Approximating L_1-distances between mixture distributions using random projections

Satyaki Mahalanabis Daniel Štefankovič

Department of Computer Science
University of Rochester
Rochester, NY 14627
{smahalan,stefanko}@cs.rochester.edu

Abstract

We consider the problem of computing L_1-distances between every pair of probability densities from a given family. We point out that the technique of Cauchy random projections [Ind06] in this context turns into stochastic integrals with respect to Cauchy motion.

For piecewise-linear densities these integrals can be sampled from if one can sample from the stochastic integral of the function $x \mapsto (1, x)$. We give an explicit density function for this stochastic integral and present an efficient (exact) sampling algorithm. As a consequence we obtain an efficient algorithm to approximate the L_1-distances with a small relative error.

For piecewise-polynomial densities we show how to approximately sample from the distributions resulting from the stochastic integrals. This also results in an efficient algorithm to approximate the L_1-distances, although our inability to get exact samples worsens the dependence on the parameters.

1 Introduction

Consider a finite class $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$ of probability densities. We want to compute the distance between every pair of members of $\mathcal{F}$. We are interested in the case where each member of $\mathcal{F}$ is a mixture of finitely many probability density functions, each having a particular functional form (e. g., uniform, linear, exponential, normal, etc.). Such classes of distributions are frequently encountered in machine learning (e. g., mixture models, see [Bis06]) and nonparametric density estimation (e. g., histograms, kernels, see [DL01]). The number of distributions in a mixture gives a natural measure of complexity which we use to express the running time of our algorithms.

For some classes of distributions exact algorithms are possible, for example, if each distribution in $\mathcal{F}$ is a piecewise linear function consisting of n pieces then we can compute the distances between all pairs in time $\Theta(m^2 n)$. For other classes of distributions (for example, mixtures of normal distributions) exact computation of the distances might not be possible. Thus we turn to randomized approximation algorithms. A (δ, ε)-*relative-error approximation scheme* computes $D_{jk},\ j, k \in [m]$ such that with probability at least $1 - \delta$ we have

$$(\forall j, k \in [m]) \quad (1 - \varepsilon) D_{jk} \leq \|f_j - f_k\|_1 \leq (1 + \varepsilon) D_{jk}.$$

A (δ, ε)-*absolute-error approximation scheme* computes $D_{jk},\ j,k \in [m]$ such that with probability at least $1-\delta$ we have

$$(\forall j,k \in [m])\ \ D_{jk} - \varepsilon \leq \|f_j - f_k\|_1 \leq D_{jk} + \varepsilon.$$

A direct application of the Monte Carlo method ([MU49], see [Met87]) immediately yields the following absolute-error approximation scheme. Let X_{jk} be sampled according to f_j and let $Y_{jk} = \mathrm{sgn}(f_j(X_{jk}) - f_k(X_{jk}))$, where $\mathrm{sgn} : \mathbb{R} \to \{-1, 0, 1\}$ is the sign function. The expected value of $Y_{jk} + Y_{kj}$ is equal to $\|f_j - f_k\|_1$, indeed

$$E[Y_{jk} + Y_{kj}] = \int (f_j(x) - f_k(x))\,\mathrm{sgn}(f_j(x) - f_k(x))\,\mathrm{d}x = \|f_j - f_k\|_1.$$

Thus, to obtain a (δ, ε)-absolute-error approximation scheme it is enough to approximate each Y_{jk} with absolute error $\varepsilon/2$ and confidence $1 - \delta/m^2$. By the Chernoff bound $O(\varepsilon^{-2}\ln(m^2/\delta))$ samples from each Y_{jk} are enough. (The total number of samples from the f_j is $O(m\varepsilon^{-2}\ln(m^2/\delta)$, since we can use the same sample from f_j for $Y_{j1}, \dots, Y_{jm}$. The total number of evaluations is $O(m^2\varepsilon^{-2}\ln(m^2/\delta)$.) The running time of this algorithm will compare favorably with the exact algorithm if *sampling* from the densities and *evaluation* of the densities at a point can be done fast. (For example, for piecewise linear densities both sampling and evaluation can be done in $O(\log n)$ time, using binary search.) Note that the evaluation oracle is essential (cf. [BFR$^+$00] who only allow use of sampling oracles).

In the rest of the paper we will focus on the harder relative-error approximation schemes (since the L_1-distance between two distributions is at most 2, a relative-approximation scheme immediately yields an absolute-error approximation scheme). Our motivation comes from an application (density estimation) which requires a relative-error scheme [MŠ07].

Now we outline the rest of the paper. In Section 2 we review Cauchy random projections; in Section 3 we point out that for density functions Cauchy random projections become stochastic integrals; in Section 4 we show that for piecewise linear functions we can sample from these integrals (using rejection sampling, with bivariate student distribution as the envelope) and as a consequence we obtain efficient approximation algorithm for relative-error all-pairs-L_1-distances. Finally, in Section 5, we show that for piecewise polynomial functions one can approximately sample from the integrals, leading to slightly less efficient approximation algorithms.

2 Cauchy random projections

Dimension reduction (the most well-known example is the Johnson-Lindenstrauss lemma for L_2-spaces [JL84]) is a natural technique to use here. We are interested in L_1-spaces for which the analogue of the Johnson-Lindenstrauss lemma is not possible [BC05, NL04] (that is, one cannot project points into a low dimensional L_1-space and preserve distances with a small relative error). However one can still project points to short vectors from which L_1-distances between the original points can be approximately recovered using *non-linear* estimators [LHC07, Ind06].

A particularly fruitful view of the dimensionality "reduction" (with non-linear estimators) is through stable distributions ([JS82, Ind06]): given vectors $v_1, \dots, v_m$ one defines (dependent) random variables $X_1, \dots, X_m$ such that the distance of v_j and v_k can be recovered from $X_j - X_k$ (for all $j, k \in [m]$). For example, in the case of L_1-distances $X_j - X_k$ will be from Cauchy distribution $C(0, \|v_j - v_k\|_1)$, and hence the recovery problem is to estimate the scale parameter R of Cauchy distribution $C(0, R)$. This is a well-studied problem (see, e. g., [HBA70]). We can, for example, use the following nonlinear estimator (other estimators, e. g., the median are also possible [Ind06]):

Lemma 2.1 (Lemma 7 of [LHC07])**.** *Let $X_1, X_2, \dots, X_t$ be independent samples from the Cauchy distribution $C(0, D)$. Define the geometric mean estimator without bias-correction $\hat{D}_{gm}$ as*

$$\hat{D}_{gm} = \prod_{j=1}^{t} |X_j|^{1/t}.$$

Then for each $\varepsilon \in [0, 1/2]$, we have

$$P\left(\hat{D}_{gm} \in [(1-\varepsilon)D, (1+\varepsilon)D]\right) \geq 1 - 2\exp(-t\varepsilon^2/8).$$

We first illustrate how Cauchy random projections immediately give an efficient relative-error approximation scheme for piecewise uniform distributions.

Let $\mathcal{F}$ consist of m piecewise uniform densities, that is, each member of $\mathcal{F}$ is a mixture of n distributions each uniform on an interval. Let $a_1, \dots, a_s$ be the endpoints of all the intervals that occur in $\mathcal{F}$ sorted in the increasing order (note that $s \leq 2mn$). Without loss of generality, we can assume that each distribution $f_j \in \mathcal{F}$ is specified by n pairs $(b_{j1}, c_{j1}), \dots, (b_{jn}, c_{jn})$ where $1 \leq b_{j1} < c_{j1} < \cdots < b_{jn} < c_{jn} \leq s$, and for each pair $(b_{j\ell}, c_{j\ell})$ we are also given a number $\alpha_{j\ell}$ which is the value of f_j on the interval $[a_{b_{j\ell}}, a_{c_{j\ell}})$.

Now we will use Cauchy random projections to compute the pairwise L_1-distances between the f_j efficiently. For $\ell \in \{1, \dots, s-1\}$ let Z_ℓ be independent from the Cauchy distribution $C(0, a_{\ell+1} - a_\ell)$. Let $Y_\ell = Z_1 + \cdots + Z_{\ell-1}$, for $\ell = 1, \dots, s$. Finally, let

$$X_j := \sum_{\ell=1}^{n} \alpha_{j\ell}(Y_{c_{j\ell}} - Y_{b_{j\ell}}) = \sum_{\ell=1}^{n} \alpha_{j\ell}(Z_{b_{j\ell}} + \cdots + Z_{c_{j\ell}-1}). \tag{1}$$

Note that X_j is a sum of Cauchy random variables and hence has Cauchy distribution (in fact it is from $C(0, 1)$). Thus $X_j - X_k$ will be from Cauchy distribution as well. The coefficient of Z_ℓ in $X_j - X_k$ is the difference of f_j and f_k on interval $[a_\ell, a_{\ell+1})$. Hence the contribution of Z_ℓ to $X_j - X_k$ is from Cauchy distribution $C(0, \int_{a_\ell}^{a_{\ell+1}} |f_j(x) - f_k(x)|\, \mathrm{d}x)$, and thus $X_j - X_k$ is from Cauchy distribution $C(0, \|f_j - f_k\|_1)$.

Remark 2.2. In the next section we will generalize the above approach to piecewise degree-d-polynomial densities. In this case for each $(b_{j\ell}, c_{j\ell})$ we are given a vector $\alpha_{j\ell} \in \mathbb{R}^{d+1}$ such that the value of f_j on interval $[a_{b_{j\ell}}, a_{c_{j\ell}})$ is given by the following polynomial (written as an inner product):

$$f_j(x) = (1, x, \dots, x^d) \cdot \alpha_{j\ell}.$$

3 Cauchy motion

A natural way of generalizing the algorithm from the previous section to arbitrary density functions is to take infinitesimal intervals. This leads one to the well-studied area of stochastic integrals w.r.t. symmetric 1-stable Lévy motion (also called Cauchy motion). Cauchy motion is a stochastic process $\{X(t), t \in \mathbb{R}\}$ such that $X(0) = 0$, X has independent increments (i. e., for any $t_1 \le t_2 \le \cdots \le t_k$ the random variables $X(t_2) - X(t_1), \dots, X(t_k) - X(t_{k-1})$ are independent), and $X(t) - X(s)$ is from Cauchy distribution $C(0, |t - s|)$. Intuitively, stochastic integral of a *deterministic function* w.r.t. Cauchy motion is like a regular integral, except one uses $X(t) - X(s)$ instead of $t - s$ for the length of an interval (see section 3.4 of [ST94] for a readable formal treatment).

We will only need the following basic facts about stochastic integrals of deterministic functions w.r.t. Cauchy motion (which we will denote $\mathrm{d}\mathcal{L}(x)$), see [ST94], Chapter 3.

Fact 3.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a (Riemann) integrable function. Let $X = \int_a^b f(x)\,\mathrm{d}\mathcal{L}(x)$. Then X is a random variable from Cauchy distribution $C(0, R)$ where*

$$R = \int_a^b |f(x)|\,\mathrm{d}x. \tag{2}$$

Fact 3.2. *Let $f_1, \dots, f_d : \mathbb{R} \to \mathbb{R}$ be (Riemann) integrable functions. Let $\phi = (f_1, \dots, f_d) : \mathbb{R} \to \mathbb{R}^d$. Let $(X_1, \dots, X_d) = \int_a^b \phi(x)\,\mathrm{d}\mathcal{L}(x)$. Then $(X_1, \dots, X_d)$ is a random variable with characteristic function*

$$\hat{f}(c_1, \dots, c_d) = \exp\left(-\int_a^b |c_1 f_1(x) + \cdots + c_d f_d(x)|\,\mathrm{d}x\right).$$

Fact 3.3. *Let $f, g : \mathbb{R} \to \mathbb{R}$ be (Riemann) integrable functions. Let $a < b, \alpha, \beta \in R$. Then*

$$\int_a^b (\alpha f + \beta g)\,\mathrm{d}\mathcal{L}(x) = \alpha \int_a^b f\,\mathrm{d}\mathcal{L}(x) + \beta \int_a^b g\,\mathrm{d}\mathcal{L}(x).$$

Let $h(x) = f(a + (b - a)x)$. Then

$$\int_a^b f(x)\,\mathrm{d}\mathcal{L}(x) = (b - a) \int_0^1 h(x)\,\mathrm{d}\mathcal{L}(x).$$

From facts 3.1 and 3.3 it follows that the problem of approximating the L_1-distances between densities can be solved if we can evaluate stochastic integrals w.r.t. Cauchy motion; we formalize this in the following observation.

Observation 3.4. *Let $f_1, \dots, f_m : \mathbb{R} \to \mathbb{R}$ be probability densities. Let $\phi : \mathbb{R} \to \mathbb{R}^m$ be defined by $\phi(x) = (f_1(x), \dots, f_m(x))$. Consider*

$$(X_1, \dots, X_m) = \int_{-\infty}^{\infty} \phi(x)\,\mathrm{d}\mathcal{L}(x). \tag{3}$$

For all $j, k \in [m]$ we have that $X_j - X_k$ is from Cauchy distribution $C(0, \|f_j - f_k\|_1)$.

Note that the X_j defined by (1) are in fact computing the integral in (3). For piecewise uniform densities it was enough to sample from the Cauchy distribution to compute the integral. For piecewise degree-d-polynomial densities it will be enough to sample from the following distribution.

Definition 3.5. Let $\phi : \mathbb{R} \to \mathbb{R}^{d+1}$ be defined by $\phi(x) = (1, x, x^2, \ldots, x^d)$. Let $\mathrm{CI}_d(a, b)$ be the distribution of Z, where

$$Z := (Z_0, \ldots, Z_d) := \int_a^b \phi(x)\, \mathrm{d}\mathcal{L}(x).$$

Note that given a sample from $\mathrm{CI}_d(0, 1)$, using $O(d^2)$ arithmetic operations we can obtain a sample from $\mathrm{CI}_d(a, b)$, using Fact 3.3.

Lemma 3.6. *Let $\mathcal{F}$ consist of m piecewise degree-d-polynomial densities, each consisting of n pieces (given as in Remark 2.2). Let $t \geq (8/\varepsilon)^2 \ln(m^2/\delta)$ be an integer. Assume that we can sample from $\mathrm{CI}_d(0, 1)$ using T_d operations. We can obtain (δ, ε)-relative-error approximation of L_1-distances between all pairs in $\mathcal{F}$, using $O((d^2 + T_d)mnt + m^2t)$ arithmetic operations.*

Proof :
For $\ell \in \{1, \ldots, s-1\}$ let Z_ℓ be independent from $\mathrm{CI}_d(a_\ell, a_{\ell+1})$ distribution. Let $Y_\ell = Z_1 + \cdots + Z_{\ell-1}$, for $\ell = 1, \ldots, s$. Finally, for each $j \in [m]$, let

$$X_j := \sum_{\ell=1}^{n} \alpha_{j\ell} \cdot (Y_{c_{j\ell}} - Y_{b_{j\ell}}) = \sum_{\ell=1}^{n} \alpha_{j\ell} \cdot (Z_{b_{j\ell}} + \cdots + Z_{c_{j\ell}-1}). \tag{4}$$

Note that $Y_{c_{j\ell}} - Y_{b_{j\ell}}$ is from $C(a_{b_{j\ell}}, a_{c_{j\ell}})$ and hence

$$\alpha_{j\ell} \cdot (Y_{c_{j\ell}} - Y_{b_{j\ell}}) = \int_{a_{b_{j\ell}}}^{a_{c_{j\ell}}} f_j(x)\, \mathrm{d}\mathcal{L}(x).$$

Thus $(X_1, \ldots, X_m)$ defined by (4) compute (3).

For every $j, k \in [m]$ we have that $X_j - X_k$ is from Cauchy distribution

$$C(0, \|f_j - f_k\|_1).$$

If we have t samples from each $X_1, \ldots, X_m$ then using Lemma 2.1 and union bound with probability $\geq 1 - \delta$ we recover all $\|f_j - f_k\|_1$ with relative error ε.

Note that $s \leq 2mn$ and hence for the Z_ℓ we used $\leq 2mnt$ samples from $\mathrm{CI}(0, 1)$ distribution, costing us $O((d^2 + T_d)mnt)$ arithmetic operation. Computing the Y_ℓ takes $O(mnt)$ operations. Computing the X_j takes $O(mnt)$ operations. The final estimation of the distances takes $O(m^2t)$ operations. ■

4 Piecewise linear functions

The density function of $\mathrm{CI}_1(0, 1)$ can be computed explicitly, using the inverse Fourier transform; the proof is deferred to the appendix. The expression for the density allows us to construct efficient sampling algorithm, which in turn yields an efficient approximation algorithm for all-pairs-L_1-distances for piecewise linear densities. We obtain the following result.

Theorem 4.1. *Let $\mathcal{F}$ consist of m piecewise linear densities, each consisting of n pieces (given as in Remark 2.2). We can obtain (δ, ε)-relative-error approximation of L_1-distances between all pairs in $\mathcal{F}$, using $O(m(m+n)\varepsilon^{-2}\ln(m/\delta))$ arithmetic operations.*

Now we state the density of $\mathrm{CI}_1(0,1)$. In the following $\Re(x)$ denotes the real part of a complex number x.

Theorem 4.2. *Let $\phi : \mathbb{R} \to \mathbb{R}^2$ be the function $\phi(x) = (1, x)$. Let*

$$Z = (X_1, X_2) = \int_0^1 \phi(z)\,\mathrm{d}\mathcal{L}(z).$$

For $x_1 \neq 2x_2$ the density function of Z is given by

$$f(x_1, x_2) = \frac{4/\pi^2}{1 + 6x_1^2 + x_1^4 - 16x_1x_2 + 16x_2^2} + \frac{2}{\pi^2}\Re\left(\frac{\operatorname{atan}(iQ/(x_1 - 2x_2))}{Q^{3/2}}\right), \tag{5}$$

where

$$Q = 1 - 2ix_1 + x_1^2 + 4ix_2. \tag{6}$$

For $x_1 = 2x_2$ the density is given by

$$f(x_1, x_2) = \frac{4/\pi^2}{(1 + x_1^2)^2} + \frac{1}{\pi(1 + x_1^2)^{3/2}}. \tag{7}$$

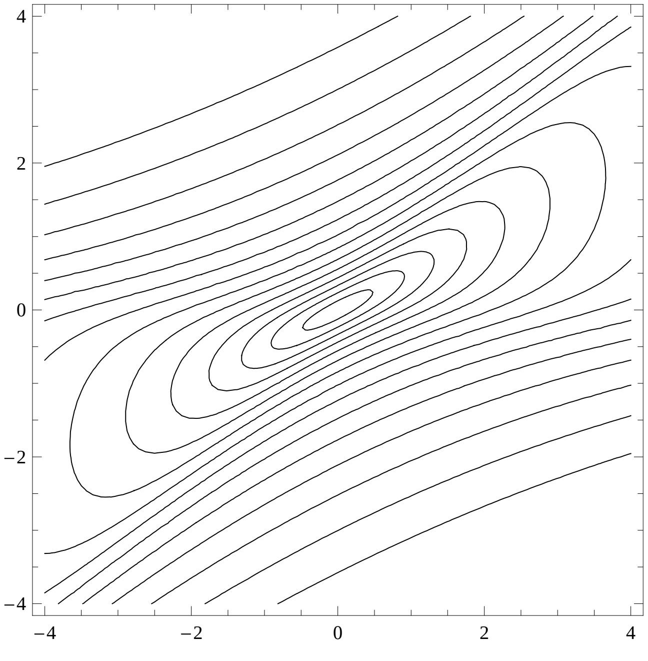

Figure 1: The density plot of $(X_1, X_2) = \int_0^1 (1, z)\,\mathrm{d}\mathcal{L}(z)$. The contours are at levels $2^{-15}, 2^{-14}, \ldots, 2^{-1}$.

Next we show how to efficiently sample from the $\mathrm{CI}_1(0,1)$ distribution by rejection sampling using the bivariate student distribution as the envelope.

Let Σ be a positive-definite 2×2 matrix. The bivariate student distribution with 1 degree of freedom is given by the following formula (see, e. g., [ES00], p. 50)

$$g(\mathbf{x}) = \frac{|\det(\Sigma)|^{-1/2}}{2\pi} \left(1 + \frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2}\right)^{-3/2}.$$

It is well-known how to sample X from this distribution: let $X = \Sigma^{1/2} Y / \sqrt{W}$, where Y, W are independent with $Y \sim N_2(0, I)$ (the two dimensional gaussian) and $W \sim \chi^2(1)$ (chi-squared distribution with 1 degree of freedom).

We are going to use the bivariate student distribution with the following density

$$g(\mathbf{x}) = \frac{1}{\pi} \left(1 + x_1^2 + (2x_2 - x_1)^2\right)^{-3/2}. \tag{8}$$

We show that the density function of the $\mathrm{CI}_1(0,1)$ distribution is bounded by a constant multiple of (8) (the proof is deferred to the appendix).

Lemma 4.3. *Let $f(\mathbf{x})$ be given by* (5) *and* (7)*. Let $g(\mathbf{x})$ be given by* (8)*. For every $\mathbf{x} \in \mathbb{R}^2$ we have*

$$f(\mathbf{x}) \le \frac{C}{\pi} \cdot g(\mathbf{x}),$$

where $C = 25$.

As an immediate corollary of Lemma 4.3 we obtain an efficient sampling algorithm for $\mathrm{CI}_1(0,1)$ distribution, using rejection sampling (see, e. g., [ES00]).

Corollary 4.4. *There is a sampler from $\mathrm{CI}_1(0,1)$ which uses a constant number of samples from from $N(0,1)$ and $\chi^2(1)$ (in expectation).*

Proof of Theorem 4.1:
The theorem follows from Corollary 4.4 and Lemma 3.6. ■

Remark 4.5. Lemma 4.3 is true with $C = \pi 2^{3/2}$ (we skip the technical proof). The constant $\pi 2^{3/2}$ is tight (see equation (40) with $\alpha \to 0$ and $T \to 1$).

5 Piecewise polynomial functions

Some kernels used in machine learning (e. g., the Epanechnikov kernel, see [DL01], p.85) are piecewise polynomial. Thus it is of interest to extend the result from the previous section to higher-degree polynomials.

For $d > 1$ we do not know how to sample from distribution $\mathrm{CI}_d(0,1)$ exactly. However we can still approximately sample from this distribution, as follows. Let r be an integer. Let $Z_1, \ldots, Z_r$ be independent from Cauchy distribution $C(0, 1/r)$. Consider the following distribution, which we call r-approximation of $\mathrm{CI}_d(0,1)$:

$$(X_0, \ldots, X_d) = \sum_{j=1}^{r} Z_j \cdot (1, (j/r), (j/r)^2, \ldots, (j/r)^d). \tag{9}$$

Now we show that if r is large enough then the distribution given by (9) can be used instead of distribution $\mathrm{CI}_d(0,1)$ for our purpose. As a consequence we will obtain the following.

Theorem 5.1. *Let $\mathcal{F}$ consist of m piecewise degree-d-polynomial densities, each consisting of n pieces (given as in Remark 2.2). We can obtain (δ, ε)-relative-error approximation of L_1-distances between all pairs in $\mathcal{F}$, using $O(m(m+n)d^3\varepsilon^{-3}\ln(m/\delta))$ arithmetic operations.*

Remark 5.2. Note that for $d = 1$ Theorem 5.1 gives worse (in ε) running time that Theorem 4.1. This slowdown is caused by the additional integration used to simulate $\mathrm{CI}_d(0, 1)$.

The proof of Theorem 5.1 will be based on the following result which shows that (9) is in some sense close to $\mathrm{CI}_d(0, 1)$.

Lemma 5.3. *Let $p = a_0 + a_1x + \cdots + a_dx^d$ be a polynomial of degree d. Let $(X_0, \ldots, X_d)$ be sampled from the distribution given by (9), with $r \geq cd^2/\varepsilon$ (where c is an absolute constant). Let $W = a_0X_0 + \cdots + a_dX_d$. Then W is from the Cauchy distribution $C(0, R)$, where*

$$(1-\varepsilon)\int_0^1 |p(x)|\,\mathrm{d}x \leq R \leq (1+\varepsilon)\int_0^1 |p(x)|\,\mathrm{d}x. \tag{10}$$

We defer the proof of Lemma 5.3 to the end of this section. Note that having (10) instead of (2) (which sampling from $\mathrm{CI}_d(0, 1)$ would yield) will introduce small relative error to the approximation of the L_1-distances.

Proof of Theorem 5.1:
The proof is analogous to the proof of Lemma 3.6. Let $r \geq cd^2/\varepsilon$. For $\ell \in \{1, \ldots, s-1\}$ let Z_ℓ be independent from r-approximation of $\mathrm{CI}_d(a_\ell, a_{\ell+1})$ distribution. Let $Y_\ell = Z_1 + \cdots + Z_{\ell-1}$, for $\ell = 1, \ldots, s$. Finally, for each $j \in [m]$, let

$$X_j := \sum_{\ell=1}^{n} \alpha_{j\ell} \cdot (Y_{c_{j\ell}} - Y_{b_{j\ell}}) = \sum_{\ell=1}^{n} \alpha_{j\ell} \cdot (Z_{b_{j\ell}} + \cdots + Z_{c_{j\ell}-1}).$$

By Lemma 5.3, for every $j, k \in [m]$ we have that $X_j - X_k$ is from Cauchy distribution $C(0, R)$ where $(1-\varepsilon)\|f_j - f_k\|_1 \leq R \leq (1+\varepsilon)\|f_j - f_k\|_1$.

If we have $t \geq (8/\varepsilon)^2\ln(m^2/\delta)$ samples from each $X_1, \ldots, X_m$ then using Lemma 2.1 and union bound with probability $\geq 1 - \delta$ we recover all $\|f_j - f_k\|_1$ with relative error $\approx 2\varepsilon$.

Note that $s \leq 2mn$ and hence for the Z_ℓ we used $\leq 2mnt$ samples from r-approximation of $\mathrm{CI}(0, 1)$ distribution, costing us $O((d^3/\varepsilon)mnt)$ arithmetic operation. Computing the Y_ℓ takes $O(mnt)$ operations. Computing the X_j takes $O(mnt)$ operations. The final estimation of the distances takes $O(m^2t)$ operations. ■

To prove Lemma 5.3 we will use the following Bernstein-type inequality from [Erd00].

Theorem 5.4. *(Theorem 3.1 of [Erd00]) There exists a constant $c > 0$ such that for any degree d polynomial p,*

$$\int_0^1 |p'(x)|dx \leq cd^2\int_0^1 |p(x)|dx.$$

We have the following corollary of Theorem 5.4.

Lemma 5.5. *There exists a constant c such that for any polynomial p of degree d, any $r \geq cd^2$, any $0 = x_0 < x_1 < x_2, \ldots < x_t = 1$ with $\max_j |x_j - x_{j-1}| \leq 1/r$, and any $\theta_1 \in [x_0, x_1], \theta_2 \in [x_1, x_2], \ldots, \theta_t \in [x_{t-1}, x_t]$, we have*

$$(1 - cd^2/r) \int_0^1 |p(x)| dx \leq \sum_{j=1}^{t} (x_j - x_{j-1}) |p(\theta_j)| \leq (1 + cd^2/r) \int_0^1 |p(x)| dx. \tag{11}$$

Proof :
We will use induction on the degree d of the polynomial. For $d = 0$ the sum and the integrals in (11) are equal.

Now assume $d \geq 1$. For each $j \in [t]$, we use the Taylor expansion of $p(x)$ about θ_j for $x \in (x_{j-1}, x_j]$. This yields for each $x \in (x_{j-1}, x_j]$, $p(x) = p(\theta_j) + (x - \theta_j)p'(\theta'_{j,x})$, where $\theta'_{j,x} \in (x_{j-1}, x_j]$. Let β_j be the point $y \in (x_{j-1}, x_j]$ that maximizes $p'(y)$. We have

$$\begin{aligned} \left| \sum_{j=1}^{t} (x_j - x_{j-1}) |p(\theta_j)| - \int_0^1 |p(x)| \, \mathrm{d}x \right| &\leq \sum_{j=1}^{t} \int_{x_{j-1}}^{x_j} |p(x) - p(\theta_j)| \, \mathrm{d}x \\ &\leq \sum_{j=1}^{t} \int_{x_{j-1}}^{x_j} |(x - \theta_j) p'(\theta'_{j,x})| \, \mathrm{d}x \leq \frac{1}{2r} \sum_{j=1}^{t} (x_j - x_{j-1}) |p'(\beta_j)|. \end{aligned} \tag{12}$$

Since p' is of degree $d - 1$, by induction hypothesis the right-hand side of (12) is bounded as follows

$$\begin{aligned} \frac{1}{2r} \sum_{j=1}^{t} (x_j - x_{j-1}) |p'(\beta_j)| &\leq \frac{1}{2r} (1 + c(d-1)^2 \varepsilon) \int_0^1 |p'(x)| \mathrm{d}x \\ &\leq (1/r) \int_0^1 |p'(x)| \mathrm{d}x \leq (cd^2/r) \int_0^1 |p(x)| \mathrm{d}x. \end{aligned}$$

where in the last inequality we used Theorem 5.4. Hence the lemma follows. ■

Proof of Lemma 5.3:
We have

$$W = (a_0, \ldots, a_d) \cdot \sum_{j=1}^{r} Z_j (1, (j/r), (j/r)^2, \ldots, (j/r)^d) = \sum_{j=1}^{r} Z_j \, p(j/r),$$

where Z_j are from Cauchy distribution $C(0, 1/r)$. Thus W is from Cauchy distribution $C(0, R)$, where

$$R = \frac{1}{r} \sum_{j=1}^{r} |p(j/r)|.$$

Using Lemma 5.5 we obtain (10). ■

Remark 5.6. An alternate view of Lemma 5.5 is that a piecewise degree-d-polynomial density with n pieces can be approximated by a piecewise uniform density with $O(nd^2/\varepsilon)$ pieces. The approximation distorts L_1-distances between any pair of such densities by a factor at most $1 \pm \varepsilon$. To obtain a relative-approximation of the L_1-distances in a family $\mathcal{F}$ one can now directly use the algorithm from Section 2 without going through the stochastic integrals approach (for $d = 1$ the price for this method is a $1/\varepsilon$ factor slowdown).

Remark 5.7. (on L_2-distances) For L_2-distances the dimension reduction uses normal distribution instead of Cauchy distribution. For infinitesimal intervals the corresponding process is Brownian motion, which is much better understood than Cauchy motion. Evaluation of a stochastic integral of a deterministic function $\mathbb{R} \to \mathbb{R}^d$ w.r.t. Brownian motion is a d-dimensional gaussian (whose covariance matrix is easy to obtain), for example

$$\int_0^1 (1, x, \ldots, x^d)\, \mathrm{d}\mathcal{L}_{\mathrm{Brown}}(x)$$

is from $N(0, \Sigma)$ where Σ is the $(d+1) \times (d+1)$ Hilbert matrix (that is, the ij-th entry of Σ is $1/(i+j-1)$).

Question 5.8. *How efficiently can one sample from $\mathrm{CI}_d(0,1)$ distribution? A reasonable guess seems to be that one can sample from a distribution within L_1-distance δ from $\mathrm{CI}_d(0,1)$ using $d^2 \ln(1/\delta)$ samples.*

Acknowledgement

The authors would like to thank Carl Mueller for advice on stochastic integrals.

References

[BC05] Bo Brinkman and Moses Charikar. On the impossibility of dimension reduction in l_1. *Journal of the ACM*, 52(5):766–788, 2005.

[BFR$^+$00] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *FOCS*, pages 259–269, 2000.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer-Verlag, New York, 2006.

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation.* Springer Series in Statistics. Springer-Verlag, New York, 2001.

[Erd00] Tamás Erdélyi. Markov- and Bernstein-type inequalities for Müntz polynomials and exponential sums in l_p. *J. Approx. Theory*, 104(1):142–152, 2000.

[ES00] Michael Evans and Tim Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford University Press, 2000.

[GR07] Israil S. Gradshteyn and Iosif M. Ryzhik. *Table of Integrals, Series, and Products, 7th edition.* Academic Press, New York, 2007.

[HBA70] Gerald Haas, Lee Bain, and Charles Antle. Inferences for the cauchy distribution based on maximum likelihood estimators. *Biometrika*, 57:403–408, 1970.

[Ind06] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM*, 53(3):307–323, 2006.

[JL84] William B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[JS82] William B. Johnson and Gideon Schechtman. Embedding l_p^m into l_1^n. *Acta Math.*, 149(1-2):71–85, 1982.

[LHC07] Ping Li, Trevor J. Hastie, and Kenneth W. Church. Nonlinear estimators and tail bounds for dimension reduction. *Journal of Machine Learning Research*, 8:2497–2532, 2007.

[Met87] Nicholas Metropolis. The beginning of the Monte Carlo method. *Los Alamos Sci.*, (15, Special Issue):125–130, 1987. Stanislaw Ulam 1909–1984.

[MŠ07] Satyaki Mahalanabis and Daniel Štefankovič. Density estimation in linear time. *arXiv.org,* `http://arxiv.org/abs/0712.2869`, December 2007.

[MU49] Nicholas Metropolis and S. Ulam. The Monte Carlo method. *J. Amer. Statist. Assoc.*, 44:335–341, 1949.

[NL04] Assaf Naor and James R. Lee. Embedding the diamond graph in l_p and dimension reduction in l_1. *Geometric and Functional Analysis*, 14(4):745–747, 2004.

[ST94] Gennady Samorodnitsky and Murad S. Taqqu. *Stable non-Gaussian random processes : stochastic models with infinite variance*. Stochastic modeling. Chapman & Hall, New York, 1994.

6 Appendix

6.1 Stochastic integral of (constant, linear) function

In this section we give an explicit formula for the density function of the random variable

$$(X,Y) = \int_0^1 \phi(z)\,\mathrm{d}\mathcal{L}(z),$$

where $\phi(z) = (1, z)$, and $\mathrm{d}\mathcal{L}(z)$ is the Cauchy motion.

We will obtain the density function from the characteristic function. The following result will be used in the inverse Fourier transform. (We use $\Re$ to denote the real part of a complex number.)

Lemma 6.1. *Let $\phi = (\phi_1, \dots, \phi_n) : \mathbb{R} \to \mathbb{R}^n$. Let*

$$Z = (X_1, \dots, X_n) = \int_0^1 \phi(x)\ \mathrm{d}\mathcal{L}(x),$$

where $\mathcal{L}$ is the Cauchy motion. The density function f of Z is given by

$$\Re\left(\frac{(n-1)!}{(2\pi)^n}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\frac{2}{(A+iB)^n}\mathrm{d}b_1\dots\mathrm{d}b_{n-1}\right), \tag{13}$$

where

$$A = A(b_1, \dots, b_{n-1}) := \int_0^1 \left|b_1\phi_1(x) + \dots + b_{n-1}\phi_{n-1}(x) + \phi_n(x)\right|, \tag{14}$$

and

$$B = B(b_1, \dots, b_{n-1}, x_1, \dots, x_n) := b_1x_1 + \dots + b_{n-1}x_{n-1} + x_n. \tag{15}$$

Proof :
The characteristic function of Z is (see, e. g., proposition 3.2.2 of [ST94]):

$$\hat{f}(a_1,\ldots,a_n) = E[\exp(i(a_1X_1+\cdots+a_nX_n))] = \exp\left(-\int_0^1 |a_1\phi_1(x)+\cdots+a_n\phi_n(x)|\right).$$

We will use the following integral, valid for any $A>0$ (see, e. g., [GR07]):

$$\int_0^\infty t^{n-1}\exp(-At)\cos(Bt)\,\mathrm{d}t = \frac{(n-1)!}{2}\left(\frac{1}{(A-iB)^n}+\frac{1}{(A+iB)^n}\right). \tag{16}$$

We would like to compute the inverse Fourier transform of $\hat{f}$, which, since $\hat{f}$ is symmetric about the origin, is given by

$$f(x_1,\ldots,x_n) = \frac{2}{(2\pi)^n}\int_0^\infty\int_{-\infty}^\infty\cdots\int_{-\infty}^\infty \hat{f}(a_1,\ldots,a_n)\cos(a_1x_1+\cdots+a_nx_n)\mathrm{d}a_1\ldots\mathrm{d}a_{n-1}\mathrm{d}a_n. \tag{17}$$

Substitution $a_n = t, a_{n-1} = b_{n-1}t, \ldots, a_1 = b_1t$ into (17) yields

$$\begin{aligned}&f(x_1,\ldots,x_n) = \\ &\frac{2}{(2\pi)^n}\int_{-\infty}^\infty\cdots\int_{-\infty}^\infty\left(\int_0^\infty t^{n-1}\exp\left(-t\int_0^1|b_1\phi_1(x)+\cdots+b_{n-1}\phi_{n-1}(x)+\phi_n(x)|\right)\right.\\ &\qquad\left.\cos\left(t(b_1x_1+\cdots+b_{n-1}x_{n-1}+x_n)\right)\mathrm{d}t\right)\mathrm{d}b_1\ldots\mathrm{d}b_{n-1}.\end{aligned}$$

Note that the inner integral has the same form as (16) and hence we have

$$\begin{aligned}f(x_1,\ldots,x_n) &= \frac{(n-1)!}{(2\pi)^n}\int_{-\infty}^\infty\cdots\int_{-\infty}^\infty\frac{1}{(A-iB)^n}+\frac{1}{(A+iB)^n}\mathrm{d}b_1\ldots\mathrm{d}b_{n-1}\\ &= \Re\left(\frac{(n-1)!}{(2\pi)^n}\int_{-\infty}^\infty\cdots\int_{-\infty}^\infty\frac{2}{(A+iB)^n}\mathrm{d}b_1\ldots\mathrm{d}b_{n-1}\right),\end{aligned} \tag{18}$$

where A and B are given by (14) and (15). The last equality in (18) follows from the fact that the two summands in the integral are conjugate complex numbers. ■

Now we apply Lemma 6.1 for the case of two functions, one constant and one linear.

Proof of Theorem 4.2:
Plugging $n=2$, $\phi_1(x)=1$, and $\phi_2(x)=x$ into (14) and (15) we obtain

$$B(b_1,x_1,x_2) = b_1x_1+x_2 \tag{19}$$

and

$$A(b_1) = \begin{cases} b_1+1/2 & \text{if } b_1\geq 0,\\ -b_1-1/2 & \text{if } b_1\leq -1,\\ b_1^2+b_1+1/2 & \text{otherwise.}\end{cases} \tag{20}$$

Our goal now is to evaluate the integral (13). We split the integral into 3 parts according to the behavior of $A(b_1)$.

We will use the following integral

$$\int \frac{1}{(Sz+T)^2}\,\mathrm{d}z = -\frac{1}{S(T+Sx)}. \tag{21}$$

For $B = b_1x_1 + x_2$ and $A = b_1 + 1/2$ we have $A + iB = b_1(1 + ix_1) + (1/2 + ix_2)$. Using (21) for A and B given by (19) and (20)) we obtain

$$\int_0^\infty \frac{1}{(A+iB)^2}\mathrm{d}b_1 = \frac{2}{(ix_1+1)(2ix_2+1)}, \tag{22}$$

and

$$\int_{-\infty}^{-1} \frac{1}{(A-iB)^2}\mathrm{d}b_1 = \frac{2}{(ix_1-1)(2i(x_1-x_2)-1)}. \tag{23}$$

We have (see, e. g., [GR07]))

$$\int \frac{1}{(z^2+Sz+T)^2}\,\mathrm{d}z = \frac{S+2z}{(4T-S^2)(T+Sz+z^2)} + \frac{4\mathrm{atan}\left((S+2z)/\sqrt{4T-S^2}\right)}{(4T-S^2)^{3/2}}. \tag{24}$$

For $A = b_1^2 + b_1 + 1/2$ and $B = b_1x_1 + x_2$ we have $A + iB = b_1^2 + b_1(1 + ix_1) + (1/2 + x_2)$. Using (24) we obtain

$$\int_{-1}^{0} \frac{1}{(A+iB)^2}\mathrm{d}b_1 = \frac{2(ix_1+1)}{(2ix_2+1)Q} + \frac{2(ix_1-1)}{(2i(x_1-x_2)-1)Q} + 4\frac{\mathrm{atan}\left(\frac{ix_1+1}{\sqrt{Q}}\right) - \mathrm{atan}\left(\frac{ix_1-1}{\sqrt{Q}}\right)}{Q^{3/2}}, \tag{25}$$

where Q is given by (6).

Summing (22), (23), and (25) we obtain

$$\int_{-\infty}^{\infty} \frac{1}{(A+iB)^2}\mathrm{d}b_1 = \frac{8}{Q(1+x_1^2)} + 4\frac{\mathrm{atan}\left(\frac{ix_1+1}{\sqrt{Q}}\right) - \mathrm{atan}\left(\frac{ix_1-1}{\sqrt{Q}}\right)}{Q^{3/2}}. \tag{26}$$

We have

$$\left|\frac{ix_1 \pm 1}{\sqrt{Q}}\right|^4 = \frac{(1+x_1^2)^2}{(1+x_1^2)^2 + (2x_1-4x_2)^2} \le 1.$$

with equality only if $x_1 = 2x_2$. Hence if $x_1 \neq 2x_2$ then using (42) we have

$$\mathrm{atan}\left(\frac{ix_1+1}{\sqrt{Q}}\right) - \mathrm{atan}\left(\frac{ix_1-1}{\sqrt{Q}}\right) = \mathrm{atan}(iQ/(x_1-2x_2)),$$

and by applying

$$\Re\left(\frac{8}{Q(1+x_1^2)}\right) = \frac{8}{1+6x_1^2+x_1^4-16x_1x_2+16x_2^2}$$

in (26) we obtain (5).

If $x_1 = 2x_2$ then $Q = 1 + x_1^2$ and using

$$\mathrm{atan}\left(\frac{ix_1+1}{\sqrt{Q}}\right) - \mathrm{atan}\left(\frac{ix_1-1}{\sqrt{Q}}\right) = \pi/2$$

in (26) we obtain (7). ■

6.2 Bounding the $\mathrm{CI}_1(0,1)$-distribution

Now we prove that the multivariate student distribution gives an efficient envelope for the $\mathrm{CI}_1(0,1)$-distribution.

Proof of Lemma 4.3:
To simplify the formulas we use the following substitutions: $x_1 = u$ and $x_2 = w + u/2$. The density g becomes

$$g'(u,v) := \frac{1}{\pi}\left(1+u^2+4w^2\right)^{-3/2}.$$

For $w = 0$ (which corresponds to $x_1 = 2x_2$) the density f becomes

$$\frac{4/\pi^2}{(1+u^2)^2} + \frac{1}{\pi(1+u^2)^{3/2}}, \tag{27}$$

and hence Lemma 4.3 is true, as

$$(27) \le (4/\pi + 1)\left(\frac{1}{\pi}\left(1+u^2\right)^{-3/2}\right).$$

For $w \neq 0$, density (5) becomes

$$f'(u,v) := \frac{1}{\pi^2}\left(\frac{4}{(1+u^2)^2+(4w)^2} + \frac{\operatorname{atan}(iM/(2w))}{M^3} - \frac{\operatorname{atan}(iM'/(2w))}{M'^3}\right),$$

where $M = (1+u^2-4iw)^{1/2}$ and $M' = (1+u^2+4iw)^{1/2}$. We are going to show

$$\pi^2 f'(u,v) \le C\pi g'(u,v). \tag{28}$$

Note that both sides of (28) are unchanged when we flip the sign of u or the sign of w. Hence we can, without loss of generality, assume $u \ge 0$ and $w > 0$.

There are unique $a > 0$ and $b > 0$ such that $w = ab/2$ and $u = \sqrt{a^2-b^2-1}$ (to see this notice that substituting $b = 2w/a$ into the second equation yields $u^2+1 = a^2 - 4w^2/a^2$, where the right-hand side is a strictly increasing function going from $-\infty$ to ∞). Note that $M = a - ib$ and $M' = a + ib$. Also note that

$$a^2 \ge b^2 + 1. \tag{29}$$

After the substitution equation (28) simplifies as follows

$$\frac{4}{(a^2+b^2)^2} + \frac{1}{(a^2+b^2)^3}\Bigg((a+ib)^3 \operatorname{atan}\left(\frac{1}{a}+\frac{i}{b}\right) \\ +(a-ib)^3 \operatorname{atan}\left(\frac{1}{a}-\frac{i}{b}\right)\Bigg) \le \frac{C}{(a^2-b^2+a^2b^2)^{3/2}}. \tag{30}$$

Now we expand $(a+ib)^3$ and $(a-ib)^3$ and simplify (30) into

$$\frac{4}{(a^2+b^2)^2} + \frac{1}{(a^2+b^2)^3}\Bigg((a^3-3ab^2)\left(\operatorname{atan}\left(\frac{1}{a}+\frac{i}{b}\right) + \operatorname{atan}\left(\frac{1}{a}-\frac{i}{b}\right)\right) \\ -i(b^3-3a^2b)\left(\operatorname{atan}\left(\frac{1}{a}+\frac{i}{b}\right) - \operatorname{atan}\left(\frac{1}{a}-\frac{i}{b}\right)\right)\Bigg) \le \frac{C}{(a^2-b^2+a^2b^2)^{3/2}}. \tag{31}$$

Now we substitute $a = 1/A$ and $b = 1/B$ into (31) and obtain

$$\begin{aligned}\frac{4A^4B^4}{(A^2+B^2)^2} + \frac{A^3B^3}{(A^2+B^2)^3}\Bigg(&(B^3-3A^2B)\big(\operatorname{atan}(A+iB)+\operatorname{atan}(A-iB)\big) \\ &-i(A^3-3AB^2)\big(\operatorname{atan}(A+iB)-\operatorname{atan}(A-iB)\big)\Bigg) \le \frac{C\cdot A^3B^3}{(B^2-A^2+1)^{3/2}}.\end{aligned} \tag{32}$$

Note that $A > 0$ and $B > 0$ and the constraint (29) becomes

$$B^2 \ge A^2(1+B^2). \tag{33}$$

Multiplying both sides of (32) by $(A^2+B^2)^3/(AB)^3$ we obtain

$$\begin{aligned}4AB(A^2+B^2) + (B^3-3A^2B)&\big(\operatorname{atan}(A+iB)+\operatorname{atan}(A-iB)\big) \\ -i(A^3-3AB^2)&\big(\operatorname{atan}(A+iB)-\operatorname{atan}(A-iB)\big) \le \frac{C\cdot(A^2+B^2)^6}{(B^2-A^2+1)^{3/2}}.\end{aligned} \tag{34}$$

Finally, we substitute $A = T\sin\alpha$ and $B = T\cos\alpha$ with $T \ge 0$. Note that the constraint (33) becomes

$$(T\sin\alpha)^2 \le \frac{\cos(2\alpha)}{(\cos\alpha)^2}, \tag{35}$$

and hence α is restricted to $[0, \pi/4)$.

Equation (34) then becomes

$$\begin{aligned}2T^4\sin(2\alpha) + T^3\cos(3\alpha)&\left(\operatorname{atan}(A+iB)+\operatorname{atan}(A-iB)\right) \\ +iT^3\sin(3\alpha)&\left(\operatorname{atan}(A+iB)-\operatorname{atan}(A-iB)\right) \le \frac{C\cdot T^6}{(T^2\cos(2\alpha)+1)^{3/2}}.\end{aligned} \tag{36}$$

We prove (36) by considering three cases.

CASE: $T < 1$. We can use (42) to simplify (36) as follows

$$2T\sin(2\alpha) + \cos(3\alpha)\operatorname{atan}\left(\frac{2T\sin(\alpha)}{1-T^2}\right) - \sin(3\alpha)\operatorname{atanh}\left(\frac{2T\cos(\alpha)}{1+T^2}\right) \le \frac{C\cdot T^3}{(T^2\cos(2\alpha)+1)^{3/2}}. \tag{37}$$

For $z \ge 0$ we have $\operatorname{atanh}(z) \ge z \ge \operatorname{atan}(z)$ and hence to prove (37) it is enough to show

$$\begin{aligned}2T\sin(2\alpha)(1-T^4) + (1+T^2)\cos(3\alpha)&\ (2T\sin(\alpha)) \\ -(1-T^2)\sin(3\alpha)&\ (2T\cos(\alpha)) \le \frac{C\cdot T^3(1-T^4)}{(T^2\cos(2\alpha)+1)^{3/2}},\end{aligned} \tag{38}$$

which is implied by the following inequality which holds for $T \le 8/9$:

$$-2T^2\sin(2\alpha) + 2\sin(4\alpha) \le 2 \le \frac{C\cdot 2465/6561}{2^{3/2}} \le \frac{C\cdot(1-T^4)}{(T^2\cos(2\alpha)+1)^{3/2}}. \tag{39}$$

For $1 > T \geq 8/9$ we directly prove (38)

$$2T\sin(2\alpha) + \cos(3\alpha)\,\mathrm{atan}\left(\frac{2T\sin(\alpha)}{1-T^2}\right) - \sin(3\alpha)\,\mathrm{atanh}\left(\frac{2T\cos(\alpha)}{1+T^2}\right)$$

$$\leq 2 + \pi/2 \leq \frac{C \cdot 512/729}{2^{3/2}} \leq \frac{C \cdot T^3}{(T^2\cos(2\alpha)+1)^{3/2}}.$$

CASE: $T > 1$. We can use (43) and (44) to simplify (36) as follows

$$2T\sin(2\alpha) + \cos(3\alpha)\left(\pi + \mathrm{atan}\left(\frac{2T\sin(\alpha)}{1-T^2}\right)\right) - \sin(3\alpha)\,\mathrm{atanh}\left(\frac{2T\cos(\alpha)}{1+T^2}\right) \leq \frac{C \cdot T^3}{(T^2\cos(2\alpha)+1)^{3/2}}. \tag{40}$$

From (35) we have $T\sin(\alpha) \leq 1$ and hence $2T\sin(2\alpha) \leq 4$. Therefore (40) can be proved as follows.

$$2T\sin(2\alpha) + \cos(3\alpha)\left(\pi + \mathrm{atan}\left(\frac{2T\sin(\alpha)}{1-T^2}\right)\right) - \sin(3\alpha)\,\mathrm{atanh}\left(\frac{2T\cos(\alpha)}{1+T^2}\right)$$

$$\leq 4 + 3\pi/2 \leq \frac{C}{2^{3/2}} \leq \frac{C \cdot T^3}{(T^2\cos(2\alpha)+1)^{3/2}}.$$

CASE: $T = 1$. Equation (36) simplifies as follows

$$2\sin(2\alpha) + (\pi/2)\cos(3\alpha) - \sin(3\alpha)\,\mathrm{atanh}\left(\cos(\alpha)\right) \leq \frac{C}{(\cos(2\alpha)+1)^{3/2}}. \tag{41}$$

The left-hand side is bounded from above by $2 + \pi/2$ which is less than $C/2^{3/2}$ which lower-bounds the right-hand side of (41). ■

6.3 Basic properties of trigonometric functions

In this section we list the basic properties of trigonometric functions that we used. For complex parameters these are multi-valued functions for which we choose the branch in the standard way. The *logarithm* of a complex number $z = (\cos\alpha + i\sin\alpha)\mathrm{e}^t$, where $\alpha \in (-\pi, \pi]$, and $t \in \mathbb{R}$ is $i\alpha + t$. The *inverse tangent* of a complex number $z \in \mathbb{C} \setminus \{\pm i\}$ is the solution of $\tan(x) = z$ with $\Re(x) \in (-\pi/2, \pi/2)$. In terms of the logarithm we have

$$\mathrm{atan}(z) := \frac{1}{2} i \left(\ln(1 - iz) - \ln(1 + iz)\right).$$

The inverse hyperbolic tangent function is defined analogously, for $z \in \mathbb{C} \setminus \{\pm 1\}$ we have

$$\mathrm{atanh}(z) := \frac{1}{2}\left(\ln(1+z) - \ln(1-z)\right) = -i\,\mathrm{atan}(iz).$$

For non-negative real numbers z we have the following inequality

$$\mathrm{atanh}(z) \geq z \geq \mathrm{atan}(z).$$

The atan function (even as a multi-valued function) satisfies

$$\tan(\operatorname{atan}(x) + \operatorname{atan}(y)) = \frac{x+y}{1-xy},$$

for any values of $x, y \in \mathbb{C} \setminus \{\pm i\}$, with $xy \neq 1$.

For $a^2 + b^2 < 1$ the real part of $\operatorname{atan}(a + bi)$ is from $(-\pi/4, \pi/4)$. Hence

$$|x| < 1 \ \wedge |y| < 1 \quad \Longrightarrow \quad \operatorname{atan}(x) + \operatorname{atan}(y) = \operatorname{atan}\left(\frac{x+y}{1-xy}\right). \tag{42}$$

For $a \geq 0$ and $a^2 + b^2 \geq 1$ the real part of $\operatorname{atan}(a + bi)$ is from $[\pi/4, \pi/2)$.

$$a \geq 0 \wedge a^2 + b^2 > 1 \implies \operatorname{atan}(a + bi) + \operatorname{atan}(a - bi) = \pi + \operatorname{atan}(2a/(1 - a^2 - b^2)). \tag{43}$$

For $a \geq 0$ the real part of $\operatorname{atan}(a+bi)$ is from $[0, \pi/2)$. Hence for any a, b with $a + ib \neq \pm i$ we have

$$\operatorname{atan}(a + bi) - \operatorname{atan}(a - bi) = \operatorname{atan}\left(\frac{2ib}{1 + a^2 + b^2}\right). \tag{44}$$