 0.55(4) | 17(1) | 0.13(8) | 10.6(1) |
| MSJD Opt. | 0.975(6) | 66.2(9) | 6.1(4) | 10.3(3) | 0.979(6) | 67(1) | 7.1(4) | 12.4(6) |
| L2HMC Obj. | 0.974(3) | 65.7(5) | 6.1(4) | 10.3(3) | 0.974(5) | 66(1) | 6.1(3e-4) | 12.6(4) |
| GSM-30 | 0.921(1) | 1.865 | 5.04(4) | 1.56(2) | 0.90(2) | 1.80(4) | 5.3(4e-4) | 1.56(2) |
| GSM-60 | 0.65(6) | 21(2) | 0.52(2) | 7.3(3) | 0.60(1) | 1.41(1) | 6(1e-4) | 3.4(4) |
| GSM-80 | 0.31(1) | 16.7(4) | 0.14(2) | 7.3(3) | 0.302(6) | 10.5(4) | 0.13(2) | 39.4(4) |
| I.I.D. Resample | 1.00 | 34.4(3) | 0.96(5) | 1.00 | 33.9(3) | 0.90(7) |

#### Notes

5 Calculations performed on a 2080 RTX GPU and a Xeon Gold 6152 CPU.

6 Augmentation is performed by appending $a$ dimensions to a $d$ dimensional dataset. For proposals, these $a$ augmenting dimensions are resampled according to a standard multivariate gaussian. Augmenting variables do not factor into the efficiency measures listed in Table III. For the augmented multi-scheme proposal optimization experiment, we have $N = 100$, $M = 8$, and perform optimization for 40,000 gradient steps with a learning rate of 3e-4, $\mu, \Sigma$, etc. have $(I = a + d, H = 3(a + d), D = 4, O = a + d)$. $T_0$ is specified by 5 coupling layers following $(I = a + d, H = 3(a + d), D = 4, O = \frac{a + d}{2})$. 

---
### TABLE III
Comparison of Sampling Performance Obtained Optimizing Augmented Multi-scheme Proposals of Equation (3) Using Various Objective Functions Targeting Posterior Distributions of Parameters for Logistic Regression of UCI Datasets. Value Means Are Reported to at Most the First Significant Digit of Standard Error (Reported in Parentheses).

| Objective Function | German Credit (d=21) | Heart Disease (d=14) |
|--------------------|----------------------|----------------------|
|                    | Acc. | MSJD | ESS | Grad. | Acc. | MSJD | ESS | Grad. |
| Ab Initio, Eq. (2) | 0.81(1) | 0.30(1) | 4.97(7) | 11.5(6) | 0.86(2) | 1.03(4) | 0.58(7) | 11.7(4) |
| MSJD Opt.          | 0.77(2) | 0.32(1) | 0.41(1) | 15.1(1) | 0.81(1) | 1.40(3) | 0.32(6) | 15.0(3) |
| L2HMC Multi        | 0.76(2) | 0.32(1) | 0.48(5) | 14.8(9) | 0.82(1) | 1.43(4) | 0.28(7) | 14.7(7) |
| I.I.D. Resample    | 1.00   | 0.3700(3) | 0.99(1) | 1.00   | 1.222(1) | 0.97(2) |

### TABLE IV
Comparison of Sampling Performances and Ab Initio Losses (Lower is Better) Obtained from Various MCMC Schemes Targeting the Posterior Distributions of the Parameters for the Logistic Regression of MNIST Digits (d=785). Value Means Are Reported to at Most the First Significant Digit of Standard Error (Reported in Parentheses).

| Model          | Obj. Func. | Acc. | MSJD | ESS | Grad. | Samples |
|----------------|------------|------|------|-----|-------|---------|
| Precond. MALA  | Eq. (2)    | 0.50(1) | 129(2) | 10(1) | 41(1) | 20(1) |
| Precond. MALA  | MSJD       | 0.54(3) | 241(1) | 1(1e-10) | 7(2e7) | 2(5e5) |
| Precond. RWM   | Eq. (2)    | 0.239(4) | 1.36(2) | 7(2e7) | 7(2e7) | 2(5e5) |
| Multi-Scheme.  | Eq. (3)    | 0.24(4) | 1.2(1) | 5(2e7) | 2(5e5) | 47(54) |

Equation (2) to be a robust measure of MCMC efficiency that may be used to evaluate and compare the efficiency of multiple MCMC schemes in addition to its utility as an objective function for optimizing the individual parameters of highly expressive MCMC schemes.

### VIII. Conclusion and Future Work
In this work, we have shown how to construct Ab Initio objective functions that are suited for the optimization of highly expressive proposal distributions. We find that Ab Initio objectives are suitably robust to proposal complexity to enable the optimization of highly expressive neural MCMC proposal distributions. Our experimental results show that Ab Initio objective functions can maintain favorable performance and preferable optimization behavior compared to existing objective functions [13], [27], [28]. By design, Ab Initio objective functions are approximations of our notion of MCMC efficiency and we do not presume that the particular example Ab Initio objective function of Equation (2) will be absolutely universal. However, should future experimental or theoretical analysis demonstrate some sub-optimal properties of Equation (2), our proposed Ab Initio procedure allows for further improvements by considering alternative component functions and coefficient fitting procedures, which we plan to explore in future work.

The experimental methodology this work is intended to isolate fundamental effects of the choice of objective function on MCMC optimization. How to best train proposals distributions in an online adaptive setting, how to attain computationally efficient samples, and how to handle scaling into higher dimensions remain important practical considerations for MCMC optimization. We argue that these considerations are primarily influenced by proposal architecture and training procedure selection. We will therefore investigate these questions in a future work focusing on comparisons among specific proposal architectures and training procedures.

### Acknowledgements and Disclosure of Funding
This work was supported by the Office of Naval Research (ONR) under grant number N00014-18-1-2244.

### References
[1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of State Calculations by Fast Computing Machines,” *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.
