Cactus Mechanisms: Optimal Differential Privacy Mechanisms in the Large-Composition Regime

Wael Alghamdi[⋆⋆], Shahab Asoodeh[‡], Flavio P. Calmon[⋆], Oliver Kosut[†], Lalitha Sankar[†], and Fei Wei[†]

***Abstract*—Most differential privacy mechanisms are applied (i.e., composed) numerous times on sensitive data. We study the design of optimal differential privacy mechanisms in the limit of a large number of compositions. As a consequence of the law of large numbers, in this regime the best privacy mechanism is the one that minimizes the Kullback-Leibler divergence between the conditional output distributions of the mechanism given two different inputs. We formulate an optimization problem to minimize this divergence subject to a cost constraint on the noise. We first prove that additive mechanisms are optimal. Since the optimization problem is infinite dimensional, it cannot be solved directly; nevertheless, we quantize the problem to derive near-optimal additive mechanisms that we call "cactus mechanisms" due to their shape. We show that our quantization approach can be arbitrarily close to an optimal mechanism. Surprisingly, for quadratic cost, the Gaussian mechanism is strictly sub-optimal compared to this cactus mechanism. Finally, we provide numerical results which indicate that cactus mechanisms outperform Gaussian and Laplace mechanisms for a finite number of compositions.**

The full proofs can be found in the appendices. This paper is Part I in a pair of papers, where Part II is [1].

I. Introduction

Likelihood ratios are at the heart of most privacy metrics. Consider the problem of quantifying the privacy loss suffered by a sensitive variable X given an observation of a disclosed variable Y. For example, X may represent a dataset and Y a randomized function computed over X. Privacy can be measured in terms of properties of the *privacy loss random variable*, defined as

$$L_{x,x'} \coloneqq \log \frac{dP_{Y|X=x}}{dP_{Y|X=x'}}(Y), \tag{1}$$

where $Y \sim P_{Y|X=x}$ and $x, x' \in \mathcal{X} \coloneqq \mathsf{supp}(X)$. The channel $P_{Y|X}$ is often referred to as a *privacy mechanism*.

Today, the most popular privacy definition (including, in practice [2]–[4]) is *differential privacy* (DP), which quantifies privacy in terms of $L_{x,x'}$ when x, x' are close or "neighboring." Thus, given a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $P_{Y|X}$ is said to be (ε, δ)-differentially private ((ε, δ)-DP) [5] if

$$\sup_{d(x,x')\le s} \sup_{A\subset\mathcal{Y}} \left[P_{Y|X=x}(A) - e^{\varepsilon} P_{Y|X=x'}(A)\right] \le \delta, \tag{2}$$

where s determines when inputs x and x' are neighboring, and $\mathcal{Y} \coloneqq \mathsf{supp}(Y)$. Intuitively, if a mechanism is (ε, δ)-differentially private for sufficiently small ε and δ, then an adversary observing Y cannot accurately distinguish between small changes in X.

Most privacy mechanisms are applied several times on sensitive data. Quantifying privacy guarantees under multiple *compositions* of a mechanism is a challenging problem. In the simple case where the same mechanism $P_{Y|X}$ is independently applied n times on data X generating output Y^n, i.e., $P_{Y^n|X} = \prod_{i=1}^n P_{Y_i|X}$, the privacy loss random variable is given by

$$L^n_{x,x'} \coloneqq \sum_{i=1}^n \log \frac{dP_{Y_i|X=x}}{dP_{Y_i|X=x'}}(Y_i), \tag{3}$$

where $Y_i \sim P_{Y_i|X=x}$. Differential privacy can be cast in terms of the privacy loss random variable. The reader can directly verify that n independent applications of a mechanism $P_{Y|X}$ is (ε, δ)-DP if

$$\sup_{d(x,x')\le s} \mathbb{E}\left[\left(1 - e^{-(L^n_{x,x'}-\varepsilon)}\right)^+\right] \le \delta. \tag{4}$$

From the law of large numbers, the distribution of $L^n_{x,x'}/n$ will concentrate around its mean, the KL-divergence, as

$$\frac{1}{n}\mathbb{E}\left[L^n_{x,x'}\right] = D\left(P_{Y|X=x} \| P_{Y|X=x'}\right). \tag{5}$$

Since the function $f(u) \coloneqq (1-e^{-nu+\varepsilon})^+$ is non-decreasing, in the limit of large compositions, privacy mechanisms with lower values of $D(P_{Y|X=x}\|P_{Y|X=x'})$ will enjoy stronger (ε, δ)-DP guarantees. Thus, regardless of the exact distribution of the privacy loss random variable, its mean (5) plays a central role in the privacy guarantees offered after many compositions. In applications such as privacy-ensuring machine learning, the number of compositions frequently exceeds $n = 10^3$.

We study the design of privacy mechanisms with favorable (ε, δ)-DP guarantees under a large number of compositions. Our approach departs from previous work in that we focus on the large-composition regime instead of optimizing (2). Since after many compositions, privacy will be mostly determined by the mean of the privacy loss random variable (5), we solve

⋆Corresponding author, remaining authors in alphabetical order. ⋆W. Alghamdi and F. P. Calmon are with the School of Engineering and Applied Science, Harvard University (emails: alghamdi@g.harvard.edu, flavio@seas.harvard.edu)

‡ S. Asoodeh is with the Department of Computing and Software, McMaster University (email: asoodehs@mcmaster.ca)

† O. Kosut, L. Sankar, and F. Wei are with the School of Electrical, Computer, and Energy Engineering, Arizona State University (emails: {okosut,lsankar,fwei16}@asu.edu)

This material is based upon work supported by the National Science Foundation under Grant Nos. CAREER-1845852, CIF-1900750, CIF-1815361, CIF-1901243, CIF-1908725, CIF-2007688, CIF-2134256, and CIF-2031799.

the optimization problem

$$\begin{aligned} \inf_{P_{Y|X}\in\mathscr{R}} \quad & \sup_{|x-x'|\le s} D(P_{Y|X=x}\|P_{Y|X=x'}) \\ \text{subject to} \quad & \sup_{x\in\mathbb{R}} \mathbb{E}[c(Y-x)\mid X=x]\le C, \end{aligned} \tag{6}$$

where $c:\mathbb{R}\to[0,\infty)$ is a pre-specified cost function, $s, C>0$ are constants, and $\mathscr{R}$ is the set of all Markov kernels on $\mathbb{R}$. Note that the cost function is critical: without the constraint, (6) can be trivially solved by any mechanism that is independent of X.

Our main contributions are as follows:

1) We show (Thm. 1) that additive mechanisms—i.e., where $Y = X + Z$ for a noise variable Z independent of X—suffice to minimize (6).
2) Even restricting to additive mechanisms, (6) is an infinite-dimensional optimization problem, so it cannot be solved directly. Instead, we formulate an approximate problem that is finite dimensional and can be solved efficiently. We prove (Thm. 3) that this approximate problem can get arbitrary close to optimal.
3) We solve the approximate problem to derive (near) optimal mechanisms for the quadratic cost function, i.e., $c(x)=x^2$. We dub the resulting mechanism the "cactus mechanism" due to the shape of the distribution (see Fig. 1). Surprisingly, the Gaussian distribution is strictly sub-optimal for (6), as the cactus mechanism achieves a smaller KL divergence for the same variance.
4) We bound the (ε,δ)-DP for the cactus mechanism in the context of sub-sampled stochastic gradient descent using the moments accountant method. Compared to the same analysis applied to a Gaussian mechanism, our approach does better for a reasonable number of compositions.

A. Related Work

Identifying optimal mechanisms is a fundamental and challenging problem in the domain of differential privacy. There have been several works in the literature that have attempted to address this problem. For instance, within the class of additive noise mechanisms and under the single shot setting (i.e., no composition), Ghosh et al. [6] showed that the geometric mechanism is universally optimal for $(\varepsilon,0)$-DP in a Bayesian framework, and Gupte and Sundararajan [7] derived the optimal noise distribution in a minimax cost framework. For a rather general cost function, the optimal noise distribution was shown to have a staircase-shaped density function [8]–[10]

Geng and Viswanath [11] showed that for (ε,δ)-DP and integer-valued query functions, in the single-shot setting, the discrete uniform noise distribution and the discrete Laplacian noise distribution are asymptotically optimal (for L^1 and L^2 costs) within a constant multiplicative gap in the high privacy regime (i.e., both ε and δ approach zero). Geng et al. [12] studied the same setting except for real-valued query functions and identified truncated Laplace distribution is asymptotically optimal in various high privacy regimes. Finally, Geng et al. [13] showed that the optimal noise distribution for real-valued query and $(0,\delta)$-DP is uniform with probability mass at the origin. Our work differs from these works in that we focus on the optimal mechanisms under a large number of compositions, rather than the single shot setting.

When considering a composition of n mechanisms, an important line of research has been to derive tighter composition results: relationships between the DP parameters of the composed mechanism and the parameters of each constituent mechanism. There are several composition results in the literature, such as [14]–[19]. More recently, Dong et al. [20] have proposed a composition result for large n and for a new variant of DP, called Gaussian-DP, that leverages the central limit theorem. These results can be sub-optimal (see, for example, [21, Fig. 1]). Consequently, numerical composition results have gained increasing traction as they lead to easier, yet powerful, methods for accounting the privacy loss in composition [21]–[24]. In particular, Koskela et al. [22] obtained a numerical composition result based on a numerical approximation of an integral that gives the DP parameters of the composed mechanism. The approximation is carried out by discretizing the integral and by evaluating discrete convolutions via the fast Fourier transform algorithm. The running time and memory needed for this approximation were subsequently improved [21]. While our work shares the focus on the large composition regime, we are primarily interested in synthesizing optimal mechanisms rather than analyzing existing mechanisms.

B. Notation

The Lebesgue measure on $\mathbb{R}$ is denoted by λ. We denote by $\mathscr{R}$ the set of all Markov kernels[1] on $\mathbb{R}$, i.e., conditional distributions $P_{Y|X}$ for $\mathbb{R}$-valued X and Y such that $x\mapsto P_{Y|X=x}(B)$ is a Borel function for all Borel sets $B\subset\mathbb{R}$. The set $\mathscr{B}$ denotes all Borel probability measures on $\mathbb{R}$. We fix a real-valued random variable X throughout, and let $P_X\in\mathscr{B}$ be its induced Borel probability measure. The KL-divergence is denoted by $D(P\|Q)$, and also by $D(p\|q)$ if $P,Q\ll\lambda$ with densities p and q. The expectation is denoted by $\mathbb{E}_P[f]:=\int_{\mathbb{R}} f\,dP$, and also by $\mathbb{E}_p[f]$ if $P\ll\lambda$ has probability density function (PDF) p. We let T_a denote the shift operator, i.e., for a function f of a real variable the function $T_a f$ is defined as $(T_a f)(x):=f(x-a)$, and for a measure P the measure $T_a P$ is defined by $(T_a P)(B):=P(B-a)$.

II. Optimality of Additive Continuous Channels

We start by deriving characterizations of solutions to the optimization problem (6). The difficulty of this problem lies in the fact that we are optimizing over all conditional distributions. This not only makes the problem infinite-dimensional, but it also renders direct approaches ineffective. The main result of this section, shown in Theorem 1, is that it suffices to consider continuous additive channels. In other words, the optimization in (6) may be restricted to conditional distributions of the form $P_{Y|X=x}=T_xP$ for some Borel probability measure P on $\mathbb{R}$ that is absolutely continuous with respect to the Lebesgue measure. Equipped with this reduction, we build in the next section an explicit family of finitely-parametrized distributions that are also optimal in (6).

[1]It is true that any conditional distribution from $\mathbb{R}$ into $\mathbb{R}$ has a version that is a Markov kernel [25, Chapter 4, Theorem 2.10].

A. Assumptions and Definitions

Throughout the paper, we require the cost function to satisfy the following properties.

Assumption 1. *The cost function $c : \mathbb{R} \to \mathbb{R}$ satisfies:*

- Positivity: $c(x) \geq 0$ *for all* $x \in \mathbb{R}$, *and* $c(0) = 0$.
- Symmetry: $c(x) = c(-x)$ *for all* $x \in \mathbb{R}$.
- Monotonicity: $c(x) \leq c(x')$ *if* $|x| \leq |x'|$.
- Continuity: c *is continuous over* $\mathbb{R}$.
- Tail regularity: *There exist* $\alpha, \beta > 0$ *such that* $c(x) \sim \beta x^{\alpha}$ *as* $x \to \infty$.

A natural choice of cost function is the quadratic cost $c(x) = x^2$, but we allow $c(x)$ to be any function that satisfies the above assumptions. For example, $c(x) = |x|^{\alpha}$ for any positive α is a natural family of cost functions.

Let $\mathscr{P} \subset \mathscr{R}$ be the set of conditional distributions $P_{Y|X}$ satisfying the cost constraint in (6), i.e., set

$$\mathscr{P} := \left\{ P_{Y|X} \in \mathscr{R} \; ; \; \sup_{x \in \mathbb{R}} \mathbb{E}[c(Y - x) \mid X = x] \leq C \right\}. \tag{7}$$

The infimal value in (6) is then

$$\mathrm{KL}^{\star} := \inf_{P_{Y|X} \in \mathscr{P}} \sup_{x, x' \in \mathbb{R} : |x - x'| \leq s} D(P_{Y|X=x} \| P_{Y|X=x'}). \tag{8}$$

We are interested in computing $\mathrm{KL}^{\star}$, as well as mechanisms $P_{Y|X}$ that approach this optimal value. Note that, for clarity of presentation, we suppress the dependence on (s, c, C) in the notations $\mathscr{P}$ and $\mathrm{KL}^{\star}$.

In the main problem (6), we allow $P_{Y|X}$ to be any mechanism that produces Y given X. A more restrictive but natural and easy-to-implement class of mechanisms is the *additive* mechanism class. An additive mechanism is given by $P_{Y|X=x}(B) = T_x P(B)$ where P is a Borel probability measure on $\mathbb{R}$. In other words, an additive mechanism $P_{Y|X}$ has Y of the form $Y = X + Z$ for some noise random variable $Z \sim P \in \mathscr{B}$ that is independent of the input X. Let $\mathscr{P}_{\text{add}} \subset \mathscr{B}$ be the set of additive mechanisms satisfying the cost constraint in (6),

$$\mathscr{P}_{\text{add}} := \{ P \in \mathscr{B} \; ; \; \mathbb{E}_P[c] \leq C \}. \tag{9}$$

Since the KL-divergence is shift-invariant, restricting the optimization (6) to additive mechanisms amounts to considering the simplified optimization problem

$$\mathrm{KL}^{\star}_{\text{add}} := \inf_{P \in \mathscr{P}_{\text{add}}} \sup_{a \in \mathbb{R} : |a| \leq s} D(P \| T_a P). \tag{10}$$

Of course, it is immediate that $\mathrm{KL}^{\star} \leq \mathrm{KL}^{\star}_{\text{add}}$. In fact, we will show below that these quantities are the same, meaning that there is no loss in restricting to additive mechanisms.

B. Optimality of Continuous Additive Mechanisms

The optimization problem in (6) is a convex problem, but the fact that the feasible set $\mathscr{P}$ is of infinite dimension means it cannot be solved directly, nor do the tractable properties one expects of a convex optimization problem necessarily follow. For example, in any finite dimensional convex optimization problem, a symmetry in the problem leads to the same symmetry in the solution. In this problem, one can see that shifting the mechanism—i.e., given $P_{Y|X}$, construct $Q_{Y|X=x}(B) = P_{Y|X=x+z}(B + z)$ for some z—does not change the cost constraint nor the objective value in (6). Thus, one might be inclined to conclude that the optimal mechanism is invariant to a shift (i.e., is an additive mechanism). Unfortunately, the infinite-dimensional nature of the problem means that this conclusion is not immediate. We resolve this issue in the following theorem which states that additive mechanisms are in fact optimal in (6).

Theorem 1. *We have that*

$$\mathrm{KL}^{\star} = \mathrm{KL}^{\star}_{\text{add}}, \tag{11}$$

and there exists a $P^{\star} \in \mathscr{P}_{\text{add}}$ *achieving this value. Further, any such* $P^{\star}$ *is necessarily absolutely continuous.*

Proof sketch: The proof is given in the extended paper [26, Appendix A]. We give here only a high level description of the approach. Let $P^{(k)}_{Y|X}$ be a sequence achieving $\mathrm{KL}^{\star}$. We make these mechanisms increasingly closer to being additive, while sacrificing neither feasibility nor utility, by considering the convex combinations

$$\overline{P}^{(k)}_{Y|X=x}(B) := \mathbb{E}\left[P^{(k)}_{Y|X=x+Z_k}(B + Z_k) \right] \tag{12}$$

where $Z_k \sim \mathrm{Unif}([-k, k])$. Specifically, one can invoke Prokhorov's theorem on the $\overline{P}^{(k)}_{Y|X}$, thereby extracting a probability measure $P^{\star}$ such that $\overline{P}^{(k)}_{Y|X=x} \to T_x P^{\star}$ weakly for each fixed x. Finally, we show that the mechanism $P^{\star}$ is optimal by invoking joint convexity and lower-semicontinuity of the KL-divergence. ∎

Remark 1. The proof of $P^{\star} \ll \lambda$ only relies on the property that $P^{\star} \ll T_a P^{\star}$ for every $|a| \leq s$, which holds in view of $\mathrm{KL}^{\star} < \infty$. Therefore, any *feasible* additive mechanism must be absolutely continuous with respect to the Lebesgue measure, i.e., if $\mu \in \mathscr{B}$ satisfies $\sup_{|a| \leq s} D(\mu \| T_a \mu) < \infty$ then we necessarily have $\mu \ll \lambda$.

III. Numerical Approximation: The Cactus Distribution

The optimization problem over additive mechanisms in (10) is infinite-dimensional, so it cannot be solved numerically as-is, and it appears to have no closed-form solution for non-trivial cost functions. The lack of closed-form solution is true even for the simple case of $c(x) = x^2$: to our surprise, as will be illustrated later, the Gaussian mechanism is not optimal![2] In our companion paper [1], we explore the regime where $s \to 0^+$; in this limit, we show that the optimal distribution can be determined exactly, and in fact for quadratic cost the limiting optimal distribution *is* Gaussian—although for other costs the optimal distribution is much more surprising.

In the regime of fixed positive s, to find practically achievable near-optimal mechanisms, we resort to numerical approximation of (10). In this section, we fix $s = 1$. We can do

[2]Of course, simply because Gaussian is not optimal does not imply that there is no closed-form solution. It is possible to write a set of KKT conditions for (10), which we have omitted from this paper in the interest of space. This set of KKT conditions cannot be solved in closed-form.

this without loss of generality simply by scaling: that is, the optimization problem in (10) with sensitivity s and cost function $c(x)$ is equivalent to the same problem with sensitivity 1 and cost function $c(sx)$.

To approximate (10) by a numerically tractable problem, we (i) quantize the distribution, and (ii) only explicitly parameterize the distribution in a certain interval. Specifically, we construct a mapping from finite-length vectors to continuous distributions as follows.

Definition 1. Fix two positive integers n and N, and a constant $r \in (0,1)$. Consider the partition of $\mathbb{R}$ by intervals $\{\mathcal{J}_{n,i}\}_{i\in\mathbb{Z}}$ defined by: $\mathcal{J}_{n,0} := [-1/(2n), 1/(2n)]$ and

$$\mathcal{J}_{n,i} := \begin{cases} \left(\frac{i-1/2}{n}, \frac{i+1/2}{n}\right], & \text{if } i > 0, \\ \left[\frac{i-1/2}{n}, \frac{i+1/2}{n}\right), & \text{if } i < 0. \end{cases} \tag{13}$$

We associate to each vector $\boldsymbol{p} = (p_0, p_1, \ldots, p_N) \in [0,1]^{N+1}$ a piecewise constant function that is defined by

$$f_{n,r,\boldsymbol{p}}(x) = \begin{cases} np_{|i|}, & \text{if } x \in \mathcal{J}_{n,i}, \text{with } |i| < N, \\ np_N r^{|i|-N}, & \text{if } x \in \mathcal{J}_{n,i}, \text{with } |i| \geq N. \end{cases} \tag{14}$$

We also associate with $f_{n,r,\boldsymbol{p}}$ the Borel measure $P_{n,r,\boldsymbol{p}}$, where

$$P_{n,r,\boldsymbol{p}}(B) := \int_B f_{n,r,\boldsymbol{p}}(x)\,dx. \tag{15}$$

Remark 2. Note that

$$\int_{\mathbb{R}} f_{n,r,\boldsymbol{p}}(x)\,dx = p_0 + \sum_{i=1}^{N-1} 2p_i + \frac{2p_N}{1-r} =: S_{r,\boldsymbol{p}}. \tag{16}$$

If $S_{r,\boldsymbol{p}} = 1$, then $P_{n,r,\boldsymbol{p}}$ is a probability measure with density $f_{n,r,\boldsymbol{p}}$. This distribution is symmetric around the origin, i.e., $f_{n,r,\boldsymbol{p}}(x) = f_{n,r,\boldsymbol{p}}(-x)$. Further, its tails decay almost geometrically: for $(N+1/2)/n < x_1 < x_2$ one has $f_{n,r,\boldsymbol{p}}(x_2) = r^{nk} f_{n,r,\boldsymbol{p}}(x_1)$ where $k = \left(\lceil nx_2 + 1/2\rceil - \lceil nx_1 + 1/2\rceil\right)/n \approx x_2 - x_1$.

The main results of this section are: we show that the distribution family introduced in Definition 1 is optimal for (6), and we show that the optimal distribution *within* this family (which we will call the *cactus distribution*) is obtainable via a tractable finite-dimensional convex optimization problem.

We use the following notation. Consider the restriction of (10) to the mechanisms constructible by Definition 1. For a fixed triplet $(n, N, r) \in \mathbb{N}^2 \times (0,1)$, denote the set of such mechanisms by $\mathscr{C}_{n,N,r} \subset \mathscr{B}$, i.e.,

$$\mathscr{C}_{n,N,r} := \left\{P_{n,r,\boldsymbol{p}} \;;\; \boldsymbol{p} \in [0,1]^{N+1}, S_{r,\boldsymbol{p}} = 1\right\}. \tag{17}$$

(Recall the definition of $S_{r,\boldsymbol{p}}$ from (16).) Denote the optimal value achievable by the class $\mathscr{C}_{n,N,r}$ by

$$\mathrm{KL}^\star_{n,N,r}(C) := \inf_{\substack{P\in\mathscr{C}_{n,N,r} \\ \mathbb{E}_P[c]\leq C}} \sup_{|a|\leq 1} D(P\|T_a P). \tag{18}$$

We show next that we may restrict the shift a in (18) to take values over the finite set $\{1/n, 2/n, \cdots, 1\}$ (rather than varying over the whole interval $[-1,1]$), thereby rendering (18) a finite-dimensional optimization problem amenable to standard numerical convex-programming methods. For each $i \in \mathbb{Z}$, we denote the constants

$$c_{n,i} := \int_{\mathcal{J}_{n,i}} nc(x)\,dx. \tag{19}$$

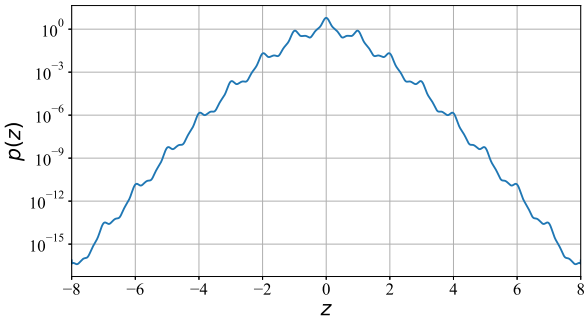


Fig. 1: The optimal distribution $p(z)$, found by solving (20) (and dubbed the *cactus distribution*), plotted on a semi-log scale. The cost function is $c(z) = z^2$, and the parameters are: $s = 1$, $C = 0.25$, $n = 200$, $N = 1600$, and $r = 0.9$.

Theorem 2. *Fix $r \in (0,1)$, and positive integers $n < N$. The minimization (18) can be recast as the following convex program over the variable $\boldsymbol{p} = (p_0, \cdots, p_N) \in \mathbb{R}^{N+1}$*

$$\begin{aligned} \underset{\boldsymbol{p}}{\text{minimize}} \quad & \max_{k\in\{1,\ldots,n\}} \frac{1}{2} \sum_{i=-N+1}^{N-k-1} (p_{|i|} - p_{|i+k|}) \log \frac{p_{|i|}}{p_{|i+k|}} \\ & + \sum_{i=N-k}^{N-1} (p_i - p_N r^{i+k-N}) \log \frac{p_i}{p_N r^{i+k-N}} \\ & + p_N \frac{1-r^k}{1-r} k \log r^{-1} \\ \text{subject to} \quad & p_0 c_{n,0} + \sum_{i=1}^{N-1} 2p_i c_{n,i} + 2p_N \sum_{i=N}^{\infty} c_{n,i} r^{i-N} \leq C, \\ & p_0 + \sum_{i=1}^{N-1} 2p_i + \frac{2p_N}{1-r} = 1, \\ & p_i \geq 0 \text{ for all } i \in \{0,\ldots,N\}. \end{aligned} \tag{20}$$

Figure 1 shows an example of the distribution that results from the finite-dimensional optimization problem in (20) with a quadratic cost. The shape of this distribution[3] has inspired the name the "cactus distribution."

The following result shows that cactus mechanisms derived from the optimization problem (20) are in fact globally optimal for the main optimization problem (6).

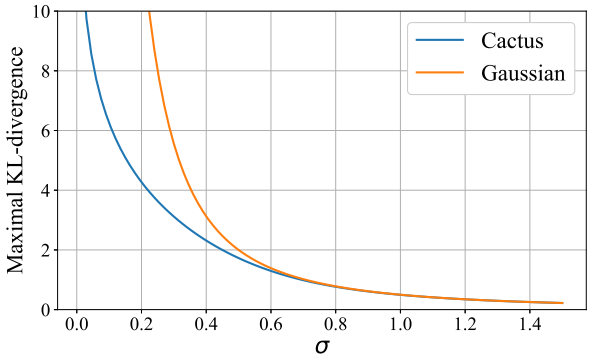
Theorem 3. *Denote the optimal value a cactus distribution can achieve by*

$$\mathrm{KL}^\star_{\text{Cactus}} := \lim_{\varepsilon\to 0^+} \inf_{(n,N,r)\in\mathbb{N}^2\times(0,1)} \mathrm{KL}^\star_{n,N,r}(C+\varepsilon). \tag{21}$$

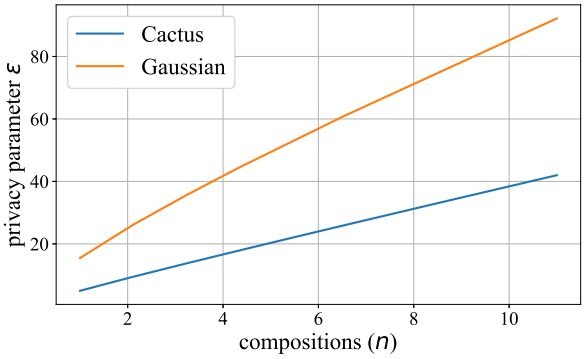
We have that $\mathrm{KL}^\star = \mathrm{KL}^\star_{\text{Cactus}}$.

Remark 3. The proof of Theorem 3 gives some guidelines for choosing the parameters (n, N, r). For example, optimal

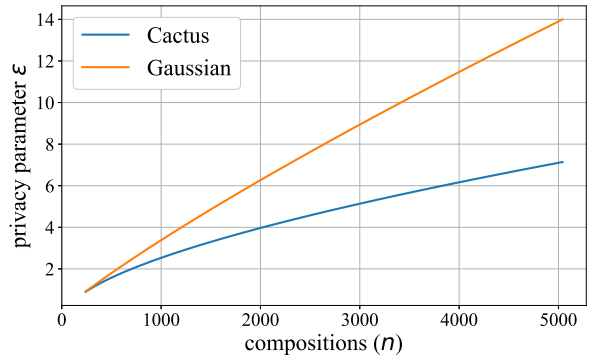
[3] In addition to the state of Arizona being home of several of the authors.



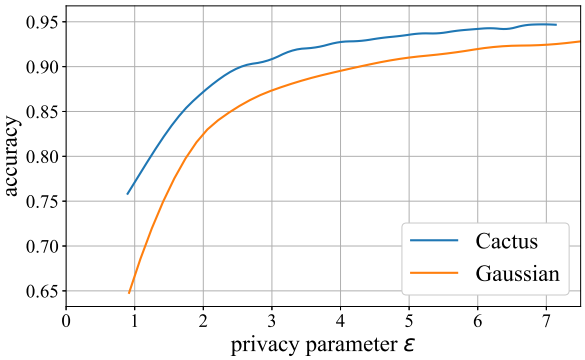
(a) Achieved maximal KL-divergence $\sup_{|a|\le s} D(p\|T_a p)$ versus σ, the (quadratic) cost constraint is of the form $\mathbb{E}[Z^2] \le \sigma^2 = C$ with fixed sensitivity $s = 1$.



(b) Privacy parameter ε versus the number of compositions, computed via the moments accountant, where $\delta = 10^{-3}$, and quadratic cost $C = 0.1$ with fixed sensitivity $s = 1$.



(c) Privacy parameter ε versus the number of compositions, computed via the moments accountant, where $\delta = 10^{-5}$, subsampling rate $q \approx 0.00417$, and quadratic cost $C = 0.1$ with fixed sensitivity $s = 1$.



(d) Model accuracy versus privacy parameter ε. The settings are the same as in Figure 2c and experiment details are given in Section IV.

Fig. 2: Comparison between the Gaussian and cactus mechanisms.

cactus distributions can be obtained by restricting the ratio N/n (chosen sufficiently large), and choosing $r = 1 - \Theta_\alpha(N^{-1})$.

IV. Numerical Results

We solve the optimization problem (20) using an interior-point method. An example of the cactus distribution for quadratic cost is shown in Figure 1. Figure 2a compares the maximal KL-divergence achieved by the cactus to that of Gaussian distributions for fixed sensitivity $s = 1$ and various σ. As noted above, varying σ with fixed s is equivalent to varying s with fixed σ. The KL-divergence for cactus is computed numerically, and for Gaussian mechanisms the KL-divergence is exactly $\frac{1}{2\sigma^2}$. The cactus distribution outperforms the Gaussian distribution in terms of KL-divergence for all values of σ, although the difference decreases as σ grows such that for larger values of σ it is difficult to discern any gap between the curves in Figure 2a. (Our companion paper [1] gives a theoretical explanation for why Gaussian is so close to optimal as s/σ decreases.) To illustrate that this improvement in KL-divergence leads to an improvement in (ε, δ)-DP, we compute the achieved privacy via moments accountant [17] for each mechanism. Figure 2b shows the resulting ε value as a function of the number of compositions, for fixed $\delta = 10^{-3}$. Indeed, the cactus mechanism does better than Gaussian.

To give a reasonable comparison in the context of machine learning, we modified the tutorial code in TensorFlow-Privacy [27], which implements the DP-stochastic gradient descent (SGD) algorithm with a Gaussian mechanism on a convolutional neural network (CNN) model. We use the training results from the original tutorial as a benchmark, then replace the Gaussian mechanism with our cactus mechanism, and train the model using the renewed setting. We select a noise level $\sigma = \sqrt{0.1}$. We test the original and modified model on a popular image dataset, MNIST, which is of size 60000. We choose a batch-size 250, such that each epoch consists of 240 iterations (i.e., compositions) and the sub-sampling rate[4] is $q = 250/60000 \approx 0.00417$. Figure 2c shows the achieved (ε, δ)-DP as computed by the moments account in this setting. Fixing $\delta = 10^{-5}$, Figure 2d shows the tradeoff between privacy ε and accuracy of the resulting CNN as the number of training iterations increases. One can see that for a fixed privacy budget (i.e., fixed ε and δ), the cactus mechanism allows more training iterations and, thus, better accuracy.

[4]The cactus mechanism is not optimized for subsampling. Nevertheless, we observe numerical performance of the cactus mechanism in the subsampling setting outperforming that of the Gaussian mechanism.

REFERENCES

[1] W. Alghamdi, S. Asoodeh, F. Calmon, O. Kosut, L. Sankar, and F. Wei, "Schrödinger mechanisms: Optimal differential privacy mechanisms for small sensitivity," 2022. [Online]. Available: https://github.com/WaelAlghamdi/DP-Schrodinger

[2] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. ACM, 2014, pp. 1054–1067.

[3] Differential privacy team Apple, "Learning with privacy at scale," 2017.

[4] D. Kifer, S. Messing, A. Roth, A. Thakurta, and D. Zhang, "Guidelines for implementing and auditing differentially private systems," *ArXiv*, vol. abs/2002.04049, 2020.

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory of Cryptography (TCC)*, Berlin, Heidelberg, 2006, pp. 265–284.

[6] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," *SIAM Journal on Computing*, vol. 41, no. 6, pp. 1673–1693, 2012.

[7] M. Gupte and M. Sundararajan, "Universally optimal privacy mechanisms for minimax agents," in *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2010, p. 135–146.

[8] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 925–951, 2015.

[9] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1176–1184, 2015.

[10] J. Soria-Comas and J. Domingo-Ferrer, "Optimal data-independent noise for differential privacy," *Information Sciences*, vol. 250, no. Complete, pp. 200–214, 2013.

[11] Q. Geng and P. Viswanath, "Optimal noise adding mechanisms for approximate differential privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 952–969, 2016.

[12] Q. Geng, W. Ding, R. Guo, and S. Kumar, "Tight analysis of privacy and utility tradeoff in approximate differential privacy," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108, 2020, pp. 89–99.

[13] ——, "Optimal noise-adding mechanism in additive differential privacy," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 11–20. [Online]. Available: https://proceedings.mlr.press/v89/geng19a.html

[14] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 51–60.

[15] J. Murtagh and S. Vadhan, "The complexity of computing the optimal composition of differential privacy," in *Proc. Int. Conf. Theory of Cryptography*, 2016, pp. 157–175.

[16] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., vol. 37, 2015, pp. 1376–1385.

[17] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[18] S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar, "Three variants of differential privacy: Lossless conversion and applications," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 208–222, 2021.

[19] S. Meiser and E. Mohammadi, "Tight on budget? tight bounds for r-fold approximate differential privacy," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18, 2018, p. 247–264.

[20] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *arXiv preprint arXiv:1905.02383*, 2019.

[21] S. Gopi, Y. T. Lee, and L. Wutschitz, "Numerical composition of differential privacy," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[22] A. Koskela, J. Jälkö, and A. Honkela, "Computing tight differential privacy guarantees using fft," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2560–2569.

[23] A. Koskela, J. Jälkö, L. Prediger, and A. Honkela, "Tight differential privacy for discrete-valued mechanisms and for the subsampled gaussian mechanism using fft," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 3358–3366. [Online]. Available: https://proceedings.mlr.press/v130/koskela21a.html

[24] Y. Zhu, J. Dong, and Y.-X. Wang, "Optimal accounting of differential privacy via characteristic function," *arXiv preprint arXiv:2106.08567*, 2021.

[25] E. Çınlar, *Probability and Stochastics*. New York, NY: Springer, 2011.

[26] W. Alghamdi, S. Asoodeh, F. Calmon, O. Kosut, L. Sankar, and F. Wei, "Cactus mechanisms: Optimal differential privacy mechanisms in the large-composition regime," 2022. [Online]. Available: https://github.com/WaelAlghamdi/DP-Cactus

[27] TensorFlow-Privacy tutorial, https://github.com/tensorflow/privacy.git/.

[28] E. Posner, "Random coding strategies for minimum entropy," *IEEE Transactions on Information Theory*, vol. 21, no. 4, pp. 388–391, 1975.

[29] V. I. Bogachev, *Measure theory*. Berlin: Springer, 2007.

APPENDIX A
PROOF OF THEOREM 1: OPTIMALITY OF ADDITIVE CONTINUOUS CHANNELS

Let $F : \mathscr{R} \to [0,\infty]$ denote the objective function in (6), i.e.,

$$F(P_{Y|X}) := \sup_{|u-v|\le s} D(P_{Y|X=u}\|P_{Y|X=v}). \tag{22}$$

Thus,

$$\mathrm{KL}^\star = \inf_{P_{Y|X}\in\mathscr{P}} F(P_{Y|X}). \tag{23}$$

Fix a sequence of conditional distributions

$$\left\{P_{Y|X}^{(k)}\right\}_{k\in\mathbb{N}} \subset \mathscr{P} \tag{24}$$

satisfying

$$\mathrm{KL}^\star = \lim_{k\to\infty} F\left(P_{Y|X}^{(k)}\right). \tag{25}$$

Recall that by assumption, the version of each conditional distribution $P_{Y|X}^{(k)}$ we choose is regular, i.e., $x \mapsto P_{Y|X=x}^{(k)}(B)$ is a Borel function for each Borel set $B \subset \mathbb{R}$. Note that $\mathrm{KL}^\star < \infty$ since, e.g., the Gaussian mechanism is feasible. Throwing away the first few elements in the sequence, we assume that $F\left(P_{Y|X}^{(k)}\right) < \infty$ for each $k \in \mathbb{N}$.

We break the proof down into several steps:

1) Introduce Markov kernels $\overline{P}_{Y|X}^{(k)}$ as "continuous" convex combinations of the $P_{Y|X}^{(k)}$.
2) The $\overline{P}_{Y|X}^{(k)}$ also satisfy the cost constraint.
3) The $\overline{P}_{Y|X}^{(k)}$ asymptotically achieve $\mathrm{KL}^\star$.
4) The $\overline{P}_{Y|X=x}^{(k)}$ are asymptotically shifted versions $T_x P^\star$ of a fixed $P^\star \in \mathscr{B}$.
5) $P^\star$ achieves $\mathrm{KL}^\star$.

• *Step 1*: Averaging the $P_{Y|X}^{(k)}$.

For $k \in \mathbb{N}$, we will define the Markov kernel $\overline{P}_{Y|X}^{(k)} \in \mathscr{R}$ by

$$\overline{P}_{Y|X=x}^{(k)}(B) := \frac{1}{2k}\int_{-k}^{k} P_{Y|X=x+z}^{(k)}(B+z)\,dz. \tag{26}$$

Of course, we need to check that (26) indeed yields a Markov kernel $\overline{P}_{Y|X}^{(k)}$. In view of Fubini's theorem, it suffices to check that the map $(x,z) \mapsto P_{Y|X=x+z}^{(k)}(B+z)$ is jointly Borel (for every fixed Borel set $B \subset \mathbb{R}$). This joint measurability is not self-evident, so we check next that it indeed holds.

Let the transition probability kernel $L^{(k)} : \mathbb{R}^2 \times \mathcal{B}(\mathbb{R}) \to [0,1]$ be defined by

$$L^{(k)}((x,z),A) := P_{Y|X=x+z}^{(k)}(A). \tag{27}$$

Let $N^{(k)} : \mathbb{R}^2 \times \mathcal{B}(\mathbb{R}) \to [0,1]$ denote the map

$$N^{(k)}((x,z),B) := P_{Y|X=x+z}^{(k)}(B+z). \tag{28}$$

For each $(x,z) \in \mathbb{R}^2$ and Borel set $B \subset \mathbb{R}$, we may write $N^{(k)}((x,z),B)$ as the integral of a nonnegative Borel function against $L((x,z),dy)$, namely,

$$N^{(k)}((x,z),B) = \int_{\mathbb{R}} 1_B(y-z)\,L((x,z),dy). \tag{29}$$

Hence (see, e.g., [25, Chapter I, Proposition 6.9]) $(x,z) \mapsto N^{(k)}((x,z),B)$ is a Borel function. Hence, $\overline{P}_{Y|X}^{(k)}$ as given by (26) is indeed a well-defined Markov kernel on $\mathbb{R}$.

For the next steps, we will use the following notation

$$R_{Y|X=z}^{(k,x)}(B) := P_{Y|X=x+z}^{(k)}(B+z), \tag{30}$$

$$P^{(k,x)}(B) := \overline{P}_{Y|X=x}^{(k)}(B), \tag{31}$$

$$U^{(k)}(B) := \frac{1}{2k}\cdot\lambda(B\cap[-k,k]). \tag{32}$$

Note that $R_{Y|X}^{(k,x)} \in \mathscr{R}$ and $P^{(k,x)} \in \mathscr{B}$ for each fixed $(k,x) \in \mathbb{N}\times\mathbb{R}$, and (26) may be rewritten as

$$P^{(k,x)} = R_{Y|X}^{(k,x)} \circ U^{(k)}. \tag{33}$$

• *Step 2*: The $\overline{P}_{Y|X}^{(k)}$ satisfy the cost constraint.

Fix $k \in \mathbb{N}$, and we will show next that $\overline{P}_{Y|X}^{(k)} \in \mathscr{P}$, i.e., that $\overline{P}_{Y|X}^{(k)}$ satisfies the cost constraint. Recall that a Markov kernel $P_{Y|X} \in \mathscr{R}$ belongs to $\mathscr{P}$ if and only if it satisfies

$$\sup_{x\in\mathbb{R}} \mathbb{E}_{P_{Y|X=x}}[T_x c] \le C. \tag{34}$$

By the assumption that $P_{Y|X}^{(k)} \in \mathscr{P}$, we have that

$$\mathbb{E}_{P_{Y|X=x}^{(k)}}[T_x c] \le C \tag{35}$$

for every $x \in \mathbb{R}$. Shifting the variable of integration in (35) by a fixed constant $-z$, we obtain that

$$\mathbb{E}_{T_{-z}P_{Y|X=x}^{(k)}}[T_{x-z}c] \le C \tag{36}$$

for every $(x,z) \in \mathbb{R}^2$. Replacing x by $x+z$ in (36), we conclude that (see (30))

$$\mathbb{E}_{R_{Y|X=z}^{(k,x)}}[T_x c] \le C \tag{37}$$

for every $(x,z) \in \mathbb{R}^2$. We proceed via the following standard approximation by simple functions argument.

Fix $x \in \mathbb{R}$, and let $\sum_j a_j 1_{B_j}(y)$ be a nonnegative simple function upper bounded by $(T_x c)(y)$. Integrating against $R_{Y|X=z}^{(k,x)}(dy)$ we deduce from (37) that

$$\sum_j a_j R_{Y|X=z}^{(k,x)}(B_j) \le C \tag{38}$$

for every $z \in \mathbb{R}$. Integrating (38) against $U^{(k)}(dz)$, and noting that $P^{(k,x)} = R_{Y|X}^{(k,x)} \circ U^{(k)}$ (see (33)), we deduce that

$$\sum_j a_j P^{(k,x)}(B_j) \le C. \tag{39}$$

Now, as (39) holds for all nonnegative simple functions below $T_x c$, taking an increasing sequence of nonnegative simple function converging pointwise to $T_x c$ we conclude that

$$\mathbb{E}_{P^{(k,x)}}[T_x c] \le C. \tag{40}$$

In other words (see (31)),

$$\mathbb{E}_{\overline{P}_{Y|X=x}^{(k)}}[T_x c] \le C. \tag{41}$$

As (41) holds for all $x \in \mathbb{R}$, we have shown that $\overline{P}_{Y|X}^{(k)} \in \mathscr{P}$.

• *Step 3*: The $\overline{P}^{(k)}_{Y|X}$ are asymptotically optimal.

Next, we use monotonicity of the KL-divergence under conditioning (see Lemma 1) to show the limit

$$\mathrm{KL}^\star = \lim_{k\to\infty} F\left(\overline{P}^{(k)}_{Y|X}\right). \tag{42}$$

Shift-invariance of the KL-divergence implies that, for each $x, x', z \in \mathbb{R}$,

$$D\left(R^{(k,x)}_{Y|X=z} \| R^{(k,x')}_{Y|X=z}\right) = D\left(P^{(k)}_{Y|X=x+z} \| P^{(k)}_{Y|X=x'+z}\right). \tag{43}$$

Thus, as $(x+z)-(x'+z) = x - x'$, we conclude that

$$\sup_{\substack{|x-x'|\le s\\ z\in\mathbb{R}}} D\left(R^{(k,x)}_{Y|X=z} \| R^{(k,x')}_{Y|X=z}\right) = F\left(P^{(k)}_{Y|X}\right) \tag{44}$$

By assumption of optimality of the $P^{(k)}_{Y|X}$ (see (25)), there exists a k_0 such that for all $k \ge k_0$,

$$\sup_{\substack{|x-x'|\le s\\ z\in\mathbb{R}}} D\left(R^{(k,x)}_{Y|X=z} \| R^{(k,x')}_{Y|X=z}\right) \le \mathrm{KL}^\star + \delta. \tag{45}$$

By definition of KL-divergence, we infer $R^{(k,x)}_{Y|X=z} \ll R^{(k,x')}_{Y|X=z}$ for all $z \in \mathbb{R}$ and $|x - x'| \le s$. Also, (45) shows in particular that

$$\sup_{|x-x'|\le s} \mathbb{E}_{\xi\sim U^{(k)}}\left[D\left(R^{(k,x)}_{Y|X=\xi} \| R^{(k,x')}_{Y|X=\xi}\right)\right] \le \mathrm{KL}^\star + \delta. \tag{46}$$

Using (33), Lemma 1 yields that

$$\sup_{|x-x'|\le s} D\left(P^{(k,x)} \| P^{(k,x')}\right) \le \mathrm{KL}^\star + \delta. \tag{47}$$

Taking $\delta \to 0^+$, we see that (42) holds.

• *Step 4*: $P^{(k,x)}$ is asymptotically $T_x P^\star$ for a fixed $P^\star$.

Next, we show that there is a measure $P^\star \in \mathscr{B}$ such that, for every $x \in \mathbb{R}$, we have the weak convergence

$$P^{(k,x)} \to T_x P^\star \tag{48}$$

as $k \to \infty$.

First, for each fixed $x \in \mathbb{R}$, we establish the total-variation distance convergence

$$\lim_{k\to\infty} \left\| P^{(k,x)} - T_x P^{(k,0)} \right\|_{\mathrm{TV}} = 0. \tag{49}$$

We may write

$$\left(T_x P^{(k,0)}\right)(B) = \frac{1}{2k}\int_{-k-x}^{k-x} R^{(k,x)}_{Y|X=z}(B)\, dz. \tag{50}$$

Therefore, for any Borel set $B \subset \mathbb{R}$ we have that

$$\begin{aligned}&\left|P^{(k,x)}(B) - T_x P^{(k,0)}(B)\right| \\ &\quad\le \frac{1}{2k}\int_{[-k,k]\Delta[-k-x,k-x]} R^{(k,x)}_{Y|X=z}(B)\, dz \le \frac{|x|}{k},\end{aligned} \tag{51}$$

where Δ denotes the symmetric difference. As the bound (51) is uniform in B, we conclude that the total-variation limit in (49) holds.

The next ingredient we need is that the set $\{P^{(k,0)}\}_{k\in\mathbb{N}} \subset \mathscr{B}$ is tight, i.e., that for any $\varepsilon > 0$ there exists an $n > 0$ such that

$$\sup_{k\in\mathbb{N}} P^{(k,0)}(\mathbb{R} \setminus [-n, n]) \le \varepsilon. \tag{52}$$

Fix $\varepsilon > 0$. By the assumption that $c(x) \sim \beta|x|^\alpha$ where $\alpha, \beta > 0$, we have $\lim_{|x|\to\infty} c(x) = \infty$. Thus there exists an integer n such that $c(x) \ge C/\varepsilon$ whenever $|x| \ge n$. Then, for each $(z, k) \in \mathbb{R} \times \mathbb{N}$,

$$\begin{aligned}&R^{(k,0)}_{Y|X=z}(\mathbb{R} \setminus [-n, n]) \\ &\quad= P^{(k)}_{Y|X=z}(\mathbb{R} \setminus [-n+z, n+z]) &(53)\\ &\quad\le \int_{\mathbb{R}\setminus[-n+z,n+z]} \frac{c(y-z)}{c(n)}\, dP^{(k)}_{Y|X=z}(y) &(54)\\ &\quad\le c(n)^{-1}\mathbb{E}_{P^{(k)}_{Y|X=z}}[T_z c] &(55)\\ &\quad\le c(n)^{-1}\cdot C &(56)\\ &\quad\le \varepsilon, &(57)\end{aligned}$$

where (54) follows by monotonicity of c, (55) by nonnegativity of c, and (56) since $P^{(k)}_{Y|X} \in \mathscr{P}$. Hence,

$$\sup_{(z,k)\in\mathbb{R}\times\mathbb{N}} R^{(k,0)}_{Y|X=z}(\mathbb{R} \setminus [-n, n]) \le \varepsilon. \tag{58}$$

Averaging over z, we deduce that (52) holds, i.e., that $\{P^{(k,0)}\}_{k\in\mathbb{N}}$ is tight.

By tightness of $\{P^{(k,0)}\}_{k\in\mathbb{N}}$, we conclude via Prokhorov's theorem [25, Chapter 3, Theorem 5.13] after passing to a subsequence that there is a $P^\star \in \mathscr{B}$ such that $P^{(k,0)} \to P^\star$ weakly as $k \to \infty$, i.e., for every continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$ we have

$$\lim_{k\to\infty} \mathbb{E}_{P^{(k,0)}}[f] = \mathbb{E}_{P^\star}[f]. \tag{59}$$

This immediately implies that, for each $x \in \mathbb{R}$, we also have

$$T_x P^{(k,0)} \to T_x P^\star \tag{60}$$

weakly as $k \to \infty$. As convergence in total variation is stronger than weak convergence, we conclude from (49) and (60) that for every $x \in \mathbb{R}$

$$P^{(k,x)} \to T_x P^\star \tag{61}$$

weakly as $k \to \infty$.

• *Step 5*: The additive mechanism $P^\star$ is optimal.

The final step is showing that $P^\star$ attains $\mathrm{KL}^\star$ and satisfies the cost constraint. By joint lower-semicontinuity of the KL-divergence [28, Theorem 1], we deduce from (61) that for each $x \in \mathbb{R}$

$$D(P^\star \| T_x P^\star) \le \liminf_{k\to\infty} D\left(P^{(k,0)} \| P^{(k,x)}\right). \tag{62}$$

But we also have

$$\sup_{|x|\le s} D\left(P^{(k,0)} \| P^{(k,x)}\right) \le F\left(\overline{P}^{(k)}_{Y|X}\right). \tag{63}$$

Therefore, taking the supremum over $|x| \le s$ in (62), we infer from (42) that

$$\sup_{|x|\le s} D(P^\star \| T_x P^\star) \le \mathrm{KL}^\star. \tag{64}$$

Hence, it only remains to check that $P^\star \in \mathscr{P}_{\text{add}}$ for us to conclude that equality holds in (64).

For every $A > 0$ and $x \in \mathbb{R}$, the function $1_{[-A,A]} \cdot T_x c$ is continuous and bounded. Hence, the weak convergence $P^{(k,x)} \to T_x P^\star$ yields

$$\mathbb{E}_{T_x P^\star}\left[1_{[-A,A]} \cdot T_x c\right] = \lim_{k\to\infty} \mathbb{E}_{P^{(k,x)}}\left[1_{[-A,A]} \cdot T_x c\right]. \quad (65)$$

As $\overline{P}^{(k)}_{Y|X} \in \mathscr{P}$, nonnegativity of c implies in view of (65) that

$$\mathbb{E}_{T_x P^\star}\left[1_{[-A,A]} \cdot T_x c\right] \le C. \quad (66)$$

By the monotone convergence theorem, taking $A \to \infty$ yields

$$\mathbb{E}_{T_x P^\star}\left[T_x c\right] \le C, \quad (67)$$

In other words, $P^\star \in \mathscr{P}_{\text{add}}$. Therefore, we must have

$$\text{KL}^\star \le \text{KL}^\star_{\text{add}} \le \sup_{|x|\le s} D(P^\star \| T_x P^\star). \quad (68)$$

Combining this inequality with (64), we conclude that

$$\text{KL}^\star = \text{KL}^\star_{\text{add}} = \sup_{|x|\le s} D(P^\star \| T_x P^\star). \quad (69)$$

This completes the proof of the first statement of the theorem.

For the last statement of the theorem, we show that the relation $\mu \ll T_x\mu$ for every $|x| \le s$ (which holds for $P^\star$ by (69) and $\text{KL}^\star < \infty$) is enough to conclude that $\mu \ll \lambda$. Fix a Borel set $B \subset \mathbb{R}$ such that $\lambda(B) = 0$, and we will show that $\mu(B) = 0$. Note that the function $x \mapsto (T_x\mu)(B)$ is Borel as it is given by the convolution $1_B * \eta$ where $\eta(A) := \mu(-A)$. Then, by Tonelli's theorem and translation-invariance of the Lebesgue measure,

$$\begin{aligned}
\int_{\mathbb{R}} (T_x\mu)(B)\, d\lambda(x) &= \int_{\mathbb{R}^2} 1_{B-x}(b)\, d\mu(b)\, d\lambda(x) && (70)\\
&= \int_{\mathbb{R}^2} 1_{B-x}(b)\, d\lambda(x)\, d\mu(b) && (71)\\
&= \int_{\mathbb{R}^2} 1_{B-b}(x)\, d\lambda(x)\, d\mu(b) && (72)\\
&= \int_{\mathbb{R}} (T_b\lambda)(B)\, d\mu(b) && (73)\\
&= \int_{\mathbb{R}} \lambda(B)\, d\mu(b) = 0. && (74)
\end{aligned}$$

Thus, $(T_x\mu)(B) = 0$ for λ-almost every x. In particular, $(T_x\mu)(B) = 0$ for at least one $x \in [-s, s]$. Thus, $\mu \ll T_x\mu$ implies $\mu(B) = 0$, and the proof is complete.

Lemma 1 (Conditioning increases divergence)**.** *Let $P_{Y|X}, P'_{Y|X}$ be Markov kernels on $\mathbb{R}$ such that $P_{Y|X=x} \ll P'_{Y|X=x}$ for every $x \in \mathbb{R}$. Then, denoting the marginalizations of $P_{X,Y} := P_{Y|X} \otimes P_X, P'_{X,Y} := P'_{Y|X} \otimes P_X$ in the second coordinate by P_Y, P'_Y, we have that*

$$D\left(P_Y \| P'_Y\right) \le \mathbb{E}_{\xi\sim P_X}\left[D\left(P_{Y|X=\xi} \| P'_{Y|X=\xi}\right)\right]. \quad (75)$$

Proof: Since by assumption $P_{Y|X=x} \ll P'_{Y|X=x}$ for every $x \in \mathbb{R}$, a generalization of the Radon-Nikodym theorem by Doob (see [25, Chapter 5, Theorem 4.44]) yields the existence of a version of the Radon-Nikodym derivatives $dP_{Y|X=x}/dP'_{Y|X=x}$ such that the function

$$(x, y) \mapsto \frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y) \quad (76)$$

is jointly measurable. We show that this function is a version of $dP_{X,Y}/dP'_{X,Y}$. First, note that $P_{X,Y} \ll P'_{X,Y}$ are equivalent. Indeed, for any Borel set $E \subset \mathbb{R}$, denoting the sections by $E_x := \{y \in \mathbb{R}\ ;\ (x, y) \in E\}$, we have that $P_{X,Y}(E) = 0$ if and only if $P_{Y|X=x}(E_x) = 0$ for P_X-a.e. x, and a similar statement holds for $P'_{X,Y}$. By assumption, $P_{Y|X=x} \ll P'_{Y|X=x}$ for each x, so we obtain $P_{X,Y} \ll P'_{X,Y}$. By joint measurability and nonnegativity, using the disintegration theorem (see, e.g., [25, Chapter 1, Theorem 6.11]) we obtain that for any Borel $E \subset \mathbb{R}^2$

$$\begin{aligned}
&\int_E \frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y)\, dP'_{X,Y}(x, y)\\
&\quad= \int_{\mathbb{R}} \int_{E_x} \frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y)\, dP'_{Y|X=x}(y)\, dP_X(x) && (77)\\
&\quad= \int_{\mathbb{R}} \int_{E_x} dP_{Y|X=x}(y)\, dP_X(x) && (78)\\
&\quad= P_{X,Y}(E). && (79)
\end{aligned}$$

Thus, we have the equality

$$\frac{dP_{X,Y}}{dP'_{X,Y}}(x, y) = \frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y) \quad (80)$$

for $P'_{X,Y}$-a.e. (x, y).

Define $f : [0, \infty) \to [-1/e, \infty)$ by $f(0) = 0$ and $f(t) = t \log t$ for $t > 0$. By the disintegration theorem and (80), we have the equality

$$\begin{aligned}
&D\left(P_{X,Y} \| P'_{X,Y}\right)\\
&\quad= \int_{\mathbb{R}^2} f\left(\frac{dP_{X,Y}}{dP'_{X,Y}}\right) dP'_{X,Y} && (81)\\
&\quad= \int_{\mathbb{R}} \int_{\mathbb{R}} f\left(\frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y)\right) dP'_{Y|X=x}\, dP_X(x) && (82)\\
&\quad= \mathbb{E}_{\xi\sim P_X}\left[D\left(P_{Y|X=\xi} \| P'_{Y|X=\xi}\right)\right]. && (83)
\end{aligned}$$

On the other hand, disintegration with respect to Y yields the following bound. Denote by $P_{X|Y}, P'_{X|Y}$ the disintegrations of $P_{X,Y}, P'_{X,Y}$ with respect to P_Y, P'_Y. In particular, $P_{X|Y}$ and $P'_{X|Y}$ are Markov kernels on $\mathbb{R}$. By the disintegration theorem and Jensen's inequality,

$$\begin{aligned}
&D\left(P_{X,Y} \| P'_{X,Y}\right)\\
&\quad= \int_{\mathbb{R}^2} f\left(\frac{dP_{X,Y}}{dP'_{X,Y}}\right) dP'_{X,Y} && (84)\\
&\quad= \int_{\mathbb{R}} \int_{\mathbb{R}} f\left(\frac{dP_{X,Y}}{dP'_{X,Y}}(x, y)\right) dP'_{X|Y=x}(x)\, dP'_Y(y) && (85)\\
&\quad\ge \int_{\mathbb{R}} f\left(g(y)\right)\, dP'_Y(y) && (86)
\end{aligned}$$

where

$$g(y) := \int_{\mathbb{R}} \frac{dP_{X,Y}}{dP'_{X,Y}}(x,y)\, dP'_{X|Y=x}(x). \tag{87}$$

For this application of Jensen's inequality, we use the fact, shown next, that g is finite P'_Y-a.e. In fact, we show that g is a version of dP_Y/dP'_Y. Note that $P_{X,Y} \ll P'_{X,Y}$ implies that $P_Y \ll P'_Y$. Now, for any Borel $B \subset \mathbb{R}$, the disintegration theorem yields that

$$\int_B g\, dP'_Y = \int_B \int_{\mathbb{R}} \frac{dP_{X,Y}}{dP'_{X,Y}}(x,y)\, dP'_{X|Y=x}(x)\, dP'_Y(y) \tag{88}$$

$$= \int_{\mathbb{R}\times B} \frac{dP_{X,Y}}{dP'_{X,Y}}\, dP'_{X,Y} \tag{89}$$

$$= P_{X,Y}(\mathbb{R}\times B) = P_Y(B). \tag{90}$$

Thus, we have that

$$g(y) = \frac{dP_Y}{dP'_Y}(y). \tag{91}$$

for P'_Y-a.e. y. Hence, we obtain from inequality (86) that

$$D\left(P_{X,Y}\|P'_{X,Y}\right) \geq D\left(P_Y\|P'_Y\right). \tag{92}$$

Combining inequality (92) and equation (83) we obtain the desired inequality (75). ∎

Appendix B
Proof of Theorem 2: Finite-Dimensionality

Note that the vector $\boldsymbol{p}$ only includes p_i for $0 \leq i \leq N$. We will simplify our analysis by defining p_i for all integers i. Specifically, for $i \in \mathbb{Z} \setminus \{0, \cdots, N\}$, we denote

$$p_i := \begin{cases} p_{|i|}, & \text{if } -N \leq i \leq -1, \\ p_N r^{|i|-N}, & \text{if } |i| > N. \end{cases} \tag{93}$$

Thus we may rewrite the formula for $f_{n,r,\boldsymbol{p}}$ in (14) as

$$f_{n,r,\boldsymbol{p}}(x) = np_i \quad \text{if } x \in J_{n,i}. \tag{94}$$

We show first that

$$\sup_{a\in\mathbb{R}:|a|\leq 1} D(P_{n,r,\boldsymbol{p}}\|T_a P_{n,r,\boldsymbol{p}}) = \max_{k\in\mathbb{Z}:|k|\leq n} \sum_{i\in\mathbb{Z}} p_i \log\frac{p_i}{p_{i+k}}, \tag{95}$$

then we show that this formula is equal to the objective function in (20). For convenience, we drop the subscripts on $f_{n,r,\boldsymbol{p}}$ and $P_{n,r,\boldsymbol{p}}$ throughout this proof. We may assume $\boldsymbol{p} > \mathbf{0}$, since any vector $\boldsymbol{p}$ with some zero coordinate will be infeasible in both optimization problems (18) and (20).

Fix $a \in [-1,1]$. For each $i \in \mathbb{Z}$, let $J^\circ_{n,i} = \left(\frac{i-1/2}{n}, \frac{i+1/2}{n}\right)$ denote the interior of $J_{n,i}$. We start by showing that the function

$$F_a := f \log \frac{f}{T_{-a} f} \tag{96}$$

is integrable, which would allow us to use countable additivity of the Lebesgue integral to split $D(P\|T_{-a}P)$ into a sum of integrals over the $J^\circ_{n,i}$. Let $k \in \mathbb{Z}$ be the unique integer such that $a + \frac{1}{2n} \in J_{n,k}$, and denote $\Delta := k - an$. From

$$\frac{k-1/2}{n} \leq a + \frac{1}{2n} \leq \frac{k+1/2}{n}, \tag{97}$$

we conclude that $0 \leq \Delta \leq 1$. Consider an integer i and a real $x \in J^\circ_{n,i}$. If $x < (i - 1/2 + \Delta)/n$, then

$$x + a = x + \frac{k-\Delta}{n} < \frac{i+k-1/2}{n} = \frac{(i+k-1)+1/2}{n} \tag{98}$$

and, since $\Delta \leq 1$,

$$x + a = x + \frac{k-\Delta}{n} > \frac{i-1/2}{n} + \frac{k-1}{n} = \frac{(i+k-1)-1/2}{n}. \tag{99}$$

Inequalities (98) and (99) together imply that $x + a \in J^\circ_{n,i+k-1}$. Similarly, if $x > (i - 1/2 + \Delta)/n$ then $x + a \in J^\circ_{n,i+k}$. We may ignore the countably many cases $x = (i - 1/2 + \Delta)/n$ (as i varies over $\mathbb{Z}$) for the sake of integrating F_a. We conclude that for every $x \in \mathbb{R}$ such that $nx - \Delta + \frac{1}{2}$ is not an integer,

$$F_a(x) = \begin{cases} np_i \log\frac{p_i}{p_{i+k-1}}, & \text{if } x \in J_{n,i}, \quad x < \frac{i-1/2+\Delta}{n}, \\ np_i \log\frac{p_i}{p_{i+k}}, & \text{if } x \in J_{n,i}, \quad x > \frac{i-1/2+\Delta}{n}. \end{cases} \tag{100}$$

Since $\int_{\mathbb{R}} |F_a| = \sum_{i\in\mathbb{Z}} \int_{J^\circ_{n,i}} |F_a|$, we obtain

$$\int_{\mathbb{R}} |F_a| = \sum_{i\in\mathbb{Z}} p_i \left(\Delta \left|\log\frac{p_i}{p_{i+k-1}}\right| + (1-\Delta)\left|\log\frac{p_i}{p_{i+k}}\right|\right). \tag{101}$$

Now, we may conclude that $F_a \in L^1(\mathbb{R})$ by comparison with a geometric series. Indeed, we show the convergence of the series

$$S_\ell := \sum_{i\in\mathbb{Z}} p_i \left|\log\frac{p_i}{p_{i+\ell}}\right| \tag{102}$$

for each fixed $\ell \in \mathbb{Z}$. Consider the set of indices

$$I = \mathbb{Z} \setminus \{-N - |\ell|, \cdots, N + |\ell|\}, \tag{103}$$

and note that for each $i \in I$ we have $p_{i+j} = p_N r^{|i+j|-N}$ for both values $j \in \{0, \ell\}$. In particular, for $i \in I$ we have that

$$\left|\log\frac{p_i}{p_{i+\ell}}\right| = ||i| - |i+\ell|| \cdot \log\frac{1}{r} \leq |\ell| \cdot \log\frac{1}{r}. \tag{104}$$

Therefore, we obtain the bound

$$S_\ell \leq \frac{|\ell| p_N \log\frac{1}{r}}{r^N} \cdot \frac{1+r}{1-r} + \sum_{|i|\leq N+|\ell|} p_i \left|\log\frac{p_i}{p_{i+\ell}}\right| < \infty. \tag{105}$$

As S_k and S_{k-1} are both finite, we conclude from (101) that $F_a \in L^1(\mathbb{R})$. Therefore, by countable additivity,

$$D(P\|T_{-a}P) = \sum_{i\in\mathbb{Z}} \int_{J^\circ_{n,i}} F_a, \tag{106}$$

i.e.,

$$D(P\|T_{-a}P) = \sum_{i\in\mathbb{Z}} p_i \left(\Delta \log\frac{p_i}{p_{i+k-1}} + (1-\Delta)\log\frac{p_i}{p_{i+k}}\right). \tag{107}$$

Let B_ℓ denote the same sum as S_ℓ but without the absolute value sign,

$$B_\ell := \sum_{i\in\mathbb{Z}} p_i \log\frac{p_i}{p_{i+\ell}}. \tag{108}$$

Finiteness of the S_ℓ yields from (107) that

$$D(P\|T_{-a}P) = \Delta B_{k-1} + (1-\Delta) B_k. \tag{109}$$

Also, the relation we are aiming to prove (95) can be restated as

$$\sup_{|d|\le 1} D(P\|T_dP) = \max_{|\ell|\le n} B_\ell. \tag{110}$$

We deduce from $k = an + \Delta$, $|a| \le 1$, and $0 \le \Delta \le 1$ that we must have $-n \le k \le n+1$. If it holds that $-n+1 \le k \le n$, then what we have shown in (109) implies, in view of $0 \le \Delta \le 1$, that

$$D(P\|T_{-a}P) \le \max_{|\ell|\le n} B_\ell. \tag{111}$$

We treat the remaining two extreme cases $k \in \{-n, n+1\}$ separately. First, if $k = -n$ then $\Delta = 0$, in which case

$$D(P\|T_{-a}P) = B_{-n} \le \max_{|\ell|\le n} B_\ell. \tag{112}$$

Second, if $k = n+1$ then $\Delta = 1$, in which case

$$D(P\|T_{-a}P) = B_n \le \max_{|\ell|\le n} B_\ell. \tag{113}$$

Combining all cases, we conclude that

$$\sup_{|d|\le 1} D(P\|T_dP) \le \max_{|\ell|\le n} B_\ell. \tag{114}$$

We establish now that the reverse inequality in (114) also holds. Let $\ell \in \{0, \cdots, n\}$. The shift $a_\ell := \ell/n$ satisfies $|a_\ell| \le 1$ and $a_\ell + \frac{1}{2n} \in J_{n,\ell}$. Also, $\Delta_\ell := \ell - a_\ell n = 0$. Therefore, we conclude from (109) that

$$D(P\|T_{-a_\ell}P) = B_\ell. \tag{115}$$

This shows that

$$\sup_{|d|\le 1} D(P\|T_dP) \ge \max_{0\le \ell\le n} B_\ell. \tag{116}$$

In addition, consider $\ell \in \{-n, \cdots, -1\}$ and the shift $a'_\ell := \ell/n$. Then, in this case $a'_\ell + \frac{1}{2n} \in J_{n,\ell+1}$. Also, $\Delta'_\ell := (\ell+1) - a'_\ell n = 1$. Thus, by (109), we have that

$$D(P\|T_{-a'_\ell}P) = B_{(\ell+1)-1} = B_\ell. \tag{117}$$

Therefore,

$$\sup_{|d|\le 1} D(P\|T_{-d}P) \ge \max_{-n\le \ell\le -1} B_\ell. \tag{118}$$

Combining (116) and (118), we conclude that

$$\sup_{|d|\le 1} D(P\|T_{-d}P) \ge \max_{|\ell|\le n} B_\ell. \tag{119}$$

Inequality (119) together with the reverse inequality (114) yield that the desired equation (95) holds, i.e.,

$$\sup_{|a|\le 1} D(P\|T_aP) = \max_{|k|\le n} \sum_{i\in\mathbb{Z}} p_i \log \frac{p_i}{p_{i+k}}. \tag{120}$$

Next, we show that the expression

$$\max_{|k|\le n} \sum_{i\in\mathbb{Z}} p_i \log \frac{p_i}{p_{i+k}} \tag{121}$$

reduces to the form given in the statement of the theorem. By construction, $p_i = p_{-i}$ for each $i \in \mathbb{Z}$. Thus, we have for each $k \in \mathbb{Z}$

$$B_k = \sum_{i\in\mathbb{Z}} p_i \log \frac{p_i}{p_{i+k}} = \sum_{j\in\mathbb{Z}} p_{-j} \log \frac{p_{-j}}{p_{-j+k}} \tag{122}$$

$$= \sum_{j\in\mathbb{Z}} p_j \log \frac{p_j}{p_{j-k}} = B_{-k}. \tag{123}$$

Therefore, $B_k = (B_k + B_{-k})/2$ for every $k \in \mathbb{Z}$. Note that this is a symmetric expression in k. As $B_0 = 0$, the KL-divergence is nonnegative, and $B_k \ge 0$ for every $|k| \le n$ (see (115) and (117)), we conclude that

$$\sup_{|a|\le 1} D(P\|T_aP) = \max_{1\le k\le n} \frac{1}{2}(B_k + B_{-k}). \tag{124}$$

We now rewrite (124) in terms of p_i for only $0 \le i \le N$, by taking advantage of (93). Fix $k \in \{1, \cdots, n\}$. We may write

$$B_{-k} = \sum_{j\in\mathbb{Z}} p_j \log \frac{p_j}{p_{j-k}} = \sum_{i\in\mathbb{Z}} p_{i+k} \log \frac{p_{i+k}}{p_i}, \tag{125}$$

so

$$B_k + B_{-k} = \sum_{i\in\mathbb{Z}} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}}. \tag{126}$$

We split this sum at the points $-N, N-k$, and N. For any $k \in \{1, \ldots, n\}$, using the assumption that $n < N$, we may write

$$\sum_{i\in\mathbb{Z}} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}} = \sum_{i=-N+1}^{N-k-1} (p_{|i|} - p_{|i+k|}) \log \frac{p_{|i|}}{p_{|i+k|}}$$
$$+ \sum_{i=N-k}^{\infty} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}} + \sum_{i=-\infty}^{-N} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}}. \tag{127}$$

In fact, the third term in (127) is identical to the second. This is proved by

$$\sum_{i=-\infty}^{-N} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}} = \sum_{i=N}^{\infty} (p_{-i} - p_{-i+k}) \log \frac{p_{-i}}{p_{-i+k}} \tag{128}$$

$$= \sum_{i=N}^{\infty} (p_i - p_{i-k}) \log \frac{p_i}{p_{i-k}} \tag{129}$$

$$= \sum_{i=N-k}^{\infty} (p_{i+k} - p_i) \log \frac{p_{i+k}}{p_i} \tag{130}$$

$$= \sum_{i=N-k}^{\infty} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}}. \tag{131}$$

Moreover, we may rewrite this expression as

$$\sum_{i=N-k}^{\infty} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}} \tag{132}$$
$$= \sum_{i=N-k}^{N-1} (p_i - p_N r^{i+k-N}) \log \frac{p_i}{p_N r^{i+k-N}}$$

$$+\sum_{i=N}^{\infty}(p_N r^{i-N}-p_N r^{i+k-N})\log\frac{p_N r^{i-N}}{p_N r^{i+k-N}} \quad (133)$$

$$=\sum_{i=N-k}^{N-1}(p_i-p_N r^{i+k-N})\log\frac{p_i}{p_N r^{i+k-N}}$$

$$+p_N\sum_{i=N}^{\infty}r^{i-N}(1-r^k)\log r^{-k} \quad (134)$$

$$=\sum_{i=N-k}^{N-1}(p_i-p_N r^{i+k-N})\log\frac{p_i}{p_N r^{i+k-N}} \quad (135)$$

$$+p_N\frac{1-r^k}{1-r}k\log r^{-1}. \quad (136)$$

Putting all of the above together shows that (124) is exactly equal to the objective function in (20).

Finally, we show that the cost constraint

$$\mathbb{E}_P[c]\le C \quad (137)$$

is equivalent to the one given in (20). By nonnegativity of c, we have that

$$\mathbb{E}_P[c]=\int_{\mathbb{R}}fc=\sum_{i\in\mathbb{Z}}\int_{J_{n,i}}np_i c=\sum_{i\in\mathbb{Z}}p_i c_{n,i} \quad (138)$$

$$=p_0c_{n,0}+2\sum_{i=1}^{N-1}p_ic_{n,i}+2p_N\sum_{i=N}^{\infty}c_{n,i}r^{i-N}, \quad (139)$$

and the proof is complete.

Appendix C
Proof of Theorem 3: Optimality of Cactus

We will use the integration shorthand

$$\int_A f:=\int_A f(x)\,dx. \quad (140)$$

Define

$$\gamma:=\begin{cases}1/2 & \text{if } \alpha>1,\\ \alpha/2 & \text{otherwise.}\end{cases} \quad (141)$$

Note that $\gamma\in(0,1/2]$ and $\gamma<\alpha$. Define the PDF

$$\psi(x):=\exp\left(-|x|^\gamma\right)\cdot\chi^{-1}, \quad (142)$$

where

$$\chi:=\int_{\mathbb{R}}\exp\left(-|x|^\gamma\right)\,dx \quad (143)$$

is the normalization constant. As $\gamma\in(0,1]$, the function $z\mapsto|z|^\gamma$ is subadditive. Hence, for any $x,y\in\mathbb{R}$ we have the inequality

$$\frac{\psi(x+y)}{\psi(x)}\le\exp\left(|y|^\gamma\right). \quad (144)$$

For each $\sigma>0$, denote the dilated PDF

$$\psi^\sigma(x):=\frac{1}{\sigma}\psi\left(\frac{x}{\sigma}\right). \quad (145)$$

We denote the result of convolving a PDF q with ψ^σ by q_σ,

$$q_\sigma:=q*\psi^\sigma. \quad (146)$$

For any $a\in\mathbb{R}$, it is easy to see that

$$T_a(q_\sigma)=(T_aq)_\sigma, \quad (147)$$

so we denote this common quantity by T_aq_σ.

Due to the length of the proof, we break down some of the initial steps into the following five auxiliary lemmas. The proof resumes in the subsequent subsection.

A. Auxiliary Lemmas

The first lemma helps reduce the problem to considering only continuous PDFs. Specifically, it shows that a convolution q_σ can perform arbitrarily close to how the original PDF q does.

Lemma 2. *For any PDF q and constant $\eta>0$, there is a constant $\sigma_0\in(0,1)$ such that $\sigma\in(0,\sigma_0]$ implies the inequalities*

$$D(q_\sigma\|T_aq_\sigma)\le D(q\|T_aq),\quad \textit{for all } a\in\mathbb{R}, \quad (148)$$

$$\mathbb{E}_{q_\sigma}[c]\le\mathbb{E}_q[c]+\eta. \quad (149)$$

Proof: First, by the data-processing inequality, for any $a\in\mathbb{R}$ and $\sigma>0$,

$$D(q_\sigma\|T_aq_\sigma)\le D(q\|T_aq). \quad (150)$$

Thus, (148) always holds. We may assume that $\mathbb{E}_q[c]<\infty$, for otherwise (149) trivially holds. Now, we will establish (149) for all small σ by proving the limit

$$\lim_{\sigma\to0^+}\mathbb{E}_{q_\sigma}[c]=\mathbb{E}_q[c]. \quad (151)$$

Let $(\Omega,\mathcal{F},P)$ be a probability space and $Z,V:\Omega\to\mathbb{R}$ be independent random variables with PDFs q and ψ, respectively, with respect to λ, i.e., with $P_Z(B):=P(Z^{-1}(B))$ and $P_V(B):=P(V^{-1}(B))$ we have

$$\frac{dP_Z}{d\lambda}=q,\qquad\frac{dP_V}{d\lambda}=\psi. \quad (152)$$

Then, for any $\sigma>0$, the random variable $Z_\sigma:=Z+\sigma V$ has PDF q_σ (see equations (141)–(146)). Denote integration against P by $\mathbb{E}$; in particular,

$$\mathbb{E}[f(Z,V)]:=\int_\Omega f(Z(\omega),V(\omega))\,dP(\omega) \quad (153)$$

for any Borel function $f:\mathbb{R}^2\to\mathbb{R}$.

By Slutsky's theorem, we have that $Z_\sigma\to Z$ in distribution. By the continuous mapping theorem, we also have that $c(Z_\sigma)\to c(Z)$ in distribution. Thus, by the Lebesgue-Vitali theorem [29, Theorem 4.5.4], to conclude that (151) holds, it suffices to show uniform integrability of $\{c(Z_\sigma)\}_{0<\sigma\le1}$, i.e., it suffices to show that

$$\lim_{K\to\infty}\sup_{0<\sigma\le1}\mathbb{E}\left[c(Z_\sigma)\cdot1_{(K,\infty)}(c(Z_\sigma))\right]=0. \quad (154)$$

To establish (154), it suffices to uniformly upper bound the $c(Z_\sigma)$ (for $\sigma\in(0,1]$) by an integrable random variable. To see this, note that if

$$\sup_{0<\sigma\le1}c(Z_\sigma)\le U \quad (155)$$

for some random variable $U:\Omega\to\mathbb{R}$ with $\mathbb{E}[U]<\infty$, then

we have the inequality

$$\sup_{0<\sigma\le 1} \mathbb{E}\left[c(Z_\sigma)\cdot 1_{(K,\infty)}(c(Z_\sigma))\right] \le \mathbb{E}\left[U\cdot 1_{(K,\infty)}(U)\right], \tag{156}$$

and the limit

$$\lim_{K\to\infty} \mathbb{E}\left[U\cdot 1_{(K,\infty)}(U)\right] = 0 \tag{157}$$

follows by absolute continuity of the Lebesgue integral in view of $\mathbb{E}[U]<\infty$.

Now, we show that a uniform bound as in (155) holds. Recall that for any $(u,v)\in\mathbb{R}^2$ and $0<s<t$, denoting $\|(u,v)\|_s := (|u|^s+|v|^s)^{1/s}$, one has from Hölder's inequality that

$$\|(u,v)\|_t \le \|(u,v)\|_s \le 2^{\frac{1}{s}-\frac{1}{t}}\|(u,v)\|_t. \tag{158}$$

In particular, for any $r>0$, denoting $\ell_r := \max(1,2^{r-1})$, one has that

$$(|u|+|v|)^r \le \ell_r(|u|^r+|v|^r). \tag{159}$$

In addition, by the tail-regularity assumption on c, there is a constant $\beta_1>0$ such that

$$c(x) \le \beta_1\,(1+|x|^\alpha) \tag{160}$$

for every $x\in\mathbb{R}$. Then, for any $u,v\in\mathbb{R}$, we have that

$$c(u+v) \le \beta_1\,(1+\ell_\alpha\,(|u|^\alpha+|v|^\alpha))\,. \tag{161}$$

In particular, for every $\sigma\in(0,1]$,

$$c(Z_\sigma) \le \beta_1\,(1+\ell_\alpha\,(|Z|^\alpha+|V|^\alpha)) =: U. \tag{162}$$

Now, we have that $\mathbb{E}[|V|^\alpha]<\infty$ by definition of ψ. Further, by assumption on c, there are $A,\beta_2>0$ such that $|x|>A$ implies

$$\beta_2|x|^\alpha \le c(x). \tag{163}$$

Then, as

$$|Z|^\alpha \le A^\alpha + |Z|^\alpha\cdot 1_{\mathbb{R}\setminus[-A,A]}(Z) \le A^\alpha + c(Z)/\beta_2 \tag{164}$$

and $\mathbb{E}[c(Z)] = \mathbb{E}_q[c]<\infty$ by assumption, we also have that $\mathbb{E}[|Z|^\alpha]<\infty$. Thus, $\mathbb{E}[U]<\infty$. Hence, by absolute continuity of the Lebesgue integral, the uniform bound in (162) implies the uniform integrability of the set $\{c(Z_\sigma)\}_{0<\sigma\le 1}$, so (151) follows by the Lebesgue-Vitali theorem, and the proof is complete. ■

The following lemma shows that the integrands when computing $D(q_\sigma\|T_aq_\sigma)$ have equi-small tails as a varies over $[-1,1]$. This will allow us to focus on approximating q_σ by a cactus distribution only in a bounded interval.

Lemma 3. *If the PDF q satisfies*

$$\sup_{|a|\le 1} D(q\|T_aq) < \infty \tag{165}$$

then for any $\sigma>0$

$$\lim_{z\to\infty}\sup_{|a|\le 1}\int_{\mathbb{R}\setminus[-z,z]} q_\sigma\left|\log\frac{q_\sigma}{T_aq_\sigma}\right| = 0. \tag{166}$$

Proof: Assume that q satisfies (165). By the data processing inequality, we also have

$$\sup_{|a|\le 1} D(q_\sigma\|T_aq_\sigma) < \infty. \tag{167}$$

Suppose, for the sake of contradiction, that (166) does not hold. That is, suppose there exists $\varepsilon>0$ where

$$\limsup_{z\to\infty}\sup_{|a|\le 1}\int_{\mathbb{R}\setminus[-z,z]} q_\sigma\left|\log\frac{q_\sigma}{T_aq_\sigma}\right| = \varepsilon. \tag{168}$$

This implies that there exists a sequence $\{(z_n,a_n)\}_{n\in\mathbb{N}}$, where $z_n\nearrow\infty$ and $\sup_{n\in\mathbb{N}}|a_n|\le 1$, such that for all n

$$\int_{\mathbb{R}\setminus[-z_n,z_n]} q_\sigma\left|\log\frac{q_\sigma}{T_{a_n}q_\sigma}\right| \ge \varepsilon/2. \tag{169}$$

Since $[-1,1]$ is a compact set, there exists a convergent subsequence $\{a_{n_k}\}_{k\in\mathbb{N}}$, say $a_{n_k}\to a$ where $a\in[-1,1]$. Moreover, for any $z>0$, for sufficiently large k we have $z_{n_k}\ge z$, which implies

$$\limsup_{k\to\infty}\int_{\mathbb{R}\setminus[-z,z]} q_\sigma\left|\log\frac{q_\sigma}{T_{a_{n_k}}q_\sigma}\right| \ge \varepsilon/2. \tag{170}$$

Recall that ψ is as defined in (142) and that, as shown in (144), it satisfies the inequality

$$\frac{\psi(x+y)}{\psi(x)} \le \exp\left(|y|^\gamma\right) \tag{171}$$

for every $x,y\in\mathbb{R}$. Thus, for any $a,b,z\in\mathbb{R}$,

$$\begin{aligned}
(T_aq_\sigma)(z) &= q_\sigma(z-a) && (172)\\
&= \int_{\mathbb{R}} q(x)\frac{1}{\sigma}\psi\left(\frac{z-a-x}{\sigma}\right)dx && (173)\\
&\le e^{|a-b|^\gamma/\sigma^\gamma}\int_{\mathbb{R}} q(x)\frac{1}{\sigma}\psi\left(\frac{z-b-x}{\sigma}\right)dx && (174)\\
&= e^{|a-b|^\gamma/\sigma^\gamma}(T_bq_\sigma)(z). && (175)
\end{aligned}$$

Thus, for any $a,b\in\mathbb{R}$, we have the uniform bound

$$\left\|\log\frac{T_aq_\sigma}{T_bq_\sigma}\right\|_{L^\infty(\mathbb{R})} \le \left(\frac{|a-b|}{\sigma}\right)^\gamma. \tag{176}$$

Applying this bound to the integral in (170) gives

$$\begin{aligned}
&\int_{\mathbb{R}\setminus[-z,z]} q_\sigma\cdot\left|\log\frac{q_\sigma}{T_{a_{n_k}}q_\sigma}\right| && (177)\\
&= \int_{\mathbb{R}\setminus[-z,z]} q_\sigma\cdot\left|\log\frac{q_\sigma}{T_aq_\sigma}+\log\frac{T_aq_\sigma}{T_{a_{n_k}}q_\sigma}\right| && (178)\\
&\le \int_{\mathbb{R}\setminus[-z,z]} q_\sigma\cdot\left(\left|\log\frac{q_\sigma}{T_aq_\sigma}\right|+\left(\frac{|a_{n_k}-a|}{\sigma}\right)^\gamma\right) && (179)\\
&\le \left(\frac{|a_{n_k}-a|}{\sigma}\right)^\gamma + \int_{\mathbb{R}\setminus[-z,z]} q_\sigma\cdot\left|\log\frac{q_\sigma}{T_aq_\sigma}\right|. && (180)
\end{aligned}$$

Recalling inequality (170) and that $a_{n_k}\to a$ as $k\to\infty$, we have, for any $z>0$,

$$\int_{\mathbb{R}\setminus[-z,z]} q_\sigma\left|\log\frac{q_\sigma}{T_aq_\sigma}\right| \ge \varepsilon/2. \tag{181}$$

Finally, note that by finiteness of the KL-divergence

$D(q_\sigma \| T_a q_\sigma)$ (see (167)), we also have that

$$\int_{\mathbb{R}} q_\sigma \left| \log \frac{q_\sigma}{T_a q_\sigma} \right| < \infty. \tag{182}$$

Indeed, the function $f(t) := t \log t$ over $(0, \infty)$ is lower bounded by $-1/e$, so dividing the integration region over the two regions where f is positive or negative we obtain

$$\int_{\mathbb{R}} q_\sigma \left| \log \frac{q_\sigma}{T_a q_\sigma} \right| = \mathbb{E}_{T_a q_\sigma} \left[\left| f \circ \frac{q_\sigma}{T_a q_\sigma} \right| \right] \tag{183}$$

$$\leq D\left(q_\sigma \| T_a q_\sigma\right) + \frac{2}{e} < \infty. \tag{184}$$

Thus, by the monotone convergence theorem, we must have

$$\lim_{z \to \infty} \int_{\mathbb{R} \setminus [-z,z]} q_\sigma \left| \log \frac{q_\sigma}{T_a q_\sigma} \right| = 0. \tag{185}$$

As this contradicts (181), the lemma is proved. ∎

The following lemma gives an $\exp(-O(w^\gamma))$ lower bound on the minimum value of q_σ over $[-w, w]$ and on the probability that $Z_\sigma \sim q_\sigma$ exceeds w, both as $w \to \infty$.

Lemma 4. *For a PDF q and a constant $\sigma > 0$, we have that*

$$\int_{[w,\infty)} q_\sigma = \exp\left(-O(w^\gamma)\right) \tag{186}$$

and

$$\min_{|x| \leq w} q_\sigma(x) = \exp\left(-O(w^\gamma)\right), \tag{187}$$

both as $w \to \infty$.

Proof: First, we show that there is a bounded Borel set B with $\lambda(B) > 0$ such that

$$\mu := \inf_{x \in B} q(x) > 0. \tag{188}$$

Note that we may remove the boundedness condition on B. Indeed, if the Borel set B satisfies $\lambda(B) > 0$ and $\inf_{x \in B} q(x) > 0$, then the bounded Borel sets $A_m := B \cap [-m, m]$ also satisfy $\lambda(A_m) > 0$ and $\inf_{x \in A_m} q(x) > 0$ for all large m by continuity of λ and the definition of the infimum. Now, to see that such a B exists, consider the Borel sets $B_n := q^{-1}([1/n, \infty))$ for integers $n \geq 1$. For each $n \geq 1$, we have that $\inf_{x \in B_n} q(x) \geq 1/n$. Suppose, for the sake of contradiction, that $\lambda(B_n) = 0$ for each n. Then we would have

$$\lambda(q^{-1}((0,\infty))) = \lambda\left(q^{-1}\left(\bigcup_{n \geq 1} [1/n, \infty) \right) \right) \tag{189}$$

$$= \lambda\left(\bigcup_{n \geq 1} B_n \right) = 0. \tag{190}$$

Hence, $q = 0$ a.e. However, this would contradict that q is a PDF. Thus, we conclude that $\lambda(B_n) > 0$ for some n. In short, there must exist a bounded Borel set B with $\lambda(B) > 0$ and $\inf_{x \in B} q(x) > 0$. Fix such a B, and let $x_0 > 0$ be such that $B \subset [-x_0, x_0]$.

Recall that we define $q_\sigma = q * \psi^\sigma$ (see equations (141)–(146)). For each $w \in \mathbb{R}$, Tonelli's theorem implies that

$$\int_{[w,\infty)} q_\sigma = \int_{\mathbb{R}} q(x) \int_w^\infty \psi\left(\frac{y-x}{\sigma}\right) \frac{1}{\sigma}\, dy\, dx. \tag{191}$$

Performing a change of variable, we have for every $x, w \in \mathbb{R}$

$$\int_w^\infty \psi\left(\frac{y-x}{\sigma}\right) \frac{1}{\sigma}\, dy = \int_{[(w-x)/\sigma, \infty)} \psi. \tag{192}$$

Further, for any $z \geq 0$, by definition of ψ, we have the bound

$$\int_{[z,\infty)} \psi \geq \int_{[z,z+1]} \psi \geq \exp\left(-(z+1)^\gamma\right) \cdot \chi^{-1}, \tag{193}$$

where $\chi = \int_{\mathbb{R}} \exp(-|u|^\gamma)\, du$ is the normalization constant for ψ. Therefore, whenever $w \geq x$ we have

$$\int_{[(w-x)/\sigma, \infty)} \psi \geq \exp\left(-\left(\frac{w - x + \sigma}{\sigma} \right)^\gamma \right) \cdot \chi^{-1}. \tag{194}$$

Now, combining (191) and (192), nonnegativity of the PDFs q and ψ implies the bound

$$\int_{[w,\infty)} q_\sigma \geq \int_B q(x) \int_{[(w-x)/\sigma, \infty)} \psi(u)\, du\, dx. \tag{195}$$

Since $B \subset [-x_0, x_0]$, we conclude from (194) that for every $w \geq x_0$

$$\int_{[w,\infty)} q_\sigma \geq \int_B \mu \cdot \exp\left(-\left(\frac{w - x + \sigma}{\sigma} \right)^\gamma \right) \cdot \chi^{-1}\, dx \tag{196}$$

$$\geq \lambda(B) \mu \chi^{-1} \cdot \exp\left(-\left(\frac{w + x_0 + \sigma}{\sigma} \right)^\gamma \right). \tag{197}$$

The estimate in (186) follows by taking $w \to \infty$.

Finally, we show that (187) holds. Let $w_0 > 0$ be such that $\int_{[-w,w]} q \geq 1/2$ for every $w \geq w_0$. Then, for any $w \geq w_0$ and $x \in [-w, w]$,

$$q_\sigma(x) = \int_{\mathbb{R}} q(u) \psi^\sigma(x-u)\, du \tag{198}$$

$$= (\sigma\chi)^{-1} \int_{\mathbb{R}} q(u)\, \exp\left(-|x-u|^\gamma / \sigma^\gamma\right)\, du \tag{199}$$

$$\geq (\sigma\chi)^{-1} \int_{-w}^{w} q(u)\, \exp\left(-|x-u|^\gamma / \sigma^\gamma\right)\, du \tag{200}$$

$$\geq (\sigma\chi)^{-1}\, \exp\left(-(2/\sigma^\gamma) w^\gamma\right) \int_{[-w,w]} q \tag{201}$$

$$\geq (2\sigma\chi)^{-1}\, \exp\left(-(2/\sigma^\gamma) w^\gamma\right). \tag{202}$$

The estimate (187) follows by taking $w \to \infty$. ∎

Conversely, the following lemma gives an upper bound on the tail of any distribution that satisfies the cost constraint.

Lemma 5. *For any $P \in \mathscr{B}$, if $\mathbb{E}_P[c] < \infty$ then*

$$P(\mathbb{R} \setminus [-x, x]) = o\left(c(x)^{-1}\right) \tag{203}$$

as $x \to \infty$.

Proof: We start by showing that

$$\lim_{t \to \infty} P\left(\{c > t\}\right) \cdot t = 0. \tag{204}$$

Denote $f(t) := P(\{c > t\})$ for $t > 0$. Note that f is a decreasing nonnegative function over $(0, \infty)$. Further, f is integrable by nonnegativity of c and by the assumption in the lemma since

$$\int_0^\infty P(\{c > t\})\, dt = \mathbb{E}_P[c] < \infty. \tag{205}$$

We show that these three properties of f yield that $f(t) = o(t^{-1})$.

Suppose, for the sake of contradiction, that there is an $\varepsilon > 0$ and an increasing sequence $t_n \nearrow \infty$ of strictly positive numbers such that

$$f(t_n) \geq \frac{\varepsilon}{t_n} \tag{206}$$

for every $n \in \mathbb{N}$. Since f is decreasing, we infer from (206) that

$$f(t) \geq \sum_{n\in\mathbb{N}} \frac{\varepsilon}{t_{n+1}} \cdot 1_{(t_n, t_{n+1}]}(t) \tag{207}$$

for every $t > t_1$. Integrating both sides in (207), integrability of f implies that

$$\infty > \int_{(t_1,\infty)} f \geq \sum_{n\in\mathbb{N}} \varepsilon \left(1 - \frac{t_n}{t_{n+1}}\right). \tag{208}$$

By convergence of the series in (208), we conclude that $t_n/t_{n+1} \sim 1$ as $n \to \infty$. In particular, the constant

$$\tau := \inf_{n\in\mathbb{N}} \frac{t_n}{t_{n+1}} \tag{209}$$

satisfies $\tau \in (0, 1)$. Set $\delta = \varepsilon\tau$, and note that $\delta > 0$. Then, from (207) we obtain

$$f(t) \cdot 1_{(t_1,\infty)}(t) \geq \sum_{n\in\mathbb{N}} \frac{\delta}{\tau \cdot t_{n+1}} \cdot 1_{(t_n, t_{n+1}]}(t) \tag{210}$$

$$\geq \sum_{n\in\mathbb{N}} \frac{\delta}{t_n} \cdot 1_{(t_n, t_{n+1}]}(t) \tag{211}$$

$$\geq \sum_{n\in\mathbb{N}} \frac{\delta}{t} \cdot 1_{(t_n, t_{n+1}]}(t) = \frac{\delta}{t} \cdot 1_{(t_1,\infty)}(t). \tag{212}$$

However, (212) contradicts the integrability of f. Thus, we conclude that it must be the case that

$$f(t) = o\left(t^{-1}\right) \tag{213}$$

as $t \to \infty$.

To finish the proof of the lemma, recall by the tail-regularity assumption on c that we have

$$\lim_{|x|\to\infty} \frac{c(x)}{|x|^\alpha} = \beta \tag{214}$$

for some $\alpha, \beta > 0$. Thus there are $0 < \beta_1 < \beta < \beta_2$ such that for sufficiently large $|x|$, $\beta_1|x|^\alpha \leq c(x) \leq \beta_2|x|^\alpha$. Then, for all large enough t, we have that

$$P(\{c > t\}) \geq P\left(\mathbb{R} \setminus \left[-(t/\beta_1)^{1/\alpha}, (t/\beta_1)^{1/\alpha}\right]\right). \tag{215}$$

Writing $x = (t/\beta_1)^{1/\alpha}$, we conclude that for all large x

$$c(x)P(\mathbb{R} \setminus [-x, x]) \leq \beta_2 x^\alpha P(\mathbb{R} \setminus [-x, x]) \tag{216}$$

$$= \frac{\beta_2}{\beta_1} tP\left(\mathbb{R} \setminus \left[-(t/\beta_1)^{1/\alpha}, (t/\beta_1)^{1/\alpha}\right]\right) \tag{217}$$

$$\leq \frac{\beta_2}{\beta_1} tP(\{c > t\}). \tag{218}$$

Taking $t \to \infty$, we obtain from (213) that

$$c(x)P(\mathbb{R} \setminus [-x, x]) \to 0, \tag{219}$$

as desired. ∎

The final auxiliary lemma gives an upper bound on the tail of the cost constraint incurred by a cactus distribution.

Lemma 6. *Fix $r \in (0, 1)$ and integers $N > n \geq 1$, and set $w = (N - 1/2)/n$. Assume that $c(x) \leq \beta_1 x^\alpha$ for $x \geq w$. Then, we have the bound*

$$\sum_{i\geq N} c_{n,i} r^{i-N} \leq \beta_1 \ell_\alpha \left(\frac{w^\alpha}{1-r} + \frac{2\left(\frac{\alpha}{e}\right)^\alpha \log\frac{1}{r} + \Gamma(\alpha+1)}{rn^\alpha \left(\log\frac{1}{r}\right)^{\alpha+1}}\right), \tag{220}$$

where $\ell_\alpha := \max(1, 2^{\alpha-1})$.

Proof: By monotonicity of c,

$$\sum_{i\geq N} c_{n,i} r^{i-N} = \sum_{i\geq N} \int_{(i-1/2)/n}^{(i+1/2)/n} ncr^{i-N} \tag{221}$$

$$\leq \sum_{i\geq N} \beta_1 \left(\frac{i+1/2}{n}\right)^\alpha r^{i-N} \tag{222}$$

$$= \beta_1 \sum_{i\geq 0} \left(w + \frac{i+1}{n}\right)^\alpha r^i \tag{223}$$

$$\leq \beta_1 \ell_\alpha \left(\frac{w^\alpha}{1-r} + \frac{\mathrm{Li}_{-\alpha}(r)}{rn^\alpha}\right), \tag{224}$$

where

$$\mathrm{Li}_{-\alpha}(r) := \sum_{k\geq 1} k^\alpha r^k \tag{225}$$

is the polylogarithm function. To finish the proof of the lemma, we show next that

$$\mathrm{Li}_{-\alpha}(r) \leq 2\left(\frac{\alpha}{e\log\frac{1}{r}}\right)^\alpha + \frac{\Gamma(\alpha+1)}{\left(\log\frac{1}{r}\right)^{\alpha+1}}. \tag{226}$$

Now, consider the function $g : (0, \infty) \to (0, \infty)$ defined by

$$g(x) := x^\alpha r^x. \tag{227}$$

We have that

$$g'(x) = (\alpha + x\log r)\, x^{\alpha-1} r^x. \tag{228}$$

Thus, g increases until it reaches a maximum at $x_0 = \alpha/\log\frac{1}{r}$ then it decreases. Thus,

$$\mathrm{Li}_{-\alpha}(r) \leq g(\lfloor x_0 \rfloor) + g(\lceil x_0 \rceil) + \int_{(0,\infty)} g. \tag{229}$$

We have

$$g(\lfloor x_0 \rfloor) + g(\lceil x_0 \rceil) \leq 2g(x_0) = 2\left(\frac{\alpha}{e\log\frac{1}{r}}\right)^\alpha, \tag{230}$$

and

$$\int_{(0,\infty)} g = \frac{\Gamma(\alpha+1)}{\left(\log \frac{1}{r}\right)^{\alpha+1}}. \tag{231}$$

The proof is thus complete. ∎

B. Proof of Theorem 3

By Theorem 1, there is a PDF $q^\star$ that satisfies both

$$\sup_{|a|\le 1} D(q^\star \| T_a q^\star) = \mathrm{KL}^\star, \tag{232}$$

$$\mathbb{E}_{q^\star}[c] \le C. \tag{233}$$

We may assume that $q^\star$ is even; indeed, we may replace $q^\star$ with the even PDF $(q^\star(x)+q^\star(-x))/2$, which satisfies the cost constraint by evenness of c, and which also has a better KL-divergence than that of $q^\star$ by joint convexity of the KL-divergence. Fix arbitrary constants $\delta, \eta > 0$, and we will find a cactus distribution that attains the KL-divergence (232) to within δ and the cost (233) to within η.

By Lemma 2, there is a $\sigma > 0$ such that the PDF $q^\star_\sigma$ satisfies the bounds

$$\sup_{|a|\le 1} D(q^\star_\sigma \| T_a q^\star_\sigma) \le \mathrm{KL}^\star, \tag{234}$$

$$\mathbb{E}_{q^\star_\sigma}[c] \le C + \frac{\eta}{2}. \tag{235}$$

Throughout the proof, we will denote

$$q := q^\star_\sigma \tag{236}$$

for short. Let

$$Q(B) := \int_B q \tag{237}$$

be the probability measure induced by q. We will construct a cactus distribution that approximates q.

We first note a few properties of q. Note that q is an even PDF. Further, it is uniformly continuous, and strictly positive over $\mathbb{R}$. Thus, q is locally bounded away from zero. For each $z \ge 0$, denote the minimum

$$\mu_z := \min_{|x|\le z} q(x), \tag{238}$$

so $\mu_z > 0$ for every z. In addition, q is upper bounded: by Young's inequality, we have that

$$\|q\|_{L^\infty(\mathbb{R})} = \|q^\star * \psi^\sigma\|_{L^\infty(\mathbb{R})} \tag{239}$$

$$\le \|q^\star\|_{L^1(\mathbb{R})} \cdot \|\psi^\sigma\|_{L^\infty(\mathbb{R})} \tag{240}$$

$$= (\sigma\chi)^{-1} =: M. \tag{241}$$

In fact, q satisfies a property resembling local γ-Hölder continuity. Specifically, as in the proof of Lemma 3 (see (171)–(175)), we have that

$$q(x) \le e^{|x-y|^\gamma/\sigma^\gamma} q(y) \tag{242}$$

for every $x, y \in \mathbb{R}$. Therefore, for some $|t_{x,y}| \le 1$ we have

$$|q(x) - q(y)| = q(y)\left|e^{t_{x,y}|x-y|^\gamma/\sigma^\gamma} - 1\right| \tag{243}$$

$$\le \frac{2M}{\sigma^\gamma}|x-y|^\gamma, \tag{244}$$

where the latter inequality follows whenever $|x-y| \le \sigma$. In particular, for all $\varepsilon \in (0, 2M)$, we have that

$$|q(x)-q(y)| \le \varepsilon \quad \text{whenever} \quad |x-y| \le \sigma\cdot\left(\frac{\varepsilon}{2M}\right)^{1/\gamma}. \tag{245}$$

Before constructing the parameters (n, N, r) of the cactus distribution, we note a fundamental lower bound on n. For the cost constraint to be satisfied, we need $c_{n,0} < C$ to hold. Nevertheless, by continuity of c, every real number is a Lebesgue point of c. In particular, as 0 is a Lebesgue point of c, we obtain

$$c_{n,0} = \frac{\int_{[-1/(2n),1/(2n)]} c}{1/n} \to c(0) = 0 \tag{246}$$

as $n \to \infty$. Let $n_{\min}$ be the least positive integer such that

$$c_{n,0} < C \tag{247}$$

for every $n \ge n_{\min}$. Note that $n_{\min}$ depends only on c and C.

Now, we choose the integers n and N. Denote the constants

$$\theta_\alpha := 4\left(\frac{\alpha}{e}\right)^\alpha + 2\Gamma(\alpha+1) \tag{248}$$

$$\theta'_\alpha := (2\theta_\alpha)^{1/\alpha} \tag{249}$$

$$\gamma' := \frac{\gamma+\alpha}{2} \in (\gamma, \alpha) \tag{250}$$

$$\varepsilon_{\min} := 2M \cdot \min\left(\frac{2}{\sigma n_{\min}}, \frac{1}{\theta'_\alpha \sigma}\right)^\gamma \tag{251}$$

$$z_{\min,0} := \left(\log\left(\frac{4}{\sigma}\cdot\left(\frac{2M}{\varepsilon_{\min}}\right)^{1/\gamma}\right)\right)^{1/\gamma'} \tag{252}$$

$$z_{\min,1} := \left(\frac{\eta}{\delta}\cdot\frac{2e\theta'_\alpha}{\beta_1\ell_\alpha}\right)^{1/(\alpha+1)} \tag{253}$$

$$z_{\min,2} := \left(\frac{2^{\alpha+1}}{\beta_1\ell_\alpha}\right)^{1/(\alpha-\gamma')} \tag{254}$$

$$z_{\min} := \max\left(z_{\min,0}, z_{\min,1}, z_{\min,2}, \frac{\delta}{12M}\right). \tag{255}$$

Since $q = q^\star_\sigma$ (see (236)), Lemma 3 yields the existence of a constant $z_0 > 0$ such that $z \ge z_0$ implies the uniform bound

$$\sup_{|a|\le 1}\int_{\mathbb{R}\setminus[-z,z]} q\left|\log\frac{q}{T_a q}\right| \le \frac{\delta}{3}. \tag{256}$$

In addition, Lemma 4 yields the existence of constants $\tau, z_1 > 0$ such that $z \ge z_1$ implies (see (237) and (238))

$$\min\left(\mu_z, Q([z,\infty))\right) \ge \exp\left(-\tau z^\gamma\right). \tag{257}$$

By the tail-regularity assumption on c, there are constants $\beta_1, \beta_2, z_2 > 0$ such that

$$\beta_2 z^\alpha \le c(z) \le \beta_1 z^\alpha \tag{258}$$

for every $z \ge z_2$. By Lemma 5, we have that (see (237))

$$\lim_{z\to\infty} Q(\mathbb{R}\setminus[-z,z])c(z) = 0. \tag{259}$$

Let $z_3 > 0$ be large enough that $z \ge z_3$ implies

$$Q\left(\mathbb{R}\setminus[-z,z]\right)c(z) \le \frac{\beta_2}{\beta_1\ell_\alpha}\cdot\frac{\eta}{6}. \tag{260}$$

If $z \geq \max(z_2, z_3)$, then by (258) and (260) we may bound the tail of Q also by

$$Q\left(\mathbb{R} \setminus [-z, z]\right) \leq \frac{1}{\beta_1 \ell_\alpha z^\alpha} \cdot \frac{\eta}{6}. \tag{261}$$

Let $z_4 > 0$ be the smallest number such that both inequalities

$$e^{\tau z^\gamma} \geq \frac{\delta \beta_1 \ell_\alpha}{2M\eta} \cdot z^\alpha \tag{262}$$

$$e^{\gamma z^{\gamma'}} \geq \left(\frac{4}{\sigma}\right)^\gamma \cdot \frac{48M^2}{\delta} \cdot z e^{\tau z^\gamma} \tag{263}$$

hold for all $z \geq z_4$. Fix a rational number

$$z > \max(z_{\min}, z_0, z_1, z_2, z_3, z_4, 2\theta'_\alpha) \tag{264}$$

that is a ratio of an odd integer by an even integer, and set

$$w := z + 1. \tag{265}$$

We choose z (hence also w) here to belong in $\mathbb{N} + \frac{1}{2}$ for simplicity, but we note that any other choice (of denominator) is also valid provided that w is increased so that the subsequent choices in (270) below can be made. Set

$$\varepsilon := \frac{2^{2\gamma+1} M}{\sigma^\gamma} \cdot e^{-\gamma w^{\gamma'}}. \tag{266}$$

Denote

$$n_0 := \frac{2}{\sigma} \cdot \left(\frac{2M}{\varepsilon}\right)^{1/\gamma}, \tag{267}$$

By the uniform continuity of q shown in (245), we have that

$$|q(x) - q(y)| \leq \varepsilon \quad \text{whenever} \quad |x - y| \leq \frac{2}{n_0}. \tag{268}$$

Note that $n_{\min} < n_0$ since $\varepsilon < \varepsilon_{\min}$, which in turn follows because $w > z_{\min,0}$. We note also that $\varepsilon < \varepsilon_{\min}$ implies $2\theta'_\alpha < n_0$. Set

$$n_1 := e^{w^{\gamma'}}. \tag{269}$$

By construction, we have that $n_1 = 2n_0$. Thus, we may choose integers $n \in [n_0, n_1]$ and $N > n$ such that

$$w = \frac{2N - 1}{2n} \tag{270}$$

Next, we choose the parameter r, thereby completing the cactus distribution construction. Define, for $i \in \{0, \cdots, N-1\}$,

$$p_i := \inf_{x \in \mathcal{J}_{n,i}} \frac{q(x)}{n}. \tag{271}$$

By evenness, continuity, and strict positivity of q, we have that

$$p_0 + \sum_{i=1}^{N-1} 2p_i = \int_{[-w,w]} \sum_{|i| \leq N-1} n p_{|i|} \cdot 1_{\mathcal{J}_{n,i}} \leq \int_{[-w,w]} q < 1. \tag{272}$$

Thus, for any $r \in (0, 1)$, setting

$$p_N := \frac{1 - r}{2} \left(1 - \left(p_0 + \sum_{i=1}^{N-1} 2p_i\right)\right), \tag{273}$$

we infer from (272) that the vector $\boldsymbol{p} = (p_0, \cdots, p_N)$ belongs to $(0, 1]^{N+1}$, and by construction it satisfies $S_{r,\boldsymbol{p}} = 1$. We will choose r as

$$r := 1 - \frac{\theta'_\alpha}{wn}, \tag{274}$$

and define p_N as in (273) for this choice of r.

Therefore, $f_{n,r,\boldsymbol{p}}$ is a valid cactus distribution. By uniform continuity of q (see (268)) and by definition of the p_i (see (271)), we have that $f_{n,r,\boldsymbol{p}}$ uniformly approximates q from below over $[-w, w]$: for every $x \in [-w, w]$ we have that

$$0 \leq q(x) - f_{n,r,\boldsymbol{p}}(x) \leq \varepsilon. \tag{275}$$

We will deduce from the uniform bound (275) that $f_{n,r,\boldsymbol{p}}$ approximates q in the two senses:

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c] \leq \mathbb{E}_q[c] + \frac{\eta}{2} \tag{276}$$

and

$$\sup_{|a| \leq 1} D(f_{n,r,\boldsymbol{p}} \| T_a f_{n,r,\boldsymbol{p}}) \leq \sup_{|a| \leq 1} D(q \| T_a q) + \delta. \tag{277}$$

Combined with (234)–(235), we would conclude from (276)–(277) that

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c] \leq C + \eta \tag{278}$$

and

$$\sup_{|a| \leq 1} D(f_{n,r,\boldsymbol{p}} \| T_a f_{n,r,\boldsymbol{p}}) \leq \mathrm{KL}^\star + \delta. \tag{279}$$

Now, we show that $f_{n,r,\boldsymbol{p}}$ satisfies the cost constraint (278). Since $f_{n,r,\boldsymbol{p}}|_{[-w,w]} \leq q|_{[-w,w]}$, we have that

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c \cdot 1_{[-w,w]}] \leq \mathbb{E}_q[c] \leq C + \frac{\eta}{2}. \tag{280}$$

We show next that

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c \cdot 1_{\mathbb{R} \setminus [-w,w]}] \leq \frac{\eta}{2}. \tag{281}$$

By construction of $f_{n,r,\boldsymbol{p}}$, and since $w = (N - 1/2)/n$ (see (270)), we have the expression

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c \cdot 1_{\mathbb{R} \setminus [-w,w]}] = 2p_N \sum_{i \geq N} c_{n,i} r^{i-N}. \tag{282}$$

We bound the terms $2p_N$ and $\sum_{i \geq N} c_{n,i} r^{i-N}$ separately. By Lemma 6, we have the bound

$$\sum_{i \geq N} c_{n,i} r^{i-N} \leq \beta_1 \ell_\alpha \left(\frac{w^\alpha}{1 - r} + \frac{2\left(\frac{\alpha}{e}\right)^\alpha \log \frac{1}{r} + \Gamma(\alpha + 1)}{r n^\alpha \left(\log \frac{1}{r}\right)^{\alpha+1}}\right). \tag{283}$$

By definition of r (see (274)), and since $w \geq 1$ and $n \geq 2\theta'_\alpha$, we have that $r \geq 1/2 > 1/e$. Thus, we deduce from (283) that

$$\sum_{i \geq N} c_{n,i} r^{i-N} \leq \beta_1 \ell_\alpha \left(\frac{w^\alpha}{1 - r} + \frac{\theta_\alpha}{n^\alpha \left(\log \frac{1}{r}\right)^{\alpha+1}}\right), \tag{284}$$

where θ_α is as defined in (248). In addition, we have that (recall that we denote by $P_{n,r,\boldsymbol{p}}$ the probability measure associated with $f_{n,r,\boldsymbol{p}}$)

$$\frac{2p_N}{1 - r} = P_{n,r,\boldsymbol{p}}\left(\mathbb{R} \setminus [-w, w]\right) = 1 - P_{n,r,\boldsymbol{p}}\left([-w, w]\right). \tag{285}$$

As $f_{n,r,\boldsymbol{p}}$ uniformly approximates q from below over $[-w, w]$

to within ε (see (275)), we have that

$$P_{n,r,\boldsymbol{p}}\left([-w,w]\right) \geq Q\left([-w,w]\right) - 2\varepsilon w. \tag{286}$$

Thus, by the bound on the tail of Q in (261)

$$\frac{2p_N}{1-r} \leq Q\left(\mathbb{R}\setminus[-w,w]\right) + 2\varepsilon w \tag{287}$$

$$\leq \frac{1}{\beta_1 \ell_\alpha w^\alpha} \cdot \frac{\eta}{6} + 2\varepsilon w. \tag{288}$$

Further, combining inequalities (262)–(263) and using the definition of ε in (266), we obtain

$$\varepsilon \leq \frac{\eta}{12\beta_1 \ell_\alpha w^{\alpha+1}}. \tag{289}$$

Thus, we deduce

$$2p_N \leq \frac{\eta\cdot(1-r)}{3\beta_1\ell_\alpha w^\alpha}. \tag{290}$$

From the expression in (282), multiplying inequalities (284) and (290) and noting that $1-r \leq \log\frac{1}{r}$, we obtain

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c\cdot 1_{\mathbb{R}\setminus[-w,w]}] \leq \frac{\eta}{3}\left(1 + \frac{\theta_\alpha}{\left(wn\log\frac{1}{r}\right)^\alpha}\right). \tag{291}$$

By definition of r, we have that

$$\log\frac{1}{r} \geq 1 - r = \frac{\theta'_\alpha}{wn}. \tag{292}$$

Using inequality (292) in (291), we obtain

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c\cdot 1_{\mathbb{R}\setminus[-w,w]}] \leq \frac{\eta}{3}\cdot\frac{3}{2} = \frac{\eta}{2}, \tag{293}$$

which is inequality (281). Combining (280)–(281), we deduce (278), i.e.,

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c] \leq C + \eta. \tag{294}$$

Next, we show that $f_{n,r,\boldsymbol{p}}$ satisfies the KL bound (279). We begin by splitting the integration at the points $\pm z$. By finiteness of the considered KL-divergences, we have for each $|a|\leq 1$

$$\begin{aligned} &D(f_{n,r,\boldsymbol{p}}\|T_{-a}f_{n,r,\boldsymbol{p}}) - D(q\|T_{-a}q) \\ &\quad\leq \int_{[-z,z]}\left(f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} - q\log\frac{q}{T_{-a}q}\right) \\ &\qquad + \int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} \\ &\qquad + \int_{\mathbb{R}\setminus[-z,z]} q\log\frac{q}{T_{-a}q}. \end{aligned} \tag{295}$$

We already have a uniform bound for the last integral in (295): since $z\geq z_0$, the estimate in (256) holds and we obtain

$$\sup_{|a|\leq 1}\int_{\mathbb{R}\setminus[-z,z]} q\log\frac{q}{T_a q} \leq \frac{\delta}{3}. \tag{296}$$

We proceed to bounding the first integral in (295) uniformly by

$$\sup_{|a|\leq 1}\int_{[-z,z]}\left(f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} - q\log\frac{q}{T_{-a}q}\right) \leq \frac{\delta}{3}. \tag{297}$$

We do this via deriving an upper bound on the integrand that is uniform in both a and the variable of integration. From $w\geq\delta/(12M)$ (255), $\mu_w \geq e^{-\tau w^\gamma}$ (257), and (263), we have that

$$\varepsilon \leq \frac{\mu_w}{2}\cdot\min\left(1, \frac{\delta}{12M}\right). \tag{298}$$

Define the function $g:[-w,w]\to[0,\varepsilon]$ by

$$g := q - f_{n,r,\boldsymbol{p}}. \tag{299}$$

That the range of g is contained within $[0,\varepsilon]$ follows since $f_{n,r,\boldsymbol{p}}$ approximates q from below uniformly over $[-w,w]$ to within ε. Thus, $z = w-1$ yields

$$\sup_{|a|\leq 1}\|T_a g\|_{L^\infty([-z,z])} \leq \varepsilon. \tag{300}$$

We note that, over $[-z,z]$, the inequality

$$\begin{aligned} &f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} - q\log\frac{q}{T_{-a}q} \\ &\leq -q\log\left(1 - T_{-a}\frac{g}{q}\right) - g\log\left(1-\frac{g}{q}\right) + g\log\frac{T_{-a}q}{q} \end{aligned} \tag{301}$$

holds; that all the logarithms are well defined follows since $g\leq q$ over $[-w,w]$. Indeed, subtracting the left hand side from the right hand side in (301), we get the function

$$-q\log\left(1-\frac{g}{q}\right) - g\log\left(1 - T_{-a}\frac{g}{q}\right), \tag{302}$$

which is nonnegative over $[-z,z]$ since g is nonnegative over $[-w,w]$. Now, we bound each of the terms in (301). It is easy to see that for $0\leq t\leq 1/2$ one has

$$-\log(1-t) \leq 2t. \tag{303}$$

Now, we show that $g/q\leq 1/2$ over $[-w,w]$. Indeed, this is equivalent to $q\leq 2f_{n,r,\boldsymbol{p}}$ over $[-w,w]$. But $q-\varepsilon\leq f_{n,r,\boldsymbol{p}}$ over $[-w,w]$, which implies in view of $\varepsilon\leq\mu_w/2\leq q/2$ (over $[-w,w]$) that $q\leq 2f_{n,r,\boldsymbol{p}}$, as desired. Thus, we obtain that over $[-z,z]$

$$-q\log\left(1 - T_{-a}\frac{g}{q}\right) \leq 2qT_{-a}\frac{g}{q} \leq \frac{2M\varepsilon}{\mu_w}, \tag{304}$$

and

$$-g\log\left(1-\frac{g}{q}\right) \leq \frac{2g^2}{q} \leq \frac{2\varepsilon^2}{\mu_w} \leq \varepsilon. \tag{305}$$

It is also clear that over $[-z,z]$

$$g\log\frac{T_{-a}q}{q} \leq \varepsilon\log\frac{M}{\mu_w} \leq \varepsilon\left(\frac{M}{\mu_w} - 1\right). \tag{306}$$

Plugging in inequalities (304)–(306) into (301), we obtain the uniform bound

$$f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} - q\log\frac{q}{T_{-a}q} \leq \frac{3M\varepsilon}{\mu_w} \tag{307}$$

over $[-z,z]$. Integrating, we deduce

$$\int_{[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} - q\log\frac{q}{T_{-a}q} \leq \frac{6zM\varepsilon}{\mu_w} \tag{308}$$

$$< \frac{\delta}{3}, \tag{309}$$

where (309) follows by (298).

It remains to upper bound the middle integral in (301), for which we also derive a uniform upper bound

$$\sup_{|a|\le 1}\int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}}\le\frac{\delta}{3}. \tag{310}$$

We will further split the integration at the points $\pm(w+1)$. By evenness of $f_{n,r,\boldsymbol{p}}$, we have that this integral depends only on $|a|$, i.e., for each $a\in[-1,1]$

$$\int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}}=\int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{a}f_{n,r,\boldsymbol{p}}}. \tag{311}$$

Thus, it suffices for (310) to show that

$$\sup_{0<a\le 1}\int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}}\le\frac{\delta}{3}. \tag{312}$$

Consider first the integral

$$\int_{\mathbb{R}\setminus[-(w+1),w+1]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} \tag{313}$$

for fixed $a\in(0,1]$. From the proof of Theorem 2, we can write the integrand in (313) as follows. Extend the definition of p_i to all $i\in\mathbb{Z}$ by

$$p_i:=\begin{cases} p_{|i|}, & \text{if } -N\le i\le -1,\\ p_N r^{|i|-N}, & \text{if } |i|>N.\end{cases} \tag{314}$$

For each $i\in\mathbb{Z}$, there is an integer j with $|j|\le n$, such that we have

$$f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}}=np_i\log\frac{p_i}{p_{i+j}} \tag{315}$$

over $\mathcal{J}_{n,i}$ except possibly at a single point. By definition of w, we have that

$$\mathbb{R}\setminus[-(w+1),w+1]=\bigcup_{|i|\ge N+n}\mathcal{J}_{n,i}. \tag{316}$$

Further, if $|i|\le N+n$ and $|j|\le n$, then $|i+j|\ge N$. Hence, from (315) we have that over $\mathcal{J}_{n,i}$ with $|i|\ge N+n$

$$f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}}=np_Nr^{|i|-N}(|i|-|i+j|)\log r \tag{317}$$

$$\le n^2p_Nr^{|i|-N}\log\frac{1}{r}. \tag{318}$$

Summing over $|i|\ge N+n$, we obtain

$$\int_{\mathbb{R}\setminus[-(w+1),w+1]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}}$$

$$=\sum_{|i|\ge N+n}\int_{J_{n,i}} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} \tag{319}$$

$$\le np_N\log\frac{1}{r}\sum_{|i|\ge n}r^{|i|}=\frac{2np_Nr^n\log\frac{1}{r}}{1-r}. \tag{320}$$

Using the upper bound on p_N in (290), we obtain that

$$\int_{\mathbb{R}\setminus[-(w+1),w+1]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}}\le\frac{\eta nr^n\log\frac{1}{r}}{3\beta_1\ell_\alpha w^\alpha}. \tag{321}$$

As $1/e\le r\le 1$ and $\log\frac{1}{r}\le\frac{1}{r}-1$, using the definition of r given in (274) and $w\ge z_4$ (see (255)), we have the bound

$$\frac{\eta nr^n\log\frac{1}{r}}{3\beta_1\ell_\alpha w^\alpha}\le\frac{e\eta n(1-r)}{3\beta_1\ell_\alpha w^\alpha}\le\frac{e\eta\theta'_\alpha}{3\beta_1\ell_\alpha w^{\alpha+1}}\le\frac{\delta}{6}. \tag{322}$$

Thus, we have shown that

$$\sup_{a\in(0,1]}\int_{\mathbb{R}\setminus[-(w+1),w+1]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}}\le\frac{\delta}{6}. \tag{323}$$

The final integral bound we need is the following:

$$\sup_{0<a\le 1}\int_{w-1<|x|\le w+1} f_{n,r,\boldsymbol{p}}(x)\log\frac{f_{n,r,\boldsymbol{p}}(x)}{T_{-a}f_{n,r,\boldsymbol{p}}(x)}\,dx\le\frac{\delta}{6}. \tag{324}$$

By evenness of $f_{n,r,\boldsymbol{p}}$, we have that

$$\int_{w-1<|x|\le w+1} f_{n,r,\boldsymbol{p}}(x)\log\frac{f_{n,r,\boldsymbol{p}}(x)}{T_{-a}f_{n,r,\boldsymbol{p}}(x)}\,dx \tag{325}$$

$$=\int_{(w-1,w+1]} f_{n,r,\boldsymbol{p}}\log\frac{f^2_{n,r,\boldsymbol{p}}}{(T_{-a}f_{n,r,\boldsymbol{p}})\cdot(T_af_{n,r,\boldsymbol{p}})}. \tag{326}$$

Consider the function inside the logarithm in the integrand:

$$\rho(x;a):=\frac{f_{n,r,\boldsymbol{p}}(x)^2}{f_{n,r,\boldsymbol{p}}(x+a)f_{n,r,\boldsymbol{p}}(x-a)}. \tag{327}$$

We will prove the uniform upper bound

$$\sup_{\substack{x\in(w-1,w+1]\\ a\in(0,1]}}\rho(x;a)\le\exp\left(2w^{\gamma'}\right), \tag{328}$$

where $\gamma':=(\gamma+\alpha)/2\in(\gamma,\alpha)$ is as defined in (250). Note that

$$(w-1,w+1]=\bigcup_{i=N-n}^{N+n}\mathcal{J}_{n,i}. \tag{329}$$

For each $a\in(0,1]$ and $x\in(w-1,w+1]$, there are integers $N-n\le i\le N+n$ and $0\le j,k\le n$ such that

$$\rho(x;a)=\frac{p_i^2}{p_{i+j}p_{i-k}}. \tag{330}$$

Thus, it suffices to show that $\exp(w^{\gamma'})$ is an upper bound on each of the terms

$$\frac{p_i}{p_j},\ \frac{p_k}{p_Nr^n},\ \frac{p_N}{p_k},\ \frac{1}{r^n} \tag{331}$$

for $0\le i,j,k\le N-1$ with $|i-j|\le n$. First, for $1/r^n$, denoting $m=nw/(2\theta_\alpha)^{1/\alpha}\ge 2$, we have the bound

$$r^n=\left(\left(1-\frac{1}{m}\right)^m\right)^{(2\theta_\alpha)^{1/\alpha}/w}\ge 4^{-(2\theta_\alpha)^{1/\alpha}/w}\ge\frac{1}{2}. \tag{332}$$

Hence,

$$\frac{1}{r^n}\le 2\le e^{w^{\gamma'}}. \tag{333}$$

For p_k/p_N with $0\le k\le N-1$, we have the bound

$$\frac{p_k}{p_N}\le\frac{M}{np_N}=\frac{2M/(1-r)}{n\cdot(2p_N/(1-r))}=\frac{2M/(1-r)}{nP_{n,r,\boldsymbol{p}}(\mathbb{R}\setminus[-w,w])} \tag{334}$$

$$\le\frac{2M/(1-r)}{nQ(\mathbb{R}\setminus[-w,w])}\le\frac{M/(1-r)}{ne^{-\tau w^\gamma}}=\frac{Mwe^{\tau w^\gamma}}{\theta'_\alpha}. \tag{335}$$

Hence,

$$\frac{p_k}{p_N r^n} \le \frac{2Mwe^{\tau w^\gamma}}{\theta'_\alpha} \le e^{w^{\gamma'}}, \tag{336}$$

where the last inequality follows from (263) for all small δ, e.g., for

$$\delta \le 3 \cdot 2^{2\gamma+2} \cdot \theta'_\alpha \cdot \chi^{-1} \tag{337}$$

(alternatively, we may increase the size of w at the outset). Consider next p_i/p_j for $0 \le i, j \le N-1$ with $|i-j| \le n$. By definition of the p_k and uniform continuity of q, we have for $0 \le k \le N-2$

$$|p_k - p_{k+1}| \le \frac{\varepsilon}{n}. \tag{338}$$

By the triangle inequality, we deduce

$$|p_i - p_j| \le \frac{|i-j|\varepsilon}{n} \le \varepsilon. \tag{339}$$

Thus,

$$\frac{p_i}{p_j} \le 1 + \frac{\varepsilon}{p_j} \le 1 + \frac{n\varepsilon}{\mu_w} \le 1 + \frac{n}{2} \le e^{w^{\gamma'}}. \tag{340}$$

The last term p_N/p_k can be bounded using (290) to obtain

$$\frac{p_N}{p_k} \le \frac{\eta \cdot (1-r)/(6\beta_1 \ell_\alpha w^\alpha)}{\mu_w/n} = \frac{\eta\theta'_\alpha}{6\beta_1\ell_\alpha\mu_w w^{\alpha+1}} \tag{341}$$

$$\le \frac{\eta\theta'_\alpha}{6\beta_1\ell_\alpha} \cdot e^{\tau w^\gamma} \le e^{w^{\gamma'}}, \tag{342}$$

where the last inequality follows from (263) for all small η, e.g., for

$$\eta \le 24\beta_1\ell_\alpha \cdot \chi^{-1} \cdot (\theta'_\alpha)^{-2} \tag{343}$$

(alternatively, we may increase the size of w at the outset). Collecting (333), (336), (340), and (342), we obtain the following upper on the integral in (326):

$$P_{n,r,\boldsymbol{p}}((w-1, w+1]) \cdot 2w^{\gamma'}. \tag{344}$$

Further,

$$\begin{aligned}
&P_{n,r,\boldsymbol{p}}((w-1, w+1]) \\
&\le P_{n,r,\boldsymbol{p}}((w-1, w]) + P_{n,r,\boldsymbol{p}}((w, \infty)) && (345)\\
&\le Q((w-1, w]) + \frac{1}{2} - P_{n,r,\boldsymbol{p}}([0, w]) && (346)\\
&\le Q((w-1, w]) + \frac{1}{2} - (Q([0, w]) - \varepsilon w) && (347)\\
&= \varepsilon w + Q((z, \infty)) && (348)\\
&\le \varepsilon w + \frac{\eta}{12\beta_1\ell_\alpha z^\alpha} && (349)\\
&\le \frac{\eta}{6\beta_1\ell_\alpha z^\alpha}, && (350)
\end{aligned}$$

where the last inequality follows by (289). Hence, the integral in (326) is upper bounded by

$$\frac{2^\alpha}{3\beta_1\ell_\alpha w^{\alpha-\gamma'}} \cdot \eta \le \frac{\eta}{6}, \tag{351}$$

where the last inequality follows since $w \ge z_{\min}$ (see (255)). Thus, we have shown that (324) holds, which when combined with (323) gives (310).

Combining (296), (297), and (310) gives, in view of (295), the desired inequality (279):

$$\sup_{|a|\le 1} D(f_{n,r,\boldsymbol{p}} \| T_a f_{n,r,\boldsymbol{p}}) \le \mathrm{KL}^\star + \delta. \tag{352}$$

Recall that we showed in (278) that

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c] \le C + \eta. \tag{353}$$

To sum up, denoting (see (18)) for $C' > 0$

$$\mathrm{KL}^\star_{n,N,r}(C') := \inf_{\substack{P \in \mathscr{C}_{n,N,r} \\ \mathbb{E}_P[c] \le C'}} \sup_{|a|\le 1} D(P \| T_a P) \tag{354}$$

what we have shown above yields that

$$\mathrm{KL}^\star_{n,N,r}(C + \eta) \le \mathrm{KL}^\star + \delta. \tag{355}$$

Denoting (see (21))

$$\mathrm{KL}^\star_{\text{Cactus}}(C') := \inf_{(n,N,r) \in \mathbb{N}^2 \times (0,1)} \mathrm{KL}^\star_{n,N,r}(C'), \tag{356}$$

we conclude that

$$\mathrm{KL}^\star_{\text{Cactus}}(C + \eta) \le \mathrm{KL}^\star + \delta. \tag{357}$$

Letting $\mathrm{KL}^\star(C')$ be defined analogously, we have that

$$\mathrm{KL}^\star(C + \eta) \le \mathrm{KL}^\star_{\text{Cactus}}(C + \eta) \le \mathrm{KL}^\star(C) + \delta. \tag{358}$$

Taking $\delta \to 0^+$, we have

$$\mathrm{KL}^\star(C + \eta) \le \mathrm{KL}^\star_{\text{Cactus}}(C + \eta) \le \mathrm{KL}^\star(C). \tag{359}$$

Finally, being the infimum of a jointly convex function over a convex set, the function $C \mapsto \mathrm{KL}^\star(C)$ is convex. Since it is also finite, we see that $\mathrm{KL}^\star(C)$ is continuous over $(0, \infty)$. Thus, taking $\eta \to 0^+$, we see that

$$\mathrm{KL}^\star_{\text{Cactus}} = \mathrm{KL}^\star, \tag{360}$$

completing the proof of the theorem.