On Learning in the Presence of Friction

Ziyu Zhu

Adviser: Manolis Zampetakis, Assistant Professor of Computer Science

Contents

Acknowledgement

This paper is my Senior Project in partial fulfillment of the requirements of CSEC 491 for the degree of Bachelor of Science in Computer Science & Economics. First and foremost, I am deeply grateful to Alkis Kalavasis, Anay Mehrotra, and Professor Manolis Zampetakis for their unwavering support and valuable insights throughout the development of this paper. Their contributions were crucial to many of the results presented, and their guidance has made this a truly memorable learning experience for me as someone who is new to research. I also extend my sincere thanks to the faculty of both the Computer Science Department and the Economics Department at Yale for their encouragement and support throughout my undergraduate studies. Special thanks to our DUS Professor Philipp Strack for making all these possible.

1. Background

In statistical analysis, one common assumption we often make is that we have access to *independently and identically distributed* (i.i.d.) samples. In practice, this assumption can become fallacious. Problems arise when the data is prone to *systematic bias*, which can happen when the data collection, selection, or processing inadvertently favors certain outcomes, leading to *biased datasets*. Works have been done to identify the source of these biases and to create efficient algorithms for statistical analysis at the presence of the biases. For example, two fundamental types of biases, *truncation bias* and *self-selection bias*, are analyzed and studied in Zampetakis's paper [9]. It is proven that there exist efficient algorithms that manage to estimate the distribution with either kind of bias.

In this paper, we will introduce the *friction model*, in which we generalize the idea of systematic bias into the setting where the data we observe have been transformed through some *friction function*. Traditionally, we think of friction as insensitivity to small changes in the state of the world. This model of friction in economics was first introduced by Rosett [7] in 1959. For example, investors' behavior will not change when the yield fluctuates by a small amount because of transaction costs. In this paper, we will generalize this idea of friction into a wide family of functions, and then analyze the possibility of performing statistical analysis for problems such as Mean Estimation and Linear Regression.

Summary of Results. For both Mean Estimation and Linear Regression, this report gives a sufficient condition to identify the mean of the Gaussian or the regression parameters in the presence of friction (Theorem 8 and Proposition 17). Specifically for Mean Estimation, we also construct unidentifiable instances of friction function for an arbitrary number of Gaussians by drawing a connection to the Consensus-Halving problem (Theorem 9). Eventually, we develop an efficient algorithm that estimates the mean of the Gaussian with certain probability of success (Theorem 16).

2. Friction Model and Problems

For the purpose of this paper, we allow a friction function to be virtually any real function. We will zoom in onto specific kinds of functions in later sections.

Definition 1. A *friction function* is any real-valued function ϕ defined on $\mathbb{R}$. We make the assumption that the friction function is always known to us.

Now we introduce the two main problems this paper will try to solve, both of which are naturally given rise by the friction model.

2.1. **Problem Setting 1 – Mean Estimation.** Let $z_i \in \mathbb{R}$ be samples drawn from a univariate normal $\mathcal{N}(\mu^*, \sigma^2)$, whose variance σ^2 is known. Suppose these x values are transformed through a known friction function ϕ, resulting in observed samples $y_i = \phi(z_i)$. We are interested in (1) whether it is statistically possible to identify the population mean μ^*; and (2) if so, a computationally efficient way to construct an estimator.

Example 1 (Sigmoid)**.** The following example uses the sigmoid function as the friction function. This is the simplest example, and the easiest to deal with – if we observe the red samples, all we need to do is to transform them back to the blue ones with the inverse mapping, and then estimate the population mean.

Example 2 (Classification)**.** For a non-trivial example, consider the following friction function

$$\phi(x) = \begin{cases} 1, & x \in [-1, 1] \\ 0, & x \notin [-1, 1] \end{cases}$$

i.e., $\phi(x) = \mathbb{1}_{[-1,1]}$. This time it is not too obvious that we can still identify the mean only by observing two discrete values. We will show this in a more general setting in the next section. Also note that this friction function corresponds to the standard *classification* problem, where we observe which "class" each sample belongs to and estimate the population statistics.

2.2. **Problem Setting 2 – Linear Regression.** Let $x_i \in \mathbb{R}$ be independent variables. Let noises ζ_i be independent and identically distributed (*i.i.d*) variables drawn from a standard normal distribution $\mathcal{N}(0, 1)$. Consider the friction version of the linear regression problem: for a true parameter $\omega^* \in \mathbb{R}$, we observe the samples $\phi(\omega^* x_i + \zeta_i)$, where ϕ is some known friction function. We are still interested in finding a statistically and computationally efficient way to construct an estimator for ω^*.

Example 3 (Truncation)**.** Besides classification in Example 2, a few other classic examples of statistics using biased data can be represented by the friction models. For example, if we are only able to observe samples that fall into a measurable set $B \subseteq \mathbb{R}$, this is the problem of *truncated* or *censored* samples. In our case, the corresponding friction function is

$$\phi(x) = \mathbb{1}_B(x) \cdot x.$$

Efficient learning from censored or truncated samples has been extensively studied by a few papers. It turns out there is a polynomial-time algorithm that recovers the mean in the truncated version of Mean Estimation (and Variance Estimation as well) [3]. In the truncated version of Linear Regression, there is also a polynomial-time algorithm with success probability of at least 2/3 that provides a close estimate of the regression parameters [4]. Note that

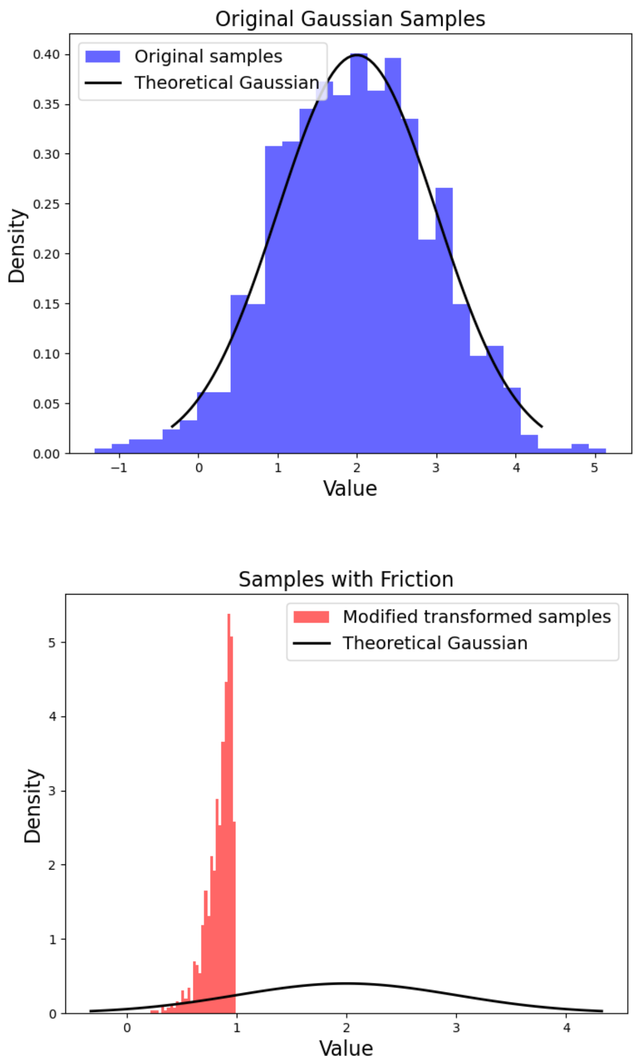

FIGURE 1. One thousand samples drawn from $\mathcal{N}(2, 1)$ with friction function $\phi = \texttt{sigmoid}$.

one more improvement the methods in these papers have over our friction model is the additional assumption that we only need an oracle access to the set B, whereas in our case the friction function is available to us.

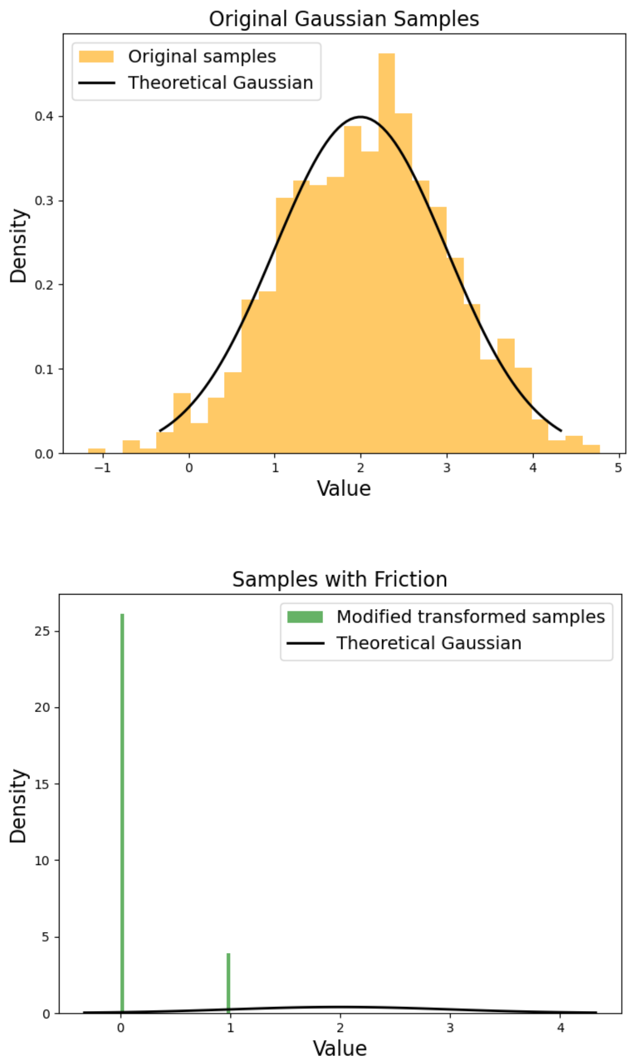

FIGURE 2. One thousand samples drawn from $\mathcal{N}(2,1)$ with friction function $\phi = \mathbb{1}_{[-1,1]}$.

3. PRELIMINARIES

We begin with a family of functions that will be the building blocks of most of our results.

Definition 2. Let $\mathcal{F}$ denote the set of real-valued functions whose level sets are all convex. In other words, if $f \in \mathcal{F}$, then the set $\{x \in \mathbb{R} : f(x) = C\}$ is convex for any $C \in f(\mathbb{R})$.

Remark 1. Trivially, any monotone function is in $\mathcal{F}$. Any 1-to-1 function is in $\mathcal{F}$. This is because for these functions, the level sets are either singletons or intervals, which are convex in $\mathbb{R}$.

Remark 2. Note the nuance between $\mathcal{F}$ and the family of *quasiconvex* functions, for which the definition is the same except that the set $\{x \in \text{dom}(f) : f(x) \leq C\}$ is considered. Not all quasiconvex function is in $\mathcal{F}$, for example, $f(x) = x^2$.

Definition 3. (*Truncated Gaussian Distribution*) Let $\mathcal{N}(\mu, \sigma^2)$ be a normal distribution with mean μ and covariance σ^2, with the probability density function

$$\mathcal{N}(\mu, \sigma^2; x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

Then let $S \subseteq \mathbb{R}^k$ be a measurable subset. We define the *probability mass* of S under this Gaussian measure by

$$\mathcal{N}(\mu, \sigma^2; S) = \int_S \mathcal{N}(\mu, \sigma; x) dx.$$

Finally, we define the *S-truncated normal distribution* as the normal distribution conditioned on that the values fall in S. Equivalently, the truncated normal distribution has the probability density function

$$\mathcal{N}(\mu, \sigma^2, S; x) = \begin{cases} \frac{1}{\mathcal{N}(\mu, \sigma^2; S)} \cdot \mathcal{N}(\mu, \sigma^2; x) & x \in S \\ 0 & x \notin S \end{cases}.$$

4. Mean Estimation

4.1. **Identifiability.** In this section, we will be exploring which kinds of friction functions will allow us to identify the population mean μ^*.

To estimate the population mean $\mu^* \in \mathbb{R}$, we construct the log-likelihood function

$$\mathcal{L}_\phi(\mu; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log\left(\Pr_{x \sim \mathcal{N}(\mu, \sigma^2)}[\phi(x) = y_i]\right),$$

which can be written in the following form assuming we draw infinitely many samples:

$$\mathcal{L}_\phi(\mu; \mu^*) = \mathbb{E}_{z \sim \mathcal{N}(\mu^*, \sigma^2)} \log\left(\Pr_{x \sim \mathcal{N}(\mu, \sigma^2)}[\phi(x) = \phi(z)]\right).$$

As we later shall prove, the Maximum Likelihood Estimator (MLE) $\tilde{\mu} = \arg\max_\mu \mathcal{L}_\phi(\mu;\mu^*)$ will recover μ^*. First, we want to find sufficient conditions under which we can find the optimizer, one of which is strong convexity of $\mathcal{L}_\phi(\mu;\mu^*)$. We wish to prove it for $\mathcal{F}$, the family of friction function ϕ defined. We start with a simple example.

Example 4 (1-to-1 Function)**.** Consider the simplest case where ϕ is any 1-to-1 function (e.g., $\phi = \text{id}$). Then

$$\begin{aligned}\mathcal{L}_\phi(\mu;\mu^*) &= \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\sigma^2)} \log\left(\Pr_{x\sim\mathcal{N}(\mu,\sigma^2)}[x=z]\right)\\ &= \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\sigma^2)} \log\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(z-\mu)^2}{2\sigma^2}}\right)\\ &= \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\sigma^2)}\left[-\frac{(z-\mu)^2}{2\sigma^2} - \log\left(\sqrt{2\pi}\sigma\right)\right].\end{aligned}$$

The first-order derivative is

$$\nabla_\mu \mathcal{L}_\phi(\mu;\mu^*) = \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\sigma^2)}\left[\frac{z-\mu}{\sigma^2}\right]$$

and the second-order derivative is

$$\nabla^2_\mu \mathcal{L}_\phi(\mu;\mu^*) = \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\sigma^2)}\left[-\frac{1}{\sigma^2}\right] = -\frac{1}{\sigma^2} \le 0,$$

so $\mathcal{L}_\phi$ is concave.

We wish to generalize this result to a more representative family of functions.

Proposition 4. *For any $\phi \in \mathcal{F}$, the corresponding log-likelihood $\mathcal{L}_\phi(\mu;\mu^*)$ is concave.*

To prove this result, we make use of the following lemma.

Theorem 5 (Brascamp-Lieb Inequality [2])**.** *Let g be convex function on $\mathbb{R}^d$ and let S be a convex set on $\mathbb{R}^d$. Let $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the Gaussian distribution on $\mathbb{R}^d$. It holds that*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim N_S}\left[g\left(\boldsymbol{x}+\boldsymbol{\mu}-\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S}[\boldsymbol{x}]\right)\right] \le \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}}[g(\boldsymbol{x})].$$

Named as the Brascamp-Lieb Inequality, this inequality was first proven by Brascamp and Lieb for the special case $g(x) = |x|^\alpha$ [2]. Later, Hargé generalized this proof to any convex function g. The proof makes use of optimal transform of measure [5], and will not be discussed in details in this paper. As we later will see, we only need the original version of the inequality proved by Brascamp and Lieb, where $\alpha = 2$.

Proof of Proposition 4: Let $s_1, s_2, \ldots$ (does not need to be finite) take on *distinct* values in the image of f. Let $S_1, S_2, \ldots$ be the corresponding preimages. By definition of $\mathcal{F}$, the S_i's are convex. The corresponding log-likelihood is

$$\begin{aligned}\mathcal{L}_\phi(\mu;\mu^*) &= \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\sigma^2)} \log\left(\Pr_{x\sim\mathcal{N}(\mu,\sigma^2)}[\phi(x)=\phi(z)]\right)\\ &= \int_i \left[\Pr[\phi(z)=s_i]\cdot \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\sigma^2)}\left[\log\left(\Pr_{x\sim\mathcal{N}(\mu,\sigma^2)}[\phi(x)=s_i]\right)\right] \mid z\in S_i\right]\end{aligned}$$

The expectations are not dependent on z, so we can omit them. The above expression simplifies to

$$\begin{aligned}\mathcal{L}_\phi(\mu;\mu^*) &= \int_i \left[\Pr_{z\sim\mathcal{N}(\mu^*,\sigma^2)}[\phi(z)=s_i]\cdot\log\left(\Pr_{x\sim\mathcal{N}(\mu,\sigma^2)}[x\in S_i]\right)\right]\\ &= \int_i \left[\Pr_{z\sim\mathcal{N}(\mu^*,\sigma^2)}[\phi(z)=s_i]\cdot\log\mathcal{N}(\mu,\sigma^2;S_i)\right]. \qquad (4.1)\end{aligned}$$

Below we derive the first-order and the second-order derivative of

$$\log\mathcal{N}(\mu,\sigma^2;S) = \log\int_S \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(z-\mu)^2}{2\sigma^2}}dz,$$

where $S \subseteq \mathbb{R}$ is convex. The first-order derivative is

$$\begin{aligned}\nabla_\mu \log\mathcal{N}(\mu,\sigma^2;S) &= \frac{\int_S \frac{z-\mu}{\sigma^2}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(z-\mu)^2}{2\sigma^2}}dz}{\int_S \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(z-\mu)^2}{2\sigma^2}}dz}\\ &= \frac{\int_S \frac{z-\mu}{\sigma^2}e^{-\frac{(z-\mu)^2}{2\sigma^2}}dz}{\int_S e^{-\frac{(z-\mu)^2}{2\sigma^2}}dz}\\ &= \mathbb{E}_{z\sim\mathcal{N}_S}\left[\frac{z-\mu}{\sigma^2}\right].\end{aligned}$$

The second-order derivative is

$$\begin{aligned}
\nabla^2_\mu \log \mathcal{N}(\mu,\sigma^2;S) &= \frac{\int_S \left(\frac{-1}{\sigma^2}+\frac{(z-\mu)^2}{\sigma^4}\right) e^{-\frac{(z-\mu)^2}{2\sigma^2}}\,dz}{\int_S e^{-\frac{(z-\mu)^2}{2\sigma^2}}\,dz} - \left(\frac{\int_S \frac{z-\mu}{\sigma^2} e^{-\frac{(z-\mu)^2}{2\sigma^2}}\,dz}{\int_S e^{-\frac{(z-\mu)^2}{2\sigma^2}}\,dz}\right)^2 \\
&= \mathbb{E}_{z\sim\mathcal{N}_S}\left[\frac{(z-\mu)^2-\sigma^2}{\sigma^4}\right] - \left(\mathbb{E}_{z\sim\mathcal{N}_S}\left[\frac{z-\mu}{\sigma^2}\right]\right)^2 \\
&= \frac{1}{\sigma^4}\left(\mathrm{Var}_{z\sim\mathcal{N}_S}(z-\mu)-\sigma^2\right) \\
&= \frac{1}{\sigma^4}\left(\mathrm{Var}_{z\sim\mathcal{N}_S}(z)-\mathrm{Var}_{z\sim\mathcal{N}}(z)\right)
\end{aligned}$$

To show this is negative, we use Lemma 5. In our case, consider the convex function $g(z)=z^2$:

$$\begin{aligned}
\mathbb{E}_{z\sim\mathcal{N}_s}\left[(z+\mu-\mathbb{E}_{z\sim\mathcal{N}_s}[z])^2\right] &\le \mathbb{E}_{z\sim\mathcal{N}}\left[z^2\right] \\
\Leftrightarrow \mathbb{E}_{z\sim\mathcal{N}_s}\left[z^2\right] + 2(\mu-\mathbb{E}_{z\sim\mathcal{N}_s}[z])\mathbb{E}_{z\sim\mathcal{N}_s}[z] + (\mu-\mathbb{E}_{z\sim\mathcal{N}_s}[z])^2 &\le \mathbb{E}_{z\sim\mathcal{N}}\left[z^2\right] \\
\Leftrightarrow \quad \mathbb{E}_{z\sim\mathcal{N}_s}\left[z^2\right] - 2\left(\mathbb{E}_{z\sim\mathcal{N}_s}[z]\right)^2 + \mu^2 + \left(\mathbb{E}_{z\sim\mathcal{N}_s}[z]\right)^2 &\le \mathbb{E}_{z\sim\mathcal{N}}\left[z^2\right] \\
\Leftrightarrow \quad \mathbb{E}_{z\sim\mathcal{N}_s}\left[z^2\right] + \mu^2 - \left(\mathbb{E}_{z\sim\mathcal{N}_s}[z]\right)^2 &\le \mathbb{E}_{z\sim\mathcal{N}}\left[z^2\right] \\
\Leftrightarrow \quad \mathbb{E}_{z\sim\mathcal{N}_s}\left[z^2\right] - \left(\mathbb{E}_{z\sim\mathcal{N}_s}[z]\right)^2 &\le \mathbb{E}_{z\sim\mathcal{N}}\left[z^2\right] - \mu^2 \\
\Leftrightarrow \quad \mathrm{Var}_{\mathcal{N}_S}(z) - \mathrm{Var}_{\mathcal{N}}(z) &\le 0
\end{aligned}$$

Then we have

$$\nabla^2_\mu \log \mathcal{N}(\mu,\sigma^2;S_i) \le 0$$

for each S_i. By equation (4.1), the second-order derivative of the log-likelihood function is a linear combination

$$\nabla^2_\mu \mathcal{L}_\phi(\mu;\mu^*) = \int_i \left[\Pr_{z\sim\mathcal{N}(\mu^*,\sigma^2)}[\phi(z)=s_i]\cdot\nabla^2_\mu \log \mathcal{N}(\mu,\sigma^2;S_i)\right] \le 0,$$

hence the log-likelihood function is concave.

□

Remark 3. The 1-to-1 function is a special case of this result where all S_i's are singletons, which are all convex in $\mathbb{R}$. These friction functions will have an infinite number of level sets. Despite identifiability, we will later discuss a way to transform 1-to-1 functions for our algorithm to run.

Note that the concavity proved above does not guarantee identifiability. We still need that the likelihood $\mathcal{L}_\phi(\mu;\mu^*)$ is *strictly concave* around μ^*. To show this, we prove the following lemma.

Lemma 6. *For any convex $S \subseteq \mathbb{R}$, we have*

$$Var_{z\sim\mathcal{N}_S}(z) - Var_{z\sim\mathcal{N}}(z) = 0$$

if and only if $S = \mathbb{R}$, where $\mathcal{N}$ denotes the Gaussian $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}_S$ the S-truncated Gaussian $\mathcal{N}(\mu, \sigma^2, S)$.

Proof: The "if" statement is immediate. To prove the other direction, without loss of generality, it suffices to show the statement for any μ and all convex S centered at 0 (since we are proving this for any μ, we can get all cases by translating S and the Gaussians). We know $\text{Var}_{z\sim\mathcal{N}}(z) = \sigma^2$. We construct the function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$,

$$\psi(s) = \text{Var}_{z\sim\mathcal{N}_{[-s,s]}}(z) - \text{Var}_{z\sim\mathcal{N}}(z) = \text{Var}_{z\sim\mathcal{N}_{[-s,s]}}(z) - \sigma^2.$$

We want to show $\psi(s) = 0$ implies $s = \infty$. Note that we can write $\text{Var}_{z\sim\mathcal{N}_{[-s,s]}}$ as

$$\text{Var}_{z\sim\mathcal{N}_{[-s,s]}}(z) = \int_{[-s,s]} (x - \mathbb{E}_{z\sim\mathcal{N}_{[-s,s]}}[z])^2 \frac{1}{\mathcal{N}(\mu, \sigma^2; [-s,s])} \mathcal{N}(\mu, \sigma^2; x) dx.$$

Let $\mu_s = \mathbb{E}_{z\sim\mathcal{N}_{[-s,s]}}[z]$. Then we can rewrite

$$\begin{aligned}\text{Var}_{z\sim\mathcal{N}_{[-s,s]}}(z) &= \frac{\int_{[-s,s]} (x-\mu_s)^2 \mathcal{N}(\mu, \sigma^2; x) dx}{\mathcal{N}(\mu, \sigma^2; [-s,s])} \\ &= \frac{\int_{[-s,s]} (x-\mu_s)^2 \mathcal{N}(\mu, \sigma^2; x) dx}{\int_{[-s,s]} \mathcal{N}(\mu, \sigma^2; x) dx}.\end{aligned}$$

First thing, we immediately know that $\psi(s)$ is continuous (and differentiable) in s by the Leibniz rule. Besides, we know $\psi(0) = 0 - \sigma^2$ and $\psi(+\infty) = 0$. To reach the desired result, we need to show that $\psi(s)$ is strictly monotone increasing. First, to get some intuition, let's visualize the behavior of the variance on a symmetric set, which is what we expected:

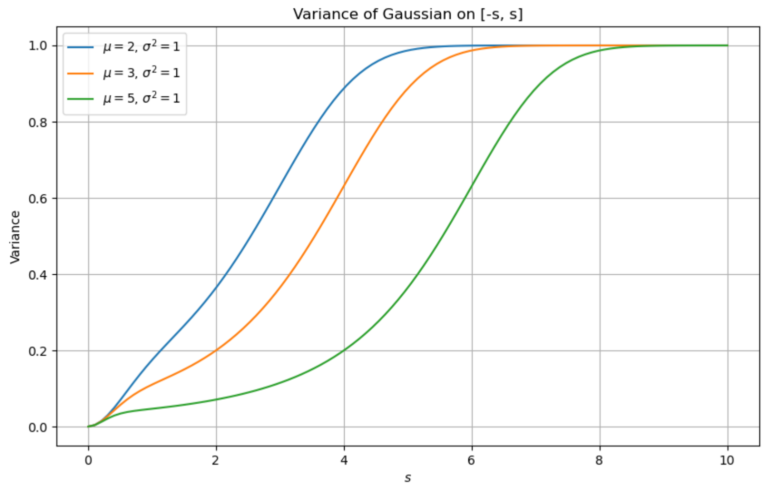

FIGURE 3. $\mathrm{Var}_{z\sim\mathcal{N}_{[-s,s]}}$ are monotone increasing.

To prove this, we keep expanding the variance:

$$\begin{aligned}\mathrm{Var}_{z\sim\mathcal{N}_{[-s,s]}}(z) &= \frac{\int_{[-s,s]}(x-\mu_s)^2\mathcal{N}(\mu,\sigma^2;x)dx}{\int_{[-s,s]}\mathcal{N}(\mu,\sigma^2;x)dx} \\ &= \frac{\int_{[-s,s]}x^2\mathcal{N}(\mu,\sigma^2;x)dx}{\int_{[-s,s]}\mathcal{N}(\mu,\sigma^2;x)dx} - 2\mu_s\frac{\int_{[-s,s]}x\mathcal{N}(\mu,\sigma^2;x)dx}{\int_{[-s,s]}\mathcal{N}(\mu,\sigma^2;x)dx} + \mu_s^2.\end{aligned}$$

Note that

$$\mu_s = \frac{\int_{[-s,s]}x\mathcal{N}(\mu,\sigma^2;x)dx}{\int_{[-s,s]}\mathcal{N}(\mu,\sigma^2;x)dx},$$

so we end up with

$$\mathrm{Var}_{z\sim\mathcal{N}_{[-s,s]}}(z) = \frac{\int_{[-s,s]}x^2\mathcal{N}(\mu,\sigma^2;x)dx}{\int_{[-s,s]}\mathcal{N}(\mu,\sigma^2;x)dx} - \mu_s^2.$$

Now taking derivative w.r.t. s:

$$\frac{\partial}{\partial s}\mathrm{Var}_{z\sim\mathcal{N}_{[-s,s]}}(z) = \frac{(\mathcal{N}(\mu,\sigma^2;s)+\mathcal{N}(\mu,\sigma^2;-s))(s^2\mathcal{N}(S)-\mathbb{E}_S[x^2])}{\mathcal{N}(S)^2} -$$

$$2\cdot\frac{\mathbb{E}_S[x]}{\mathcal{N}(S)}\frac{(\mathcal{N}(\mu,\sigma^2;s)-\mathcal{N}(\mu,\sigma^2;-s))s\mathcal{N}(S)-(\mathcal{N}(\mu,\sigma^2;s)+\mathcal{N}(\mu,\sigma^2;-s))\mathbb{E}_S[x]}{\mathcal{N}(S)^2}$$

Refactor the expression into coefficients of $\mathcal{N}(\mu, \sigma^2; s)$ and $\mathcal{N}(\mu, \sigma^2; -s)$, we get (ignoring the $\mathcal{N}(S)^2$ in the denominator):

$$\text{coef}(\mathcal{N}(\mu, \sigma^2; s)) = \frac{(s\mathcal{N}(S) - \mathbb{E}_S[x])^2 + \mathbb{E}_S[x]^2}{\mathcal{N}(S)} - \mathbb{E}_S[x^2]$$
$$\text{coef}(\mathcal{N}(\mu, \sigma^2; s)) = \frac{(s\mathcal{N}(S) + \mathbb{E}_S[x])^2 + \mathbb{E}_S[x]^2}{\mathcal{N}(S)} - \mathbb{E}_S[x^2]$$

which are both positive if we expand them back to the integral form.

□

Therefore, to get identifiability, we only need to remove all the constant functions from $\mathcal{F}$. The strict concavity points to possibility of efficiently estimating μ^* for friction functions $\phi \in \mathcal{F}$. Before moving on, we check that the maximum likelihood estimator indeed identifies the true mean.

Proposition 7. *The log-likelihood function $\mathcal{L}_\phi(\mu; \mu^*)$ achieves maximum at $\mu = \mu^*$.*

Proof: It suffices to show $\nabla_\mu \mathcal{L}_\phi(\mu; \mu^*) = 0$ at $\mu = \mu^*$. There is

$$\begin{aligned}
\nabla_\mu \mathcal{L}_\phi(\mu = \mu^*; \mu^*) &= \sum_i \left[\Pr_{z \sim \mathcal{N}(\mu^*, \sigma^2)} [\phi(z) = s_i] \cdot \nabla_\mu \log \mathcal{N}(\mu^*, \sigma^2; S_i) \right] \\
&= \sum_i \left[\mathcal{N}(\mu^*, \sigma^2; S_i) \cdot \nabla_\mu \log \mathcal{N}(\mu^*, \sigma^2; S_i) \right] \\
&= \sum_i \left[\int_{S_i} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu^*)^2}{2\sigma^2}} dz \cdot \frac{\int_{S_i} \frac{z-\mu^*}{\sigma^2} e^{-\frac{(z-\mu^*)^2}{2\sigma^2}} dz}{\int_{S_i} e^{-\frac{(z-\mu^*)^2}{2\sigma^2}} dz} \right] \\
&= \sum_i \int_{S_i} \frac{1}{\sqrt{2\pi}\sigma} \frac{z-\mu^*}{\sigma^2} e^{-\frac{(z-\mu^*)^2}{2\sigma^2}} dz \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} \frac{z-\mu^*}{\sigma^2} e^{-\frac{(z-\mu^*)^2}{2\sigma^2}} dz \\
&= \mathbb{E}_{z \sim \mathcal{N}(\mu^*, \sigma^2)} \left[\frac{z-\mu^*}{\sigma^2} \right] \\
&= 0,
\end{aligned}$$

as desired. □

Combining all the results above, we get the following theorem.

Theorem 8. *μ^* is identifiable for any friction function $\phi \in \mathcal{F}$ that are not constant.*

The construction of the family of functions $\mathcal{F}$ is fairly natural. Consider the following non-example that lies out of $\mathcal{F}$.

Example 5 (Unidentifiable)**.** Let $\phi(x) = \begin{cases} 1, & x \notin [-1,1] \\ 0, & x \in [-1,1] \end{cases}$. We see $\phi \notin \mathcal{F}$ since $\phi^{-1}(1)$ is not convex in $\mathbb{R}$. Now suppose we sample from the Gaussian $\mathcal{N}(\mu^*, 1)$ where $\mu^* > 1$ and feed the samples through ϕ. It is impossible to identify μ^* with any means, for the simple reason that ϕ is symmetric across $x = 0$ and demolishes all information about the sign of the sample. Specifically, the log-likelihood function $\mathcal{L}_\phi(\mu; \mu^*)$ cannot be concave, because if $\tilde{\mu}$ maximizes $\mathcal{L}_\phi(\mu; \mu^*)$, then there is $\mathcal{L}_\phi(\tilde{\mu}; \mu^*) = \mathcal{L}_\phi(-\tilde{\mu}; \mu^*)$, and it is easy to show that $\mathcal{L}_\phi(\tilde{\mu}; \mu^*) \neq \mathcal{L}_\phi(0; \mu^*)$.

Meanwhile, the family $\mathcal{F}$ leaves out some friction functions for which the mean is still identifiable. Consider the following non-symmetric example.

Example 6 (Asymmetric Identifiability)**.** Consider the friction function $\phi(x) = \begin{cases} 1, & x \in (-1,0] \cup [1,\infty) \\ 0, & x \in (-\infty,-1] \cup (0,1) \end{cases}$. The level sets are not constant. However, the mean is still identifiable in this case, because there is a 1-to-1 mapping between $\mu \in \mathbb{R}$ and the probability mass that $\mathcal{N}(\mu, \sigma^2)$ assigns to the set $(-1,0] \cup [1,\infty)$. In other words, one can check that the probability mass of $S = (-1,0] \cup [1,\infty)$, i.e.,

$$\mathcal{N}(\mu, \sigma^2; S) = \int_{(-1,0]} \mathcal{N}(\mu, \sigma^2; x)dx + \int_{[1,\infty]} \mathcal{N}(\mu, \sigma^2; x)dx$$

is monotone increasing in μ. Therefore, estimating $\mathcal{N}(\mu, \sigma^2; S)$ with the observed samples would allow us to recover the population mean.

4.2. **Non-Identifiability.** In the previous subsection, we tried to solve the question: what friction functions will allow us to identify the Gaussian's mean? In this section, we want study the reverse problem, that whether we can find friction functions that will make it impossible to identify the mean. This problem has a trivial answer, which is yes because we can always make two Gaussians unidentifiable by constructing a friction function that is symmetric between the two Gaussians, so that each Gaussian assign the same probability weight to each outcome.

A more interesting question is: **given $m > 2$ Gaussians, can we still find a nontrivial friction function that makes the m Gaussians unidentifiable?** Here, nontrivial means the friction function is not constant, because by the discussion above, a constant function trivially leads to unidentifiability.

We first present the final theorem of this section, and make our way up there.

Theorem 9. *Given $m > 2$ Gaussians, there always exist a friction function ϕ such that $\phi(x) \in \{0, 1\}$ and ϕ uses at most m discontinuity points, and ϕ makes the given Gaussians unidentifiable.*

This result is derivative from the famous *Consensus-Halving* (or *Cake-Cutting*) problem.

Definition 10 (Consensus-Halving)**.** Let *object* A be a measurable bounded set in $\mathbb{R}$. There are m *features*. Each feature i is a bounded continuous (w.r.t Lebesgue) measure μ_i. The *Consensus-Halving* problem studies if we can cut the object A into two parts so that each feature is divided evenly into the two parts (equal measure).

The following theorem is established by Alon and West [1] on the Consensus-Halving problem.

Lemma 11. *The object A can be divided into two portions A_+ and A_- using $m + 1$ pieces, such that $\mu_i(A_+) = \mu_i(A_-)$ for all $i \in \{1, \ldots, m\}$.*

Proof: Since A is bounded, it suffices to show the theorem for $A = [0, 1]$. We define a m-cut $x \in \mathbb{R}^m$ using $x_1, x_2, \ldots, x_m \in [-1, 1]$. The actual cuts will be the absolute values $|x_1|, |x_2|, \ldots, |x_m|$. This results in $m + 1$ intervals, which we call $S_1, S_2, \ldots, S_{m+1}$. We always assign S_{m+1} to A_- and assign S_k to A_+ if $x_k < 0$ and to A_- otherwise for all $1 \leq k \leq m$.

By this procedure, we will get the measure of the i^{th} feature on A_+ defined by the cut x as

$$\mu_i(A_+)(x) = \sum_{j=1}^{m} \mathbb{1}\{x_j < 0\} \cdot \left(\mu_i \left(\sum_{k=1}^{j} |x_k| \right) - \mu_i \left(\sum_{k=1}^{j-1} |x_k| \right) \right)$$

and similarly

$$\mu_i(A_-)(x) = 1 - \mu_i \left(\sum_{k=1}^{m} |x_k| \right) + \sum_{j=1}^{m} \mathbb{1}\{x_j \geq 0\} \cdot \left(\mu_i \left(\sum_{k=1}^{j} |x_k| \right) - \mu_i \left(\sum_{k=1}^{j-1} |x_k| \right) \right).$$ $\blacksquare$

We wish to prove the existence of $x \in \mathbb{R}^m$ such that $\mu_i(A_+)(x) = \mu_i(A_-)(x)$ for all i. Define the difference

$$\begin{aligned} f_i(x) &= \mu_i(A_-)(x) - \mu_i(A_+)(x) \\ &= 1 - \mu_i\left(\sum_{k=1}^{m}|x_k|\right) + \sum_{j=1}^{m}\text{sgn}(x_j)\cdot\left(\mu_i\left(\sum_{k=1}^{j}|x_k|\right) - \mu_i\left(\sum_{k=1}^{j-1}|x_k|\right)\right). \end{aligned}$$

Obviously f_i is continuous. Then define $F : \mathcal{B}^m(0,1) \to \mathbb{R}^m$ such that $F_i(x) = f_i(x)$. F is continuous, and note that by the property of a cut, F is defined with the boundary $S^m = \{|x_1| + \ldots + |x_m| = 1\}$. At the boundary, one can check the antipodal condition $F(x) = -F(-x)$ for all $x \in S^m$, since

$$\begin{aligned} f_i(x) &= \sum_{j=1}^{m}\text{sgn}(x_j)\cdot\left(\mu_i\left(\sum_{k=1}^{j}|x_k|\right) - \mu_i\left(\sum_{k=1}^{j-1}|x_k|\right)\right) \\ &= -\sum_{j=1}^{m}\text{sgn}(-x_j)\cdot\left(\mu_i\left(\sum_{k=1}^{j}|x_k|\right) - \mu_i\left(\sum_{k=1}^{j-1}|x_k|\right)\right) = -f_i(-x). \end{aligned}$$

By the Borusk-Ulam theorem, this indicates there exists $x^* \in B^m(0,1)$ such that

$$F(x^*) = 0, \tag{4.2}$$

then the division $A^+(x^*)$ and $A^-(x^*)$ is the desired division.

□

Remark 4. Note that this only shows existence. An efficient algorithm to construct an ε-approximate solution is given by Simmons and Su [8]. As we will see below, our theorem is a direct derivative of the Consensus-Halving problem. Hence, the construction will also help us construct a ε-approximate friction function.

Proof of Theorem 9: In our version of the problem, each Gaussian represents a feature, with $\mu_i(x) = \mathcal{N}(\mu, \sigma^2; x)$. The object to be divided is $\mathbb{R}$. The problem with applying the above lemma directly to this problem is that $\mathbb{R}$ is not bounded. A way to get around this is to transform the intervals to into their probability mass under the standard normal distribution. Concretely, for each cut length ℓ_i, we can transform it from $x_i > 0$ such that

$$\int_{x_{i-1}}^{x_i} \mathcal{N}(0,1;z)dz = |\ell_i|.$$

In this way, if we are going to partition the unbounded set $\mathbb{R}$ into intervals separated by $x_i's$, we have transformed the set into a bounded set $[0,1]$, because

$$|\ell_1| + |\ell_2| + \ldots + |\ell_m| = \sum_{i=1}^{m}\int_{x_{i-1}}^{x_i}\mathcal{N}(0,1;z)dz = \int_{-\infty}^{\infty}\mathcal{N}(0,1;z)dz = 1$$

The only remaining thing we need to do is to transform the features (which are Gaussians) onto the set $[0,1]$. Since the transformation above is bounded and continuous, the transformed features will still be bounded continuous measures. Using Lemma 11, we will get a m-cut on the set $[0,1/2]$ that assigns equal mass to each of the transformed features on the two divided subsets. Transforming the cut back using equation (4.2), we get a m-cut on $\mathbb{R}$ that assign equal mass to each of the original features. In other words, let $\mathbb{R}_+$ and $\mathbb{R}_-$ be the final divisions. Then we have

$$\mathcal{N}(\mu_i, \sigma^2; \mathbb{R}_+) = \mathcal{N}(\mu_i, \sigma^2; \mathbb{R}_-) = \frac{1}{2}$$

for all $1 \leq i \leq m$. In particular, we have

$$\mathcal{N}(\mu_i, \sigma^2; \mathbb{R}_+) = \mathcal{N}(\mu_j, \sigma^2; \mathbb{R}_+) \tag{4.3}$$

for any $i, j \in \{1, 2, \ldots, m\}$. Now consider the friction function

$$\phi(x) = \mathbb{1}_{\mathbb{R}_+}(x).$$

The m Gaussians will not be identifiable because each of them assign equal probability mass to $\mathbb{R}_+$, as shown by equation (4.3), which completes the proof.

□

4.3. **Algorithm.** Going back to the log-likelihood function in equation (4.1), in this section we propose an efficient algorithm to estimate μ^* using *Projected Stochastic Gradient Descent* (PSGD) to optimize the log-likelihood. Before we start, we yet need to refine the class of friction functions we are working with. Inside $\mathcal{F}$, consider the following class of functions.

Definition 12. Let $\mathfrak{S}$ denote the set of all the level sets of ϕ. Let μ define the distribution $\mathcal{N}_{\mathfrak{S}}(\mu)$ over $\mathfrak{S}$ that is naturally induced by the probability mass of $\mathcal{N}(\mu, \sigma^2)$ on the elements of $\mathfrak{S}$, i.e.,

$$\Pr_{T \sim \mathcal{N}_{\mathfrak{S}}(\mu)} [T = S_i] = \Pr_{z \sim \mathcal{N}(\mu)} [z \in S_i].$$

Definition 13. (Information Preservation) Let $\lambda \in [0,1]$. We say $\mathfrak{S}$ is *λ-info preserving* with regard to $\mathcal{N}(\mu^*)$ if for any μ, there is

$$\mathrm{KL}\left(\mathcal{N}_{\mathfrak{S}}(\mu) \parallel \mathcal{N}_{\mathfrak{S}}(\mu^*)\right) \geq \lambda \cdot \mathrm{KL}\left(\mathcal{N}(\mu) \parallel \mathcal{N}(\mu^*)\right),$$

where $\mathrm{KL}(\cdot \parallel \cdot)$ denotes the Kullback–Leibler divergence.

Remark 5. We consider functions with the information preservation property because this would result in strong concavity. Specifically, if $\mathcal{N}(\mu)$ and $\mathcal{N}(\mu^*)$ coincide, there must be $\mu = \mu^*$. Note that our previous class of function $\mathcal{F}$ overlaps non-trivially with these functions.

For the rest of this section, we make the following assumptions:

Assumption 1. The set of level sets of the friction function $\phi \in \mathcal{F}$ is λ-info preserving.

Assumption 2. The number of values that ϕ takes is bounded by constant. (This leaves out the most common functions such as the continuous ones. We will later discuss a way to get around this assumption.)

Assumption 3. The absolute value of μ^* is bounded by some $k \in \mathbb{R}_{\geq 0}$, i.e., $\mu^* \in [-k, k]$. The bound k is known to us.

Given these assumptions, we propose the following PSGD algorithm for Mean Estimation:

Algorithm 1 Projected SGD for Mean Estimation

Input: optimizer bounded by k, friction function ϕ, data $\phi(x_{t_i})_{i=1}^N$.

$\mu^{(0)} \leftarrow$ arbitrary point in $\mathcal{P} = [-k, k]$.
for $t = 1, 2, \ldots, T$ **do**
  for each level set S_i of ϕ **do**
    sample $(z_i)_t$ drawn from $\mathcal{N}(\mu^{(t)}, \sigma^2, S_i)$
  end for
  calculate the gradient estimate $\xi^{(t)}$ using $(z_i)_t$ and equation (4.1)
  by unbiased gradient, $\mathbb{E}\left[\xi^{(t)} \mid \mu^{(t-1)}\right] = \nabla_\mu \mathcal{L}_\phi(\mu^{t-1}; \mu^*)$
  $\chi^{(t)} \leftarrow \mu^{(t-1)} - \frac{1}{\sqrt{t}} \xi^{(t)}$
  $\mu^{(t)} \leftarrow \arg\min_{\mu \in \mathcal{P}} \left\| \mu - \chi^{(t)} \right\|$
end for
return $\mu^{(T)}$

To show this algorithm works and finishes in finite time, we have the following theorem on SGD convergence from Shamir and Zhang [6]:

Theorem 14. *Suppose that F is convex, and that for some constants D, G, it holds that $\mathbb{E}\left[\|\xi_t\|^2\right] \leq G^2$ for all t, and $\sup_{\mu, \mu' \in \mathcal{P}} \|\mu - \mu'\| \leq D$. Consider SGD with step sizes $\eta_t = c/\sqrt{t}$ where $c > 0$ is a constant. Then for any $T > 1$, it holds that*

$$\mathbb{E}\left[F\left(\mu^{(T)}\right) - F\left(\mu^*\right)\right] \leq \left(\frac{D^2}{c} + cG^2\right) \frac{2 + \log(T)}{\sqrt{T}}.$$

We check the following conditions for our PSGD algorithm:

Proposition 15. *The following conditions are met:*

(1) $\sup_{\mu,\mu'\in\mathcal{P}} |\mu - \mu'|$ *is bounded.*
(2) *The PSGD has an initial starting point in the projection set.*
(3) *The stochastic gradient is unbiased.*
(4) *The second moment of the stochastic gradient is bounded.*
(5) *The log-likelihood function is convex everywhere, and strongly convex around* μ^**.*

Proof: (1) and (2) is trivial given by Assumption 2. (3) is given by equation (4.1). (5) is immediate from Proposition 4 and the definition of information preservation. To check (4), notice that the stochastic gradient is

$$\sum_S \mathcal{N}(\mu^*, \sigma^2; S)(x - z_S),$$

where $x \sim \mathcal{N}(\mu, \sigma^2)$ and $z_S \sim \mathcal{N}(\mu, \sigma^2, S)$. Since x and z are independent, the variance of the stochastic gradient is equal to

$$\mathrm{Var}(x) + \sum_S \Big(\mathcal{N}(\mu^*, \sigma^2; S)\Big)^2 \mathrm{Var}(z_S).$$

By Proposition 4, we have $\mathrm{Var}(z_S) \le \mathrm{Var}(x) = \sigma^2$. Hence, the variance is bounded by

$$\sigma^2 \left(1 + \sum_S \mathcal{N}(\mu^*, \sigma^2; S)^2\right) = \sigma^2 \left(1 + \sum_S \mathcal{N}^*(S)^2\right).$$

Note that $\sum_S \mathcal{N}^*(S) = 1$. Since each $\mathcal{N}^*(S)$ is non-negative, we easily get $\sum_S \mathcal{N}^*(S)^2 \le 1$, hence an upper bound of the variance by $2\sigma^2$. To show the second moment is bounded, we still need to show the expected value of the stochastic gradient is bounded. This is a more tricky, and it comes down to proving that

$$\sum_S \mathcal{N}(\mu^*, \sigma^2; S)\mathbb{E}_{\mathcal{N}_S}[|x|]$$

is bounded. We can prove an approximate version of this by considering the intersection between S's and the ball $\mathcal{B}(0, k + \log\frac{1}{\delta})$ for some δ, and then all $|x|$ will have bounded values.

□

Note that Theorem 14 only shows the expected convergence. To get a certain probability of success, we use the following "boosting" trick from [3]: we can amplify the probability of success to $1 - \delta$ by repeating the optimization process $\log(\frac{1}{\delta})$ times from scratch and keep the optimizer that achieves the maximum log-likelihood. Combining these results, we have the following theorem:

Theorem 16. *Given the bound of true mean k, the bound of the second moment of stochastic gradient G^2, the friction function ϕ, and observed data $\phi(x_{t_i})_{i=1}^N$, repeat the PSGD algorithm we proposed with step sizes $\eta_t = c/\sqrt{t}$ and step numbers $T > 1$ for $\log(\frac{1}{\delta})$ times independently and keep the best optimizer $\hat{\mu}$. Then it holds with probability $1 - \delta$ that*

$$\mathcal{L}_\phi(\mu^*; \mu^*) - \mathcal{L}_\phi(\hat{\mu}; \mu^*) \leq \left(\frac{k^2}{c} + cG^2 \right) \frac{2 + \log(T)}{\sqrt{T}},$$

in other words the total runtime is $T \log(\frac{1}{\delta})$.

Remark 6. To try to loosen up our assumptions, consider any monotone continuous function, which is definitely identifiable, yet will not be accepted by our algorithm. One way to make such a function pass is by *discretizing* such a function into a step function, i.e., to give up information on the exact value of the samples we see and only observe the bins of the values. This is not optimal, and whether there exists a more efficient alternative is an open question.

5. Linear Regression

For convenience, set $z_i = \omega^* x_i + \zeta_i$ for each i, then $z_i \sim \mathcal{N}(\omega^* x_i, 1)$. We construct the log-likelihood function

$$\mathcal{L}_\phi(\omega; x_1, \ldots, x_n, \phi(z_1), \ldots, \phi(z_n)) = \sum_{i=1}^{n} \log \left(\Pr_{\alpha_i \sim \mathcal{N}(\omega x_i, 1)} [\phi(\alpha_i) = \phi(z_i)] \right),$$

which we write as $\mathcal{L}_\phi(\omega; \omega^*)$ for short. Notice that this log-likelihood function is different from the previous one in that the α_i's are drawn from different distributions rather than the same one. Still, we get similar results as before:

Proposition 17. *For $\phi \in \mathcal{F}$, the corresponding log-likelihood $\mathcal{L}_\phi(\omega; \omega^*)$ is concave.*

Proof: Write

$$\mathcal{L}_\phi(\omega; \omega^*) = \sum_{i=1}^{n} \log \left(\Pr_{\alpha_i \sim \mathcal{N}(\omega x_i, 1)} [\phi(\alpha_i) \in S_i] \right)$$

for convex $S_i = \phi^{-1}(\phi(z_i)) \subseteq \mathbb{R}$. We proved in Proposition 4 that the logarithm in the expression above is concave with respect to the mean ωx_i. Since x_i is known, the logarithm is concave with respect to ω, hence $\mathcal{L}_\phi(\omega; \omega^*)$ is concave.

□

6. Open Problems

As we mentioned in the previous sections, there are many results that have not been solved by this paper. We list a few open questions stemming from our discussions:

- Does there exist an exact way to characterize the family of friction functions that leads to identifiability in Mean Estimation?
- Besides running the PSGD algorithm on the discretized function, is there a more efficient algorithm to optimize the log-likelihood associated with functions whose number of level sets is not bounded?
- Can we develop a similar PSGD algorithm for Linear Regression? The major difficulty we will encounter is that we will have to sample without replacement at each step. An algorithm with such technique has been developed for the truncation version of regression [4].

References

[1] Alon, N., & West, D. B. (1986). The Borsuk-Ulam theorem and bisection of necklaces. *Proceedings of the American Mathematical Society, 98*(4), 623–628. https://doi.org/10.1090/s0002-9939-1986-0861764-9

[2] Brascamp, H. J., & Lieb, E. H. (1976). On extensions of the brunn-minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis, 22*(4), 366–389. https://doi.org/10.1016/0022-1236(76)90004-5

[3] Daskalakis, C., Gouleakis, T., Tzamos, C., & Zampetakis, M. (2018). Efficient statistics, in high dimensions, from truncated samples. *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. https://doi.org/10.1109/focs.2018.00067.

[4] Daskalakis, C., Gouleakis, T., Tzamos, C., & Zampetakis, M. (2019). Computationally and statistically efficient truncated regression. *Proceedings of the Thirty-Second Conference on Learning Theory,* PMLR 99, 955-960. `https://proceedings.mlr.press/v99/daskalakis19a.html`

[5] Hargé, G. (2004). A convex/log-concave correlation inequality for gaussian measure and an application to abstract Wiener Spaces. *Probability Theory and Related Fields, 130*(3), 415–440. https://doi.org/10.1007/s00440-004-0365-8

[6] Shamir, O., Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. *Proceedings of the 30th International Conference on Machine Learning, PMLR 28*(1), 71-79. `https://proceedings.mlr.press/v28/shamir13.html.`

[7] Rosett, R. N. (1959). A statistical model of friction in Economics. *Econometrica, 27*(2), 263. https://doi.org/10.2307/1909447

[8] Simmons, F. W., & Su, F. E. (2003). Consensus-halving via theorems of Borsuk-Ulam and Tucker. *Mathematical Social Sciences, 45*(1), 15–25. https://doi.org/10.1016/s0165-4896(02)00087-2

[9] Zampetakis, M. (2022). Analyzing data with systematic bias. *ACM SIGecom Exchanges*, *20*(1), 55–63. https://doi.org/10.1145/3572885.3572890