Some Notes on the Sample Complexity of Approximate Channel Simulation

Gergely Flamich
University of Cambridge
Cambridge, UK
gf332@cam.ac.uk

Lennie Wells
University of Cambridge
Cambridge, UK
ww347@cam.ac.uk

***Abstract*—Channel simulation algorithms can efficiently encode random samples from a prescribed target distribution Q and find applications in machine learning-based lossy data compression. However, algorithms that encode exact samples usually have random runtime, limiting their applicability when a consistent encoding time is desirable. Thus, this paper considers approximate schemes with a fixed runtime instead. First, we strengthen a result of Agustsson and Theis [1] and show that there is a class of pairs of target distribution Q and coding distribution P, for which the runtime of any approximate scheme scales at least super-polynomially in $D_\infty[Q \parallel P]$. We then show, by contrast, that if we have access to an unnormalised Radon-Nikodym derivative $r \propto dQ/dP$ and knowledge of $D_{\mathrm{KL}}[Q \parallel P]$, we can exploit global-bound, depth-limited A* coding [2] to ensure $D_{TV}[Q \parallel P] \leq \epsilon$ and maintain optimal coding performance with a sample complexity of only $\exp_2\big((D_{\mathrm{KL}}[Q \parallel P] + o(1))/\epsilon\big)$.**

I. Introduction

One-shot channel simulation is a communication problem between two parties, Alice and Bob, who share a probabilistic model over a pair of correlated random variables $\mathbf{x}, \mathbf{y} \sim P_{\mathbf{x},\mathbf{y}}$, as well as a source of common randomness. In one round of communication, Alice receives a sample $\mathbf{y} \sim P_{\mathbf{y}}$ and needs to send the minimum number of bits to Bob, so that he can simulate a sample $\mathbf{x} \sim P_{\mathbf{x}|\mathbf{y}}$.

Solutions to this problem provide an alternative to quantization and entropy coding for implementing transform coding. Thus, efficient channel simulation protocols have far-reaching applications in machine learning-based data compression, as we can use them to turn essentially any generative model into a lossy compression algorithm [1], [3]–[6]. Furthermore, channel simulation provides unique advantages over quantization-based transform coding in many scenarios, such as when in addition to the rate-distortion trade-off we consider realism constraints [7], [8] or differential privacy [9].

Unfortunately, under a reasonable computational model of sampling, exact channel simulation in general is hopelessly difficult. Concretely, for a given $\mathbf{y} \sim P_{\mathbf{y}}$, let us set $P \leftarrow P_{\mathbf{x}}$ and $Q \leftarrow P_{\mathbf{x}|\mathbf{y}}$ for brevity. The standard computational model of channel simulation protocols assumes that Alice and Bob's shared randomness takes the form of an infinite sequence of i.i.d. P-distributed samples $(X_1, X_2, \ldots)$ and Alice has to select an index N, such that $X_N \sim Q$. Then, under mild assumptions on the selection rule, Goc and Flamich [10] show that the sample complexity, i.e. the number of samples Alice needs to examine on average from the sequence before she can determine N, is at least $\exp_2(D_\infty[Q \parallel P])$, where $D_\infty[Q \parallel P]$ is the Rényi ∞-divergence and $\exp_2(x) = 2^x$.

It is thus natural to ask whether relaxing the requirement that the law P_{X_N} of the selected sample be exactly Q could help reduce the computational complexity of channel simulation algorithms. This question is related to previous investigations by Chatterjee and Diaconis [11], Agustsson and Theis [1], and Block and Polyanskiy [12], who considered approximate sampling without regard for how efficiently the sample can be encoded. In this paper, we build on these works and strengthen some of their relevant results. Interestingly, we find that approximate channel simulation is not harder than approximate sampling. In fact, taking inspiration from the channel simulation literature we can improve sample complexity bounds in general, as we demonstrate in Section V.

Contributions. The goal of our paper is to determine the sample complexity required for approximate channel simulation under different computational assumptions. In particular,

1) We strengthen a result of Agustsson and Theis, and show that approximate sampling is prohibitively expensive for general distributions. Concretely, we show that under the standard complexity-theoretic assumption that $P \neq RP$, there is no algorithm whose runtime scales polynomially in $D_\infty[Q \parallel P]$ and which can output an approximate sample with law $\widetilde{Q}$ such that $D_{TV}[\widetilde{Q} \parallel Q] \leq 1/12$.
2) We give an improved variant of Block and Polyanskiy's approximate rejection sampler [12], that can achieve $D_{TV}[\widetilde{Q} \parallel Q] \leq \epsilon$ with a sample complexity of
$$\ln\left(\frac{1}{(1-\gamma)\epsilon}\right)(f')^{-1}\left(\frac{D_f[Q \parallel P]}{\gamma\epsilon}\right)$$
for any f-divergence and $\gamma \in (0, 1)$. While this might seem to contradict our first result, we clarify that this requires *exact* knowledge of $D_f[Q \parallel P]$ and dQ/dP.
3) We demonstrate that global-bound, depth-limited A* coding [2] can achieve $D_{TV}[\widetilde{Q} \parallel Q] \leq \epsilon$ error with a sample complexity of $\exp_2((D_{\mathrm{KL}}[Q \parallel P] + c)/\epsilon)$ for $c = e^{-1}\log_2 e + 1$.

II. Background

Notation. For two real numbers a, b we define the infix notation $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. We denote

the base two logarithm as $\log_2$ and its inverse function as $\exp_2$. Likewise, we denote the natural logarithm as $\ln$ and its inverse function as $\exp$. Let Q and P be probability measures over the measurable space $(\Omega, \mathcal{A})$; we will always assume that Ω is Polish. Then, we define their total variation distance as $D_{TV}[Q \parallel P] = \sup_{B\in\mathcal{A}}|Q(B) - P(B)|$. Furthermore, assuming $Q \ll P$ and denoting their Radon-Nikodym derivative as $r = dQ/dP$, we define the Kullback-Leibler divergence of Q from P as $D_{\mathrm{KL}}[Q \parallel P] = \mathbb{E}_{X\sim Q}[\log_2 r(X)]$ and the Rényi ∞-divergence as $D_\infty[Q \parallel P] = \log_2\|r\|_\infty$, where $\|\cdot\|_\infty$ is the P-essential supremum of a P-measurable function. Finally, let $\mathcal{F} = \{f : [0,\infty) \to \mathbb{R}_{\geq 0} \cup \{\infty\} \mid f \text{ convex}, f(1) = 0, f'(1) = 0\}$. Then, for an $f \in \mathcal{F}$ we define the f-divergence of Q from P as $D_f[Q \parallel P] = \mathbb{E}_{X\sim P}[f(r(X))]$.

One-shot channel simulation (OSCS), also known as relative entropy coding [4] or reverse channel coding [13], is a communication problem between two parties, Alice and Bob, and is defined as follows. Let $\mathbf{x}, \mathbf{y} \sim P_{\mathbf{x},\mathbf{y}}$ be a pair of random variables, whose law is known to both parties. Furthermore, we assume that Alice and Bob share a source of common randomness $\mathcal{S}$. In one round of channel simulation, Alice receives a symbol $\mathbf{y} \sim P_{\mathbf{y}}$ and sends the minimum number of bits to Bob such that he can simulate a sample $\mathbf{x} \sim P_{\mathbf{x}|\mathbf{y}}$ using $\mathcal{S}$. Surprisingly, it can be shown that Alice needs to send only $\mathbb{I}[\mathbf{x}\,;\,\mathbf{y}] + \log_2(\mathbb{I}[\mathbf{x}\,;\,\mathbf{y}] + 1) + 4.732$ bits on average to achieve this [14].

Selection samplers and sample complexity. In practice, we are also concerned with the encoding time of the channel simulation algorithm. However, we first need to establish a reasonable model of computation within which we can make sense of runtime. One could use notions from computational complexity theory, where runtime can be associated with the number of steps taken by a universal Turing machine that executes the sampling algorithm, we consider such a framework in Section IV. However, we are also interested in purely statistical or information theoretical properties of sampling; we focus on a natural formulation of sample complexity for a certain class of sampling algorithms. We define this class of algorithms below, taking inspiration from [10, Definition A.2].

Definition II.1 (Selection samplers)**.** Let $Q \ll P$ be probability measures over some space Ω. Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. P-distributed random variables. A selection sampler *selects*an index, modeled by some random variable N over $\mathbb{N}$, and returns a sample X_N. Moreover,

- if $\mathrm{Law}\,(X_N) = Q$ we say that the sampler is **exact**.
- if $D_{TV}[\mathrm{Law}\,(X_N) \parallel Q] \leq \epsilon$ for some $\epsilon > 0$ we say that the sampler is ϵ**-approximate**.

Finally, if there is a constant $k \in \mathbb{N}$ such that $N \leq k$ then we call k the sample complexity of the sampler. If there is a stopping time K adapted to the sequence $(X_i)_{i\in\mathbb{N}}$ such that $1 \leq N \leq K$ then we say that the sampler is A*-like and call $\mathbb{E}[K]$ its sample complexity.

Intuitively, an A*-like sampler needs to examine K proposal samples, after which it *has to* select one of the samples it already examined to output a sample from the target. We can now precisely state the result of Goc and Flamich [10] mentioned in the introduction: the sample complexity of any exact A*-like sampler is at least $\exp_2(D_\infty[Q \parallel P])$.

A coding.* We now briefly describe global-bound A* coding [2], the namesake of Definition II.1, as we will utilise it in Sections V and VI. The algorithm is equivalent to the Poisson functional representation [15] and is based on A* sampling [16]. Given the shared sequence $(X_i)_{i\in\mathbb{N}}$ of i.i.d. P-distributed samples and the (potentially unnormalized) Radon-Nikodym derivative $\widetilde{r} \propto dQ/dP$, A* coding selects the sample with index $N = \arg\max_{k\in\mathbb{N}}\{\ln\widetilde{r}(X_k) + G_k\}$, where G_1 is a Gumbel random variable with mean 0 and scale 1, and for $k > 1$ each $G_k \mid G_{k-1}$ is a standard Gumbel random variable truncated to $(-\infty, G_{k-1})$. While the maximisation to select N is over all positive integers, it can be shown [17] that it is sufficient to examine the first K elements of the sequence, where K is a geometric random variable with mean $\|\widetilde{r}\|_\infty$.

III. Why characterise approximate channel simulation using total variation?

A crucial detail to consider in studying approximate channel simulation protocols is how we ought to measure the approximation error. In this section, we present some arguments for why total variation distance is an appropriate choice.

Our main motivation stems from the following "one-shot" interpretation of the total variation distance. Let a sample X follow distribution Q or $\widetilde{Q}$ with probability $1/2$. Then the probability that an optimal observer can successfully tell which of the two distribution a given sample X follows is [18], [19]:

$$p_{\text{success}} = \frac{1}{2}D_{TV}[Q \parallel \widetilde{Q}] + \frac{1}{2}.$$

Identifying Q with the target distribution of a channel simulation algorithm and $\widetilde{Q}$ with its output distribution, we see the above characterisation aligns well with our goal: we are interested in the quality of a single encoded sample, as opposed to the quality of quantities derived from samples.

Furthermore, providing guarantees on the total variation integrates well with our main application of interest: lossy data compression with realism constraints, or the rate-distortion-perception trade-off [7], [19], [20]. For this argument, we briefly describe how lossy data compression is usually implemented using transform coding. We first encode a stochastic representation $\mathbf{x} \sim Q$ of some data $\mathbf{y} \sim P_{\mathbf{y}}$ using a channel simulation algorithm, and use some measurable transformation g (usually a neural network) to recover the data: $\mathbf{y}' = g(\mathbf{x})$. Now, in addition to the rate and the distortion of our compressor, we are also interested in controlling its *output distribution*, given by the pushforward measure $P_{\mathbf{y}'} = g_*Q$. In the usual, adversarial formulation of realism we require $D_{TV}[P_{\mathbf{y}} \parallel P_{\mathbf{y}'}] \leq \delta$, where $\delta = 0$ corresponds to perfect realism [7], [19].

Now, assume that we wish to use an ϵ-approximate scheme to encode a sample from Q instead of an exact one, resulting in a sample with distribution $\widetilde{Q}$ with $D_{TV}[\widetilde{Q} \parallel Q] \leq \epsilon$. Then the output distribution is given by $g_*\widetilde{Q}$, so the realism of our transform coder depends on both the transform g and the

channel simulation protocol we use. By applying the triangle inequality, we can bound the realism error as

$$\begin{aligned} D_{TV}[P_{\mathbf{y}} \parallel g_* \widetilde{Q}] &\leq D_{TV}[P_{\mathbf{y}} \parallel g_* Q] + D_{TV}[g_* \widetilde{Q} \parallel g_* Q] \\ &\leq D_{TV}[P_{\mathbf{y}} \parallel g_* Q] + D_{TV}[\widetilde{Q} \parallel Q] \\ &\leq \delta + \epsilon, \end{aligned}$$

where the second inequality uses the data processing inequality and the third applies the assumed bounds on the TV distances. The practical significance of this decomposition is that in machine-learning-based pipelines, where g is usually a neural network and we wish to learn its parameters, it provides principled justification for optimizing $D_{TV}[P_{\mathbf{y}} \parallel g_* Q]$, which can be done relatively easily using adversarial methods [21], instead of $D_{TV}[P_{\mathbf{y}} \parallel g_* \widetilde{Q}]$.

Finally, on a more technical note, we consider an error bound proposed by Chatterjee and Diaconis [11] for self-normalized importance sampling that has been adapted to minimal random coding [3] and A* coding [2]. To state the bound, write $\widetilde{Q}_n$ for the approximate distribution of the procedure; consider a sample complexity of $n = \exp_2(D_{\mathrm{KL}}[Q \parallel P] + t)$ for some $t \geq 0$; then for any measurable function f, we have

$$\mathbb{P}\left[\left|\mathbb{E}_{\widetilde{Z}\sim\widetilde{Q}_n}[f(\widetilde{Z})] - \mathbb{E}_{Z\sim Q}[f(Z)]\right| \geq \frac{2\|f\|\epsilon}{1-\epsilon}\right] \leq 2\epsilon, \quad (1)$$

$$\epsilon = \left(2^{\frac{-t}{2}} + 2\sqrt{\mathbb{P}_{X\sim Q}\left[\log_2 \frac{dQ}{dP}(X) \geq D_{\mathrm{KL}}[Q \parallel P] + \frac{t}{2}\right]}\right)^{\frac{1}{2}}$$

and $\|f\|$ denotes the $L^2(Q)$ norm of f. It might appear that the error ϵ vanishes exponentially quickly in the quantity t, which was introduced as an additive overhead to $D_{\mathrm{KL}}[Q \parallel P]$, at odds with the results from [1], [12]. To resolve this conflict, first note that this bound applies to a test function f rather than providing a one-shot guarantee on the approximate sample. Secondly, as noted in [11], the bound is only meaningful when $\log_2 \frac{dQ}{dP}(X)$ is concentrated; we present a simple calculation in Appendix B that shows that for certain pairs of distributions, ensuring a given tolerance in fact requires t to scale with $D_{\mathrm{KL}}[Q \parallel P]$, corresponding to a much larger sample size. In Appendix B we also present a (strict) strengthening of this bound inspired by ideas that we will develop in Section V.

IV. Improving a result of Agustsson and Theis

In this section, we strengthen the result of [1] on the computational hardness of approximate sampling. However, before we begin, we make some definitions that we will use later.

Definition IV.1 (Restricted Boltzmann Machine (RBM))**.** For some integer M, let $\Omega = \{0,1\}^M$ and let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^M$ and $\mathbf{A} \in \mathbb{R}^{M\times M}$. Then, the RBM distribution Q with parameters $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{A}\}$ is given by the probability mass function

$$q(\mathbf{z}) \propto \sum_{\mathbf{h}\in\Omega} \exp\left(\mathbf{a}^\top \mathbf{z} + \mathbf{h}^\top \mathbf{A}\mathbf{z} + \mathbf{b}^\top \mathbf{h}\right). \quad (2)$$

Definition IV.2 (Efficiently evaluatable representation)**.** For positive integers N, M a function $f : \{0,1\}^N \to \{0,1\}^M$ is an efficiently evaluatable representation of a distribution Q if:

1) f consits of a Boolean circuit of $\mathrm{poly}(M)$ size and $N = \mathrm{poly}(M)$ input bits and M output bits.
2) For $B \sim \mathrm{Unif}\left(\{0,1\}^N\right)$, we have $f(B) \sim Q$.

Theorem IV.1. Consider an algorithm which receives the parameters of an arbitrary RBM Q of problem size M as input and has access to an unlimited number of i.i.d. random variables $Z_n \sim P$, where P is the uniform measure over $\{0,1\}^M$. It outputs $\widetilde{Z} \sim \widetilde{Q}$ with $D_{TV}\left[\widetilde{Q} \,\middle\|\, Q\right] \leq 1/12$. If $RP \neq NP$, there is no such algorithm with $\mathrm{poly}(D_\infty[Q \parallel P])$ time complexity.

Proof. The proof follows the proof of Agustsson and Theis [1] mutatis mutandis, which we repeat here for completeness.

The high-level idea is that for such pairs of distributions $Q \ll P$, even evaluating their Radon-Nikodym derivative $r = dQ/dP$ is difficult. To this end, we make use of the result of Theorem 13 of Long and Servedio [22]:

Theorem IV.2. If $RP \neq NP$, then there is no polynomial-time algorithm with the following property: Given parameters $\theta = (\mathbf{A}, \mathbf{a}, \mathbf{b})$ as input, the algorithm outputs an efficiently evaluatable representation of a distribution whose total variation distance from an RBM with parameters θ is at most $1/12$.

Now, fix some positive integer M and fix some θ as given in Theorem IV.2. Assume there is an algorithm $\mathcal{A}$ which outputs $\widetilde{Z} \sim \widetilde{Q}$ in $\psi(D_\infty[Q \parallel P])$ steps with $D_{TV}\left[\widetilde{Q} \,\middle\|\, Q\right] \leq 1/12$ for some polynomial ψ. We compute

$$D_\infty[Q \parallel P] = \log_2 \max_{\mathbf{z}\in\Omega}\left\{\frac{q(\mathbf{z})}{2^{-M}}\right\} \leq \log_2\left\{\frac{1}{2^{-M}}\right\} = M\,. \quad (3)$$

Then by Equation (3) the computational complexity of $\mathcal{A}$ is at most $N = \psi(M)$. In that time, $\mathcal{A}$ can examine at most N random variables Z_n; since the input random variables are i.i.d., we can assume without loss of generality that $\mathcal{A}$ examines the first N of them. Since the proposal P is the uniform measure on $\{0,1\}^M$, these N variates correspond to an input of $M \cdot \psi(M) = \mathrm{poly}(M)$ uniformly random bits. This is an efficiently evaluatable representation of $\widetilde{Q}$, which contradicts Theorem IV.2, assuming $RP \neq NP$. □

V. Improving the scheme of Block and Polyanskiy

Recently, Block and Polyanskiy [12] have shown that by modifying rejection sampling [23], we can achieve ϵ error in total variation at a sample complexity of

$$k = \frac{2}{1-\epsilon} \ln\left(\frac{2}{\epsilon}\right) (f')^{-1}\left(\frac{4 \cdot D_f[Q \parallel P]}{\epsilon}\right) \vee 2. \quad (4)$$

In particular, as a special case of the above equation we get that $k = \mathcal{O}(\exp_2(4 \cdot D_{\mathrm{KL}}[Q \parallel P]/\epsilon))$. Note that, in general, we can have $D_{\mathrm{KL}}[Q \parallel P] \ll D_\infty[Q \parallel P]$ so this might seem at odds with Theorem IV.1. To reconcile these two results, we highlight that additional assumptions were needed in Block and Polyanskiy's scheme to achieve this improved sample complexity: we need to be able to compute $D_f[Q \parallel P]$ as well as evaluate dQ/dP *exactly*. By contrast these are not given in the setup of [1], and it is shown in [22, Theorem 8] that it

Algorithm 1: Depth-limited A* coding.

Input : Sequence $(X_i)_{i\in\mathbb{N}}$ of i.i.d. P-distributed samples, target Q defined via $\widetilde{r} \propto dQ/dP$, computational budget k

$N, Y, G_0, L \leftarrow (0, \perp, \infty, -\infty)$
for $i = 1$ **to** k **do**
  $G_i \sim \text{TruncGumbel}(0,1)|_{(-\infty, G_{i-1})}$
  if $L < \ln \widetilde{r}(X_i) + G_i$ **then**
    $L \leftarrow \ln \widetilde{r}(X_i) + G_i$
    $N, Y \leftarrow i, X_i$
  end
end
return N, Y

is computationally hard to even approximate the normalizing constants corresponding to densities in Equation (2) to within an exponentially large factor!

The approximate rejection sampler: We now describe the three key ideas of Block and Polyanskiy to modify rejection sampling [23] to get an approximate scheme for some target Q and proposal P:

1) They fix a budget $k \in \mathbb{N}$ and if the rejection sampler does not terminate, they pick one of the proposed samples at random. Denoting the output distribution of the budgeted sampler as $\widetilde{Q}$ and its termination step as K, the sampler's error is
$$D_{TV}[\widetilde{Q} \parallel Q] = \mathbb{P}[K > k] D_{TV}[Q \parallel P] \leq \mathbb{P}[K > k] \quad (5)$$
2) Since $\mathbb{P}[K > k]$ in Equation (5) depends on $D_\infty[Q \parallel P]$, they propose to use a truncated target Q_M, defined via its Radon-Nikodym derivative
$$\frac{dQ_M}{dP}(x) \propto \mathbf{1}\left[\frac{dQ}{dP}(x) \leq M\right] \cdot \frac{dQ}{dP}(x). \quad (6)$$
Then, using Q_M in the rejection sampler with a budget of k samples will have distribution $\widetilde{Q}_M$.
3) They show that for a fixed $\epsilon > 0$, setting
$$M = (f')^{-1}(4 \cdot D_f[Q \parallel P]/\epsilon)$$
and k as in Equation (4) yields $D_{TV}[\widetilde{Q}_M \parallel Q_M] \leq \epsilon/2$ and $D_{TV}[Q_M \parallel Q] \leq \epsilon/2$. Combining these two inequalities and applying the triangle inequality then yields the desired guarantee $D_{TV}\left[\widetilde{Q}_M \,\middle\|\, Q\right] \leq \epsilon$ at the sample complexity given in Equation (4).

We now propose several small improvements to this scheme. In this section, we focus on the sample complexity of the improved scheme and deal with encoding the approximate samples in Section VI.

Useful quantities and identities: inspired by [10], take $w_Q(h) = \mathbb{P}_{X\sim Q}[r(X) \geq h]$, $w_P(h) = \mathbb{P}_{X\sim P}[r(X) \geq h]$, and $W_P(h) = \int_0^h w_P(\eta)\, d\eta$, $S_P(h) = 1 - W_P(h)$. Note, that by Fubini,
$$\int_0^\infty \overbrace{\mathbb{P}_{X\sim P}[r(X) \geq h]}^{=w_P(h)}\, dh = \mathbb{E}_{X\sim P}[r(X)] = 1.$$
This, taken together with the fact that $w_P \geq 0$ shows that we can interpret it as the probability density of a random variable H. Thus, we can also interpret W_P as $W_P(h) = \mathbb{P}[H \leq h]$, and similarly $S_P(h) = \mathbb{P}[H > h]$ as H's survival function. Now, for any $f \in \mathcal{F}$ and $a \geq 1$, we can bound S_P by noting
$$\begin{aligned} f'(a)\mathbb{P}[H > a] &= f'(a) \int_a^\infty w_P(h)\, dh \\ &\leq \int_a^\infty f'(h) w_P(h)\, dh \\ &\leq \int_1^\infty f'(h) w_P(h)\, dh \\ &= \int_\Omega \left(\int_1^\infty f'(h)\mathbf{1}[r(x) \geq h] dh\right) dP(x) \\ &= \int_\Omega \mathbf{1}[r(x) \geq 1] f(r(x))\, dP(x) \\ &\leq D_f[Q \parallel P], \end{aligned}$$
where the second equality follows from Fubini (positive integrand), the third equality from the fundamental theorem of calculus, and we exploit $f \geq 0 = f(1)$. Rearranging, we get
$$S_P(a) \leq D_f[Q \parallel P]/f'(a). \quad (7)$$
Next, note that
$$\begin{aligned} W_P(h) &= \int_\Omega \int_0^h \mathbf{1}[r(x) \geq \eta]\, d\eta\, dP(x) \\ &= \int_\Omega r(x) \wedge h\, dP(x) \\ &= \int_{x: r(x) \geq h} h\, dP(x) + \int_{x: r(x) \leq h} r(x)\, dP(x) \\ &= h \cdot w_P(h) + (1 - w_Q(h)). \end{aligned} \quad (8)$$
Rearranging the terms, we find
$$S_P(h) = w_Q(h) - h \cdot w_P(h). \quad (9)$$

Better approximate target distribution for a tighter bound: The truncated target in Equation (6) is a rough approximation of Q. Instead, we propose a better approximation that will also lend itself to simpler analysis and will allow us to improve Equation (4). We define this through its density
$$\frac{dQ_M}{dP}(x) = r_M(x) = \frac{r(x) \wedge M}{W_P(M)}. \quad (10)$$
That $W_P(M)$ normalises $r(x) \wedge M$ follows from Equation (8). Now, set $\widetilde{M} = M/W_P(M)$. In Appendix A we show that
$$D_{TV}[Q_M \parallel Q] = S_P(\widetilde{M}) \overset{\text{eq. (7)}}{\leq} \frac{D_f[Q \parallel P]}{f'(\widetilde{M})} \quad (11)$$
We now substitute this improved truncation into Block and Polyanskiy's scheme to improve upon Equation (4). Let ϵ be given, and let $\gamma \in (0,1)$ a constant that we have yet to choose. By a standard result for rejection sampling, K is geometrically distributed with mean $\|r_M\|_\infty = M/W_P(M) = \widetilde{M}$. Thus,
$$\mathbb{P}[K > k] = \left(1 - \frac{1}{\widetilde{M}}\right)^k \leq \exp\left(-\frac{k}{\widetilde{M}}\right). \quad (12)$$

Setting $k \geq \widetilde{M} \ln(1/(1-\gamma)\epsilon)$ thus gives $D_{TV}[\widetilde{Q}_M \parallel Q_M] \leq (1-\gamma)\epsilon$. Furthermore, by Equation (11), setting

$$\widetilde{M} = (f')^{-1}\left(\frac{D_f[Q \parallel P]}{\gamma\epsilon}\right)$$

will guarantee $D_{TV}[Q_M \parallel Q] \leq \gamma\epsilon$. Combining these facts with the triangle inequality shows that for a fixed $\epsilon > 0$, a sample complexity of

$$k \geq \ln\left(\frac{1}{(1-\gamma)\epsilon}\right)(f')^{-1}\left(\frac{D_f[Q \parallel P]}{\gamma\epsilon}\right) \tag{13}$$

is sufficient to achieve ϵ error in the TV distance. Since the dependence on $(1-\gamma)$ is only logarithmic, the bound will often be tightest by taking γ very close to 1. This improves on Equation (4) by reducing the coefficient of $D_f[Q \parallel P]$ from 4 to a constant $1/\gamma$ that can be taken arbitrarily close to 1, removing the $2/(1-\epsilon)$ coefficient in the bound and additionally removing the requirement that k be at least 2. In practice, we can compute M from $\widetilde{M}$ by numerically inverting the function $h \mapsto h/W_P(h)$. This function is strictly increasing on a suitable domain by a simple calculation in Appendix A. In the rest of Appendix A, we also derive cleaner bounds for exact rejection sampling from Q_M (allowing random runtime), and show that our truncation is optimal in a certain sense.

Replacing rejection sampling by global-bound A sampling yields identical analysis.* Indeed, if we use the truncated target Q_M, the index K at which we can guarantee termination is again geometrically distributed with mean $\|r_M\|_\infty$ [17]. Remarkably, we in fact only need access to an unnormalised version of $\tilde{r}_M \propto r_M$ to run the algorithm, though we would also need knowledge of $D_f[Q \parallel P]$ to compute the sample complexity in Equation (13).

Depth-limited A sampling yields better bounds without needing truncation.* Using depth-limited A* sampling, as described in Algorithm 1, again only requires access to an unnormalised density ratio, but now is well behaved even when this density ratio is unbounded. Indeed, there is now no need for truncation; we next obtain tighter bounds by working with the original target distribution Q and basing the bounds on the index N of the accepted sample instead of the sampler's runtime K.

As before, we use Equation (5) to obtain the bound $D_{TV}[\widetilde{Q} \parallel Q] \leq \mathbb{P}[N > k]$ for a fixed sample complexity k. Now, note that by Markov's inequality, we have

$$\begin{aligned}\mathbb{P}[N > k] &= \mathbb{P}[\log_2 N > \log_2 k] \\ &\leq \frac{\mathbb{E}[\log_2 N]}{\log_2 k} \\ &\leq \frac{D_{\mathrm{KL}}[Q \parallel P] + e^{-1}\log_2 e + 1}{\log_2 k},\end{aligned} \tag{14}$$

where the last equality follows from the identity given in Appendix A of [15]. Note that the bound in Equation (14) does not depend on $\|r\|_\infty$. Therefore, for a fixed $\epsilon > 0$, choosing

$$k = \exp_2\left(\frac{D_{\mathrm{KL}}[Q \parallel P] + e^{-1}\log_2 e + 1}{\epsilon}\right) \tag{15}$$

yields the desired bound $D_{TV}[\widetilde{Q} \parallel Q] \leq \epsilon$. Equation (15) significantly improves Equation (4), and is close to being worst-case optimal [12, Theorem 5]. It also significantly improves a result of Theis and Yosri [13, Corollary 3.2], which requires that $k = \mathcal{O}(\exp_2(D_\infty[Q \parallel P]))$ to guarantee the same ϵ total variation error.

Remark. We can obtain almost-optimal ϵ-approximate sample complexities similar to Equation (15) by using ordered random coding [13], greedy Poisson rejection sampling [24] or greedy rejection coding [25], [26] and modifying the arguments that yield Equation (14) appropriately. Furthermore, via similar arguments we can also obtain linear-in-the-KL sample complexities for the branch-and-bound variants of these algorithms when Ω is one-dimensional and $\tilde{r}$ is unimodal.

VI. Approximate Channel Simulation Using the Improved Scheme

Unfortunately, an issue with Block and Polyanskiy's sampler is that it cannot be used for channel simulation. As we show in Appendix C, $\mathbb{H}[N] \geq 4 \cdot \mathbb{I}[\mathbf{x}\,;\,\mathbf{y}]/\epsilon$, which is much worse than the optimal upper bound of $\mathbb{I}[\mathbf{x}\,;\,\mathbf{y}]+\log_2(\mathbb{I}[\mathbf{x}\,;\,\mathbf{y}]+1)+\mathcal{O}(1)$.

Thankfully, this can also be fixed by replacing rejection sampling with A* coding. For some fixed target Q and proposal P, let N denote the index returned by A* coding, and let N' denote the index returned by Algorithm 1, i.e. depth-limited A* coding with given a budget of k samples. Then,

$$\begin{aligned}N' &= \operatorname*{arg\,max}_{n \in \{1,\ldots,k\}} \{\ln \widetilde{r}(X_k) + G_k\} \\ &\leq \operatorname*{arg\,max}_{k \in \mathbb{N}} \{\ln \widetilde{r}(X_k) + G_k\} = N\end{aligned}$$

Hence, we have $\mathbb{E}[\log_2 N'] \leq \mathbb{E}[\log_2 N]$. We encode N' using the ζ-distribution $\zeta(n \mid \lambda) \propto n^{-\lambda}$. Li and El Gamal [15] show that by setting $\lambda = 1 + 1/(\mathbb{I}[\mathbf{x}\,;\,\mathbf{y}] + e^{-1}\log_2 e + 1)$ we get

$$\mathbb{H}[N'] < \mathbb{I}[\mathbf{x}\,;\,\mathbf{y}] + \log_2(\mathbb{I}[\mathbf{x}\,;\,\mathbf{y}] + 1) + 4, \tag{16}$$

meaning we can encode our sample at the optimal rate.

VII. Discussion and Future Work

To summarise our results: ϵ-approximate sampling has computational complexity super-polynomial in $D_\infty[Q \parallel P]$ without further assumptions. When $D_{\mathrm{KL}}[Q \parallel P]$ is known, one can get ϵ-approximate samples with sample complexity exponential in $D_{\mathrm{KL}}[Q \parallel P]/\epsilon$ and we can also encode these samples at the optimal rate using Algorithm 1. It would be interesting to consider what other structural assumptions may be leveraged to achieve lower sample complexity.

Author contributions statement

GF conceived the core of this work, was responsible for the majority of the content, and wrote the entire first draft. LW helped GF craft the story of the paper, reworked the document, and contributed a number of insights, results and refinements.

Acknowledgments

GF acknowledges funding from DeepMind. LW is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/V52024X/1.

REFERENCES

[1] E. Agustsson and L. Theis, "Universally quantized neural compression," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[2] G. Flamich, S. Markou, and J. M. Hernández-Lobato, "Fast relative entropy coding with A* coding," in *International Conference on Machine Learning*, pp. 6548–6577, PMLR, 2022.

[3] M. Havasi, R. Peharz, and J. M. Hernández-Lobato, "Minimal random code learning: Getting bits back from compressed model parameters," in *International Conference on Learning Representations*, 2018.

[4] G. Flamich, M. Havasi, and J. M. Hernández-Lobato, "Compressing images by encoding their latent representations with relative entropy coding," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 16131–16141, 2020.

[5] Z. Guo, G. Flamich, J. He, Z. Chen, and J. M. Hernández-Lobato, "Compression with bayesian implicit neural representations," in *Advances in Neural Information Processing Systems*, 2023.

[6] J. He, G. Flamich, Z. Guo, and J. M. Hernández-Lobato, "Recombiner: Robust and enhanced compression with bayesian implicit neural representations," in *International Conference on Learning Representations*, 2024.

[7] L. Theis and E. Agustsson, "On the advantages of stochastic encoders," in *Neural Compression Workshop at ICLR*, 2021.

[8] L. Theis, T. Salimans, M. D. Hoffman, and F. Mentzer, "Lossy compression with Gaussian diffusion." arXiv:2206.08889, 2022.

[9] A. Shah, W.-N. Chen, J. Balle, P. Kairouz, and L. Theis, "Optimal compression of locally differentially private mechanisms," in *International Conference on Artificial Intelligence and Statistics*, pp. 7680–7723, PMLR, 2022.

[10] D. Goc and G. Flamich, "On channel simulation with causal rejection samplers." arXiv:2401.16579, 2024.

[11] S. Chatterjee and P. Diaconis, "The sample size required in importance sampling," *The Annals of Applied Probability*, vol. 28, no. 2, pp. 1099–1135, 2018.

[12] A. Block and Y. Polyanskiy, "The sample complexity of approximate rejection sampling with applications to smoothed online learning," *arXiv preprint arXiv:2302.04658*, 2023.

[13] L. Theis and N. Yosri, "Algorithms for the communication of samples," in *International Conference on Machine Learning*, 2022.

[14] C. T. Li and V. Anantharam, "A unified framework for one-shot achievability via the poisson matching lemma," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2624–2651, 2021.

[15] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967–6978, 2018.

[16] C. J. Maddison, D. Tarlow, and T. Minka, "A* sampling," *Advances in neural information processing systems*, vol. 27, 2014.

[17] C. J. Maddison, "A Poisson process model for Monte Carlo," *Perturbation, Optimization, and Statistics*, pp. 193–232, 2016.

[18] F. Nielsen, "Hypothesis testing, information divergence and computational geometry," in *International Conference on Geometric Science of Information*, pp. 241–248, Springer, 2013.

[19] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

[20] L. Theis, "What makes an image realistic?," *arXiv preprint arXiv:2403.04493*, 2024.

[21] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Advances in Neural Information Processing Systems*, 2020.

[22] P. M. Long and R. A. Servedio, "Restricted boltzmann machines are hard to approximately evaluate or simulate," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, (Madison, WI, USA), p. 703–710, Omnipress, 2010.

[23] J. von Neumann, "Various techniques used in connection with random digits," *Notes by GE Forsythe*, pp. 36–38, 1951.

[24] G. Flamich, "Greedy poisson rejection sampling," in *Advances in Neural Information Processing Systems*, 2023.

[25] G. Flamich and L. Theis, "Adaptive greedy rejection sampling," *arXiv preprint arXiv:2304.10407*, 2023.

[26] G. Flamich, S. Markou, and J. M. Hernández-Lobato, "Greedy rejection coding," in *Advances in Neural Information Processing Systems*, 2023.