A Proofs and Calculations Regarding the Objective

A.1 The Truncated Negative Expected Log-Likelihood Function

The negative log-likelihood that $\mathbf{x} \in S$ is a sample of $p_{\boldsymbol{\theta}}^S(\mathbf{x})$ is

$$\underbrace{\ell(\theta, \mathbf{x})}_{-\log p_{\boldsymbol{\theta}}^S(\mathbf{x})} := -\log h(\mathbf{x}) - \boldsymbol{\theta}^\top T(\mathbf{x}) + \log \int_S h(\mathbf{x}') \exp(\theta^\top T(\mathbf{x}')) d\mathbf{x}'.$$

Its gradient w.r.t. $\boldsymbol{\theta}$ is

$$\begin{aligned}
\nabla \ell(\boldsymbol{\theta}, \mathbf{x}) &= -T(\mathbf{x}) + \frac{\int_S T(\mathbf{x}') h(\mathbf{x}') \exp(\theta^\top T(\mathbf{x}')) d\mathbf{x}'}{\int_S h(\mathbf{x}') \exp(\theta^\top T(\mathbf{x}')) d\mathbf{x}'} \\
&= -T(\mathbf{x}) + \frac{\int_S T(\mathbf{x}') h(\mathbf{x}') \exp(\theta^\top T(\mathbf{x}') - A(\boldsymbol{\theta})) d\mathbf{x}'}{\int_S h(\mathbf{x}') \exp(\boldsymbol{\theta}^\top T(\mathbf{x}') - A(\boldsymbol{\theta})) d\mathbf{x}'} \\
&= -T(\mathbf{x}) + \mathbb{E}_{\mathbf{z} \sim p_{\boldsymbol{\theta}}^S}[T(\mathbf{z})]
\end{aligned}$$

The Hessian is

$$\begin{aligned}
\nabla^2 \ell(\boldsymbol{\theta}) = & \frac{(\int_S T(\mathbf{x}) T(\mathbf{x})^\top h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) d\mathbf{x})}{(\int_S h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) d\mathbf{x})} \\
& - \frac{(\int_S T(\mathbf{x}) h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) d\mathbf{x})}{(\int_S h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) d\mathbf{x})} \cdot \left(\frac{(\int_S T(\mathbf{x}) h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) d\mathbf{x})}{(\int_S h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) d\mathbf{x})} \right)^\top \\
= & \, \mathbf{Cov}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}^S}[T(\mathbf{x}), T(\mathbf{x})]
\end{aligned}$$

We can similarly define the population negative log-likelihood as

$$\begin{aligned}
\bar{\ell}(\boldsymbol{\theta}) &:= \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}^*}^S} \left[-\log h(\mathbf{x}) - \boldsymbol{\theta}^\top T(\mathbf{x}) \right] + \log \int_S h(\mathbf{x}) \exp(\theta^\top T(\mathbf{x})) d\mathbf{x}), \\
\nabla \bar{\ell}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}^*}^S} \left[-T(\mathbf{x}) \right] + \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}^S} \left[T(\mathbf{x}) \right], \\
\nabla^2 \bar{\ell}(\boldsymbol{\theta}) &= \nabla^2 \ell(\boldsymbol{\theta})
\end{aligned}$$

A.2 Proof of Lemma 3.2

Proof. Define the following quantities:

$$\begin{aligned}
\mathbf{R}^* &= \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}} \left[(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}}[T(\mathbf{x})]) \cdot (T(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}}[T(x)])^\top \right] \\
\mathbf{R}' &= \mathbb{E}_{x \sim p_{\boldsymbol{\theta}}} \left[\Big(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}^S}[T(\mathbf{x})] \Big) \cdot \Big(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}^S}[T(\mathbf{x})] \Big)^\top \right] \\
\mathbf{R} &= \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}^S} \left[\Big(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}^S}[T(\mathbf{x})] \Big) \cdot \Big(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}^S}[T(\mathbf{x})] \Big)^\top \right]
\end{aligned}$$

Claim 2. $R' \succeq R^*$. (Proof in Appendix A.4.)

Now, let $\xi \in \mathbb{R}^k$ with $\|\xi\|_2^2 = 1$ arbitrary. Then

$$\begin{aligned}
\xi^\top \mathbf{R}^* \xi &= \xi^\top \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}} \left[(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}}[T(\mathbf{x})]) \cdot (T(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}}[T(\mathbf{x})])^\top \right] \xi = \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}} \left[p_\xi(\mathbf{x}) \right] \\
\xi^\top \mathbf{R}' \xi &= \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}} \left[p'_\xi(\mathbf{x}) \right] \\
\xi^\top \mathbf{R} \xi &= \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}^S} \left[p'_\xi(\mathbf{x}) \right]
\end{aligned}$$

where $p_\xi(\mathbf{x}), p'_\xi(\mathbf{x})$ are polynomials of degree at most $2d$ whose coefficients depend on ξ (under A3). Furthermore, note that for any $\xi \in \mathbb{R}^k$, $p_\xi(\mathbf{x}) \geq 0$ and $p'_\xi(\mathbf{x}) \geq 0$ (due to the rank one matrix inside the expectation being PSD).

First, since $\mathbf{R}' \succeq \mathbf{R}^* \iff \xi^\top \mathbf{R}' \xi \geq \xi^\top \mathbf{R}^* \xi$, we have

$$\mathbb{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}}\left[p'_\xi(\mathbf{z})\right] \geq \mathbb{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}}\left[p_\xi(\mathbf{z})\right] \geq \lambda.$$

Now define the set $A := \{\mathbf{x} : p'_\xi(\mathbf{x}) \leq \gamma\}$ for $\gamma = \left(\frac{\beta}{4Cd}\right)^{2d} \lambda$ where $p_{\boldsymbol{\theta}}(S) = \beta > 0$. Theorem 8 of [5] says

$$p_{\boldsymbol{\theta}}(A) \leq \frac{Cq\gamma^{1/(2d)}}{\left(\mathbb{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}}\left[p'_\xi(\mathbf{z})\right]^{q/2d}\right)^{1/q}} \overset{q=2d}{=} \frac{2Cd\gamma^{1/(2d)}}{\left(\underbrace{\mathbb{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}}\left[p'_\xi(\mathbf{z})\right]}_{\geq \lambda}\right)^{1/(2d)}} \leq \frac{2Cd \cdot \gamma^{1/(2d)}}{\lambda^{1/(2d)}} = \frac{\beta}{2}.$$

Now we can split $\mathbb{E}_{\mathbf{z}\sim p^S_{\boldsymbol{\theta}}}\left[p'_\xi(\mathbf{z})\right]$ into the part on $S \cap A$ and $S \cap A^c$. Note that if $p_{\boldsymbol{\theta}}(S) = \beta$ and $p_\theta(A) \leq \frac{\beta}{2}$, this implies $p_{\boldsymbol{\theta}}(S \cap A^c) \geq \frac{\beta}{2}$ as

$$p_{\boldsymbol{\theta}}(S \cap A^c) \geq p_\theta(S) + p_{\boldsymbol{\theta}}(A^c) - p_{\boldsymbol{\theta}}(S \cup A^c) \geq \beta + \left(1 - \frac{\beta}{2}\right) - 1 = \frac{\beta}{2}.$$

Then

$$\mathbb{E}_{\mathbf{z}\sim p^{S\cap A}_{\boldsymbol{\theta}}}\left[p'_\xi(\mathbf{z})\right] + E_{\mathbf{z}\sim p^{S\cap A^c}_{\boldsymbol{\theta}}}\left[p'_\xi(\mathbf{z})\right] \geq \frac{p_{\boldsymbol{\theta}}(S \cap A)}{p_{\boldsymbol{\theta}}(S)} \cdot 0 + \frac{p_{\boldsymbol{\theta}}(S \cap A^c)}{p_{\boldsymbol{\theta}}(S)} \cdot \gamma \geq \frac{1}{2}\gamma \Rightarrow \mathbb{E}_{\mathbf{z}\sim p^S_{\boldsymbol{\theta}}}\left[p'_\xi(\mathbf{z})\right] \geq \frac{1}{2}\left(\frac{\beta}{4Cd}\right)^{2d} \lambda$$

and the claim follows. □

A.3 Proof of Lemma 3.3

Proof. Similar to the proof of the previous lemma, define the following quantities:

$$\begin{aligned}
\mathbf{R}^* &= \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}\left[(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}[T(\mathbf{x})]) \cdot (T(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}[T(\mathbf{x})])^\top\right] \\
\mathbf{R}'' &= \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}\left[(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}[T(\mathbf{x})]) \cdot (T(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}[T(\mathbf{x})])^\top\right] \\
\mathbf{R} &= \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}\left[\left(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[T(\mathbf{x})]\right) \cdot \left(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[T(\mathbf{x})]\right)^\top\right]
\end{aligned}$$

Claim 3. It holds that $\mathbf{R}'' \succeq \mathbf{R}$. (Similar proof to Claim 2.)

Let $\xi \in \mathbb{R}^k$ with $\|\xi\|_2^2 = 1$ arbitrary. Then

$$\begin{aligned}
\xi^\top \mathbf{R}^* \xi &= \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}[f_\xi(\mathbf{x})] \\
\xi^\top \mathbf{R}'' \xi &= \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[f_\xi(\mathbf{x})] \\
\xi^\top \mathbf{R} \xi &= \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[f'_\xi(\mathbf{x})]
\end{aligned}$$

where $f_\xi(\mathbf{x}), f'_\xi(\mathbf{x})$ are some functions which depend on $\mathbf{x}$ and ξ (e.g., polynomials of degree at most $2d$ under A3). By the previous claim, we also have

$$\mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[f_\xi(\mathbf{x})] \geq \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[f'_\xi(\mathbf{x})].$$

Note that

$$\mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[f_\xi(\mathbf{x})] = \int_{\mathcal{X}} p^S_{\boldsymbol{\theta}}(\mathbf{x}) \cdot f_\xi(\mathbf{x}) d\mathbf{x} = \int_{\mathcal{X}} \frac{1}{p_{\boldsymbol{\theta}}(S)} p_{\boldsymbol{\theta}}(\mathbf{x}) \cdot f_\xi(\mathbf{x}) \cdot \mathbb{1}\{\mathbf{x} \in S\} d\mathbf{x} \leq \frac{1}{p_{\boldsymbol{\theta}}(S)} \underbrace{\int p_{\boldsymbol{\theta}}(\mathbf{x}) f_\xi(\mathbf{x}) d\mathbf{x}}_{=\mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}[f_\xi(\mathbf{x})]}.$$

Since $\lambda I \preceq \mathbf{R}^* \preceq LI$ by A1, it holds that $\xi^\top \mathbf{R}^* \xi = \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}[f_\xi(\mathbf{x})] \leq L$, thus the following inequalities hold:

$$\xi^\top \mathbf{R} \xi = \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[f'_\xi(\mathbf{x})] \leq \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}}}[f_\xi(\mathbf{x})] \leq \frac{1}{p_{\boldsymbol{\theta}}(S)} L.$$

□

A.4 Proof of Claim 2

We will prove a general claim which should take care of both claims in Lemmas 3.2 and 3.3.

Claim 4. Let $\mathbf{x} \sim \rho$ be a random vector with mean $\boldsymbol{\mu}$. Let $\mathbf{b}$ be another vector such that $\mathbf{b} \neq \boldsymbol{\mu}$. Then

$$\mathbf{Cov}_{\mathbf{x}\sim\rho}[\mathbf{x}, \mathbf{x}] = \mathbb{E}_{\mathbf{x}\sim\rho}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_{\mathbf{x}\sim\rho}[(\mathbf{x} - \mathbf{b})(\mathbf{x} - \mathbf{b})^\top] - (\mathbf{b} - \boldsymbol{\mu})(\mathbf{b} - \boldsymbol{\mu})^\top.$$

Proof.

$$\begin{aligned}
&\mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] \\
&= \mathbb{E}[(\mathbf{x} - \mathbf{b} + \mathbf{b} - \boldsymbol{\mu})(\mathbf{x} - \mathbf{b} + \mathbf{b} - \boldsymbol{\mu})^\top] \\
&= \mathbb{E}[(\mathbf{x} - \mathbf{b})(\mathbf{x} - \mathbf{b})^\top] + \underbrace{\mathbb{E}[(\mathbf{x} - \mathbf{b})(\mathbf{b} - \boldsymbol{\mu})^\top]}_{=(-1)\cdot\mathbb{E}[(\mathbf{b}-\boldsymbol{\mu})(\mathbf{b}-\boldsymbol{\mu})^\top]} + \underbrace{\mathbb{E}[(\mathbf{b} - \boldsymbol{\mu})(\mathbf{x} - \mathbf{b})^\top]}_{=(-1)\cdot\mathbb{E}[(\mathbf{b}-\boldsymbol{\mu})(\mathbf{b}-\boldsymbol{\mu})^\top]} + \underbrace{\mathbb{E}[(\mathbf{b} - \boldsymbol{\mu})(\mathbf{b} - \boldsymbol{\mu})^\top]}_{=\mathbb{E}[(\mathbf{b}-\boldsymbol{\mu})(\mathbf{b}-\boldsymbol{\mu})^\top]} \\
&= \mathbb{E}[(\mathbf{x} - \mathbf{b})(\mathbf{x} - \mathbf{b})^\top] - \mathbb{E}[(\mathbf{b} - \boldsymbol{\mu})(\mathbf{b} - \boldsymbol{\mu})^\top]
\end{aligned}$$

□

As a corollary, since the second term is a rank-1 matrix (thus PSD), we have that $\mathbb{E}[(\mathbf{x}-\mathbf{b})(\mathbf{x}-\mathbf{b})^\top] \succeq \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]$.

B Examples of Other Distributions which Satisfy Assumptions

Example 1 (Exponential Distribution)**.** The exponential distribution density can be written

$$p_\lambda(x) = \lambda \exp(-\lambda x) = \exp(-\lambda x + \log(\lambda)),$$

defined on $x \in \mathbb{R}^+$ which is a convex set and for $\lambda > 0$. In natural form, it is

$$p_\theta(x) = \exp(\theta x + \log(-\theta))),$$

defined for $\theta < 0$. Note that

- $T(x) = x$ is a polynomial in x.
- This is log-linear in x (so log-concave in x).
- Variance of the sufficient statistic is simply the variance, which is $1/\theta^2 > 0$ for any $\theta < 0$. If we restrict θ in a bounded set, the negative log-likelihood will be strongly convex and smooth in θ.

Example 2 (Weibull Distribution with known shape k)**.** The Weibull distribution with known shape $k > 0$ has density

$$p_\lambda(x) = \exp((k-1)\log x + \left(-\frac{1}{\lambda^k}\right) x^k + \log k - k \log \lambda)$$

defined on $x \in \mathbb{R}^+$ and $\lambda > 0$. We can re-parameterize this in terms of $\theta = -\frac{1}{\lambda^k}$ with $\theta < 0$ as

$$p_\theta(x) = x^{k-1} \exp(\theta \cdot x^k + \log k + \log(-\theta)).$$

Then

- $T(x) = x^k$ is polynomial in x.
- $p_\theta(x)$ is log-concave in x if $k > 1$ (where recall $x \in \mathbb{R}^+$ and $\theta < 0$).
- The variance of the sufficient statistic can also be found by taking the second derivative of $A(\theta) = -\log k - \log(-\theta)$ w.r.t. θ, which is also $1/\theta^2 > 0$.

Example 3 (Continuous Bernoulli)**.** The continuous Bernoulli density [32] can be written

$$p_\lambda(x) = \exp\left(\log \frac{\lambda}{1-\lambda} - \log \frac{1-2\lambda}{(1-\lambda)\log \frac{1-\lambda}{\lambda}}\right)$$

with support $x \in [0,1]$ and $\lambda \in (0,1)$. We can re-parameterize this in terms of $\theta = \log \frac{\lambda}{1-\lambda}$ with $\theta \in [0,\infty)$ so

$$p_\theta(x) = \exp\left(\theta x - \log \frac{e^\theta - 1}{\theta}\right).$$

Then

- $T(x) = x$ is polynomial in x.
- $p_\theta(x)$ is log-linear in x (so log-concave).
- The variance of sufficient statistic is simply the variance again, which is given by

  $$\mathbf{Var}(X) = \begin{cases} 1/12 & \text{if } \lambda = 1/2 \\ \frac{(\lambda-1)\lambda}{(1-2\lambda)^2} + \frac{1}{(2\tanh^{-1}(1-2\lambda))^2} & \text{otherwise} \end{cases}$$

  This is strictly positive and bounded for all values of λ (thus all values of θ).

Example 4 (Continuous Poisson)**.** A continuous version of the Poisson distribution (although there can be others [24]) can be written

$$p_\lambda(x) = \frac{1}{Z(\lambda)} \frac{e^{-\lambda}\lambda^x}{\Gamma(x+1)}$$

with support $x \in [0,\infty)$ and $\lambda \in (0,\infty)$. We can write this with $\theta = \log \lambda$ so

$$p_\theta(x) = \frac{1}{\Gamma(x+1)} \exp(\theta x - A(\theta)).$$

Then

- $T(x) = x$ is polynomial in x.
- $p_\theta(x)$ is log-concave in x for $x \in \mathbb{R}^+$.
- In λ parameters, the mean of this distribution is λ through usual calculations (e.g., similar to those of the Gamma distribution). Note: we can absorb the $e^{-\lambda}$ term into the partition function.

  $$\begin{aligned} \mathbb{E}[X] &= \frac{1}{Z(\lambda)} \int_0^\infty \frac{x\lambda^x}{\Gamma(x+1)} dx \\ &= \frac{1}{Z(\lambda)} \int_0^\infty \frac{x\lambda^x}{x \cdot \Gamma(x)} dx && \Gamma(x+1) = x \cdot \Gamma(x) \\ &= \frac{\lambda}{Z(\lambda)} \int_1^\infty \frac{\lambda^{x-1}}{\Gamma(x)} dx && \text{Partition function, change var. } z = x - 1 \\ &= \lambda \end{aligned}$$

  Similarly, we should be able to show the variance is λ as usual. In θ space, this means the variance is $\exp(\theta)$ for $\theta \in \mathbb{R}$ which is always positive. Again, we can make it bounded by restricting θ to some set.

Example 5 (Multivariate Gaussian)**.** The multivariate Gaussian also satisfies all of these properties. Recall that the sufficient statistics of the multivariate Gaussian has

- $T(\mathbf{x}) = [\mathbf{x}, \mathbf{x}\mathbf{x}^\top]$ is a polynomial in the components of $\mathbf{x}$ with degree at most 2 (where the $\mathbf{x}\mathbf{x}^\top$ term can be thought of as the vector after standard vectorization).
- The multivariate Gaussian density is strongly log-concave.

- The covariance matrix (of the sufficient statistics) has a complicated form, which the authors of [10] have analyzed the lower bound for, e.g., in their Claims 1 and 2. As before, we can restrict our parameter space to ensure upper bounds.

Example 6 (Generalized Linear Models)**.** This example is the same as the one given in [26] for generalized linear models. It is restated here for completeness.

Consider when we have some covariance, response pair (X, Y) drawn from some distribution D. Suppose that we have a family of distributions $P(\cdot \mid \theta; X)$ such that, for each X, it is an exponential family with sufficient statistic $t_{y,X}$

$$P(y \mid \theta; X) = h(y) \exp\left(\langle \theta, t_{y,X} \rangle - A(\theta, X)\right).$$

We can consider a one-dimensional exponential family q_ν with parameterization $\nu = \langle \theta, X \rangle$, then

$$P(y \mid \theta; X) = h(y) \exp\left(y \langle \theta, X \rangle - \log Z(\langle \theta, X \rangle)\right)$$

where we see that $t_{y,X} = yX$ and the log partition function $A(\theta, X) = \log Z(\langle \theta, X \rangle)$. When q_ν is Bernoulli family or unit variance Gaussian family, this corresponds to *logistic regression* or *least squares regression*, respectively.

We can appropriately generalize this to beyond linear models (e.g., polynomials) provided that we can keep the distribution log-concave.

Comment on A3. We mentioned in the main paper that this assumption combined with log-concavity provides the anti-concentration property that we need for Lemma 3.2. We assume it for simplicity of exposition, but it should be noted that as long as we have the type of anti-concentration property to control how much the covariance can shrink under truncation, we do not necessarily need $T(\mathbf{x})$ to be polynomial. However, we've provided examples of exponential families which already satisfy this above (and there are potentially more which can be addressed by this framework that do not have polynomial sufficient statistics but nonetheless exhibit similar anti-concentration properties).

C Proofs Relating Truncated and Non-Truncated Quantities

C.1 General Truncated Densities

Let ρ be a probability distribution on $\mathbb{R}^d$. Let $S \subseteq \mathbb{R}^d$ be such that $\rho(S) = \alpha$ for some $\alpha \in (0, 1]$. Let $\rho^S := \rho(\cdot \mid \cdot \in S)$ be the conditional distribution of $\mathbf{x} \sim \rho$ given that $\mathbf{x} \in S$.

$$\rho^S(\mathbf{x}) = \frac{\rho(\mathbf{x}) \cdot \mathbb{1}\{\mathbf{x} \in S\}}{\rho(S)}.$$

Note that the relative density is

$$\frac{\rho^S(\mathbf{x})}{\rho(\mathbf{x})} = \frac{\mathbf{1}\{\mathbf{x} \in S\}}{\rho(S)}.$$

Then we can compute that the Rényi divergence is a constant for any order $1 \leq q \leq \infty$.

$$\begin{aligned}
\mathsf{KL}(\rho^S \| \rho) &= \mathbb{E}_{\rho^S}\left[\log \frac{\rho^S}{\rho}\right] = \mathbb{E}_{\rho^S}\left[\log \frac{1}{\rho(S)}\right] = \log \frac{1}{\alpha}. \\
\chi^2(\rho^S \| \rho) &= \mathbb{E}_{\rho^S}\left[\frac{\rho^S}{\rho}\right] - 1 = \frac{1}{\rho(S)} - 1 = \frac{1}{\alpha} - 1. \\
\mathsf{R}_q(\rho^S \| \rho) &= \frac{1}{q-1} \log \mathbb{E}_{\rho^S}\left[\left(\frac{\rho^S}{\rho}\right)^{q-1}\right] = \frac{1}{q-1} \log \frac{1}{\rho(S)^{q-1}} = \log \frac{1}{\rho(S)} = \log \frac{1}{\alpha}. \\
\mathsf{R}_\infty(\rho^S \| \rho) &= \log \sup_x \frac{\rho^S(x)}{\rho(x)} = \log \frac{1}{\rho(S)} = \log \frac{1}{\alpha}.
\end{aligned}$$

Note $\mathsf{R}_2(\rho^S \| \rho) = \log(1 + \chi^2(\rho^S \| \rho))$.

We recall the following general estimates.

Lemma C.1. For any probability distributions ρ, π (such that the quantities below are finite):

1. $\|\mathbb{E}_\rho[\mathbf{x}] - \mathbb{E}_\pi[\mathbf{x}]\| \leq \sqrt{\chi^2(\rho\|\pi)} \cdot \sqrt{\mathbf{Var}_\pi(\mathbf{x})}$.
2. $|\mathbb{E}_\rho[\|\mathbf{x}\|^2] - \mathbb{E}_\pi[\|\mathbf{x}\|^2]| \leq \sqrt{\chi^2(\rho\|\pi)} \cdot \sqrt{\mathbb{E}_\pi[\|\mathbf{x}\|^4]}$.
3. $|\mathbf{Var}_\rho(\mathbf{x}) - \mathbf{Var}_\pi(\mathbf{x})| \leq \sqrt{(\chi^2(\rho\|\pi)+1)^2 - 1} \cdot \sqrt{2\mathbb{E}_\pi[\|\mathbf{x} - \mathbb{E}_\pi[\mathbf{x}]\|^4]}$.

Proof. The first two claims are immediate by Cauchy-Schwarz. For the third one, recall we can write

$$\mathbf{Var}_\rho(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{\rho^{\otimes 2}}[\|\mathbf{x} - \mathbf{y}\|^2].$$

Then by applying part (1) to $\rho^{\otimes 2}$ and $(\pi)^{\otimes 2}$, we get

$$\begin{aligned}
|\mathbf{Var}_\rho(\mathbf{x}) - \mathbf{Var}_\pi(\mathbf{x})| &\leq \frac{1}{2}\sqrt{\chi^2(\rho^{\otimes 2}\|\pi^{\otimes 2})} \cdot \sqrt{\mathbb{E}_{\pi^{\otimes 2}}[\|\mathbf{x} - \mathbf{y}\|^4]} \\
&= \frac{1}{2}\sqrt{(\chi^2(\rho\|\pi)+1)^2 - 1} \cdot \sqrt{2\mathbb{E}_\pi[\|\mathbf{x} - \mathbb{E}_\pi[\mathbf{x}]\|^4] + 6\mathbb{E}_\pi[\|\mathbf{x} - \mathbb{E}_\pi[\mathbf{x}]\|^2]^2} \\
&\leq \frac{1}{2}\sqrt{(\chi^2(\rho\|\pi)+1)^2 - 1} \cdot \sqrt{8\mathbb{E}_\pi[\|\mathbf{x} - \mathbb{E}_\pi[\mathbf{x}]\|^4]}.
\end{aligned}$$

□

For our application, we have the following. Given a probability distribution ρ on $\mathbb{R}^d$, we let $\mu(\rho) = \mathbb{E}_\rho[\mathbf{x}]$ be its mean, and for $k \in \mathbb{N}$,

$$M_k(\rho) := \mathbb{E}_\rho[\|\mathbf{x} - \mu(\rho)\|^k]^{1/k}.$$

So for example we have $M_2(\rho) = \sqrt{\mathbf{Var}_\rho(\mathbf{x})}$. We also have $M_k(\rho) \leq M_\ell(\rho)$ if $k \leq \ell$.

Lemma C.2. Let ρ be a probability distribution on $\mathbb{R}^d$. Let $S \subseteq \mathbb{R}^d$ with $\rho(S) = \alpha \in (0,1]$. Then

1. $\|\mathbb{E}_{\rho^S}[\mathbf{x}] - \mathbb{E}_\rho[\mathbf{x}]\| \leq \sqrt{\frac{1-\alpha}{\alpha}} \cdot \sqrt{\mathbf{Var}_\rho(\mathbf{x})}$.
2. $|\mathbf{Var}_{\rho^S}(\mathbf{x}) - \mathbf{Var}_\rho(\mathbf{x})| \leq \frac{\sqrt{2(1-\alpha^2)}}{\alpha} M_4(\rho)^2$.

In particular, if $\alpha \in (0,1]$ is such that $\frac{1}{\alpha^2} \leq 1 + \frac{c^2 M_2(\rho)^4}{2M_4(\rho)^4}$ for some $0 \leq c < 1$, then

$$\mathbf{Var}_{\rho^S}(\mathbf{x}) \geq (1-c)\mathbf{Var}_\rho(\mathbf{x}).$$

Note that the constraint on α above implies $\frac{1}{\alpha^2} \leq \frac{3}{2}$, so $\alpha \geq \sqrt{2/3}$. But if $M_2(\rho) \ll M_4(\rho)$, then $1-\alpha$ will be very small.

Recall also that under some conditions, e.g. if ρ is log-concave, then we have the reverse bound that

$$M_2(\rho) \geq C_{2,4} M_4(\rho)$$

for a universal constant $C_{2,4}$, so the constraint above is not too restrictive, as it allows $1-\alpha$ of constant size.

C.2 Exponential Families with Strongly Convex and Smooth Log-Partition Functions are Sub-Exponential

Let $\boldsymbol{\theta} \in \Theta$ such that $\boldsymbol{\theta} + \frac{1}{\beta}\mathbf{u} \in \Theta$ for some $\beta > 0$ for all unit vectors $\mathbf{u}$ and such that Assumption A1 holds for $p_{\boldsymbol{\theta}}$. Then $X := \mathbf{u}^\top(T(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}[T(\mathbf{x})])$ is $SE(L, \beta)$.

Proof. WLOG, consider $p_{\boldsymbol{\theta}}$ in the transformed space $\mathbf{x} \mapsto T(\mathbf{x})$ so that

$$p_{\boldsymbol{\theta}}(\mathbf{t}) = h(\mathbf{t})\exp(\boldsymbol{\theta}^\top \mathbf{t} - A(\boldsymbol{\theta}))d\mathbf{t},$$

where $\boldsymbol{\theta} \in \Theta$ and $A(\boldsymbol{\theta}) = \log(Z(\boldsymbol{\theta})) = \log\left(\int_{\mathcal{T}(\mathcal{X})} h(\mathbf{t})\exp(\boldsymbol{\theta}^\top \mathbf{t})d\mathbf{t}\right)$ is the log-partition function. Note that $\nabla^2 A(\boldsymbol{\theta}) = \mathbf{Cov}_{\mathbf{t}\sim p_{\boldsymbol{\theta}}(\mathbf{t})}[\mathbf{t}] = \mathbf{Cov}_{\mathbf{x}\sim p_{\theta}(\mathbf{x})}[T(\mathbf{x})]$, and by A1, $A(\boldsymbol{\theta})$ is a λ-strongly convex and L-smooth function in $\boldsymbol{\theta}$.

To show that $p_{\boldsymbol{\theta}}(\mathbf{t})$ is sub-exponential with parameters (ν^2, β) we need to show that its moment generating function satisfies $\mathbb{E}[e^{\gamma \mathbf{u}^\top(\mathbf{t}-\boldsymbol{\mu})}] \le e^{\gamma^2\nu^2/2}$, where $\boldsymbol{\mu} = \mathbb{E}_{p_{\boldsymbol{\theta}}}[\mathbf{t}]$, $\mathbf{u}$ is a unit vector, for $|\gamma| < 1/\beta$.

$$\begin{aligned}
\mathbb{E}[e^{\gamma \mathbf{u}^\top(\mathbf{t}-\boldsymbol{\mu})}] &= \int \left(e^{\gamma \mathbf{u}^\top \mathbf{t} - \gamma \mathbf{u}^\top \boldsymbol{\mu}}\right) h(\mathbf{t}) e^{\boldsymbol{\theta}^\top \mathbf{t} - A(\boldsymbol{\theta})} d\mathbf{t} \\
&= \frac{\exp(-\gamma \mathbf{u}^\top \boldsymbol{\mu})}{Z(\boldsymbol{\theta})} \int h(\mathbf{t}) \exp((\gamma \mathbf{u} + \boldsymbol{\theta})^\top \mathbf{t}) d\mathbf{t} \\
&= \frac{Z(\gamma \mathbf{u} + \boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \cdot \exp(-\gamma \mathbf{u}^\top \boldsymbol{\mu})
\end{aligned}$$

The inequality we need to show is equivalent to proving

$$\begin{aligned}
&\mathbb{E}[e^{\gamma \mathbf{u}^\top(\mathbf{t}-\boldsymbol{\mu})}] \le e^{\gamma^2\nu^2/2} \\
\iff\ & \frac{Z(\gamma \mathbf{u} + \boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \cdot e^{-\gamma \mathbf{u}^\top \boldsymbol{\mu}} \le e^{\gamma^2\nu^2/2} \\
\iff\ & \frac{Z(\gamma \mathbf{u} + \boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \le e^{\gamma \mathbf{u}^\top \boldsymbol{\mu}} \cdot e^{\gamma^2\nu^2/2} \\
\iff\ & A(\gamma \mathbf{u} + \boldsymbol{\theta}) - A(\boldsymbol{\theta}) \le \gamma \mathbf{u}^\top \boldsymbol{\mu} + \frac{\gamma^2\nu^2}{2}
\end{aligned}$$

Since $A(\boldsymbol{\theta})$ is L-smooth, we have that

$$A(\gamma \mathbf{u} + \boldsymbol{\theta}) - A(\boldsymbol{\theta}) \le \langle \underbrace{\nabla A(\boldsymbol{\theta})}_{=\boldsymbol{\mu}}, \gamma \mathbf{u} \rangle + \frac{L}{2}\|\gamma \mathbf{u}\|^2 = \gamma \mathbf{u}^\top \boldsymbol{\mu} + \frac{\gamma^2 L}{2}$$

where we've used the property of exponential families that the gradient of the log partition function is the mean sufficient statistic. Now we can see that the appropriate parameter for ν^2 is L and γ must be small enough so that $\gamma u + \theta \in \Theta$, i.e., $|\gamma| < \frac{1}{\beta}$ for some $\beta > 0$. This is possible if θ is bounded away from the boundary of Θ. □

Remark. In the above, we only needed to use that $p_{\boldsymbol{\theta}}$ is an exponential family distribution and that its log-partition function $A(\boldsymbol{\theta})$ is smooth. It is also possible to show that $p_{\boldsymbol{\theta}}$ has exponentially decreasing tails (in quantities involving $\mathbf{x}$ rather than $T(\mathbf{x})$) if it is log-concave in $\mathbf{x}$ (assumption A2), e.g., by [40].

C.3 Proof of Lemma 3.4

Let $p_{\boldsymbol{\theta}}(\mathbf{x}) = h(\mathbf{x})\exp(\langle \theta, T(\mathbf{x})\rangle - A(\boldsymbol{\theta}))$ and $A\colon \Theta \to \mathbb{R}$ is the log-partition function:

$$A(\boldsymbol{\theta}) = \int_{\mathcal{X}} h(\mathbf{x}) \exp(\langle \theta, T(\mathbf{x})\rangle) d\mathbf{x}.$$

Lemma C.3. For any $q > 1$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$:

$$\mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\left(\frac{p_{\boldsymbol{\theta}'}}{p_{\boldsymbol{\theta}}}\right)^q\right] = \exp\left((q-1)A(\boldsymbol{\theta}) - qA(\boldsymbol{\theta}') + A\left(q\boldsymbol{\theta}' - (q-1)\boldsymbol{\theta}\right)\right).$$

Proof.

$$\begin{aligned}
\mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\left(\frac{p_{\boldsymbol{\theta}'}}{p_{\boldsymbol{\theta}}}\right)^q\right] &= \int_{\mathcal{X}} h(x)\exp\left(\langle \boldsymbol{\theta}, T(x)\rangle - A(\boldsymbol{\theta})\right) \cdot \exp\left(q\langle \boldsymbol{\theta}' - \boldsymbol{\theta}, T(x)\rangle - qA(\boldsymbol{\theta}') + qA(\boldsymbol{\theta})\right) dx \\
&= \exp\left((q-1)A(\boldsymbol{\theta}) - qA(\boldsymbol{\theta}') + A\left(q\boldsymbol{\theta}' - (q-1)\boldsymbol{\theta}\right)\right).
\end{aligned}$$

□

Lemma C.4. Assume A is convex and L-smooth on Θ. For any $S \subseteq \mathcal{X}$, and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$:

$$p_{\boldsymbol{\theta}}(S) \geq p_{\boldsymbol{\theta}'}(S)^2 \cdot \exp\left(-\frac{3L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2\right).$$

Proof. By Cauchy-Schwarz,

$$\begin{aligned} p_{\boldsymbol{\theta}'}(S)^2 &= \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\frac{p_{\boldsymbol{\theta}'}}{p_{\boldsymbol{\theta}}}\mathbf{1}_S\right]^2 \\ &\leq p_{\boldsymbol{\theta}}(S) \cdot \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\left(\frac{p_{\boldsymbol{\theta}'}}{p_{\boldsymbol{\theta}}}\right)^2\right] \\ &= p_{\boldsymbol{\theta}}(S) \cdot \exp\left(A(\boldsymbol{\theta}) - 2A(\boldsymbol{\theta}') + A\left(2\boldsymbol{\theta}' - \boldsymbol{\theta}\right)\right). \end{aligned}$$

Since A is convex and L-smooth,

$$A(\boldsymbol{\theta}) \leq A(\boldsymbol{\theta}') + \langle \nabla A(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle$$

$$A(2\boldsymbol{\theta}' - \boldsymbol{\theta}) \leq A(\boldsymbol{\theta}') + \langle \nabla A(\boldsymbol{\theta}'), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2$$

Therefore,

$$\begin{aligned} A(\boldsymbol{\theta}) - 2A(\boldsymbol{\theta}') + A\left(2\boldsymbol{\theta}' - \boldsymbol{\theta}\right) &\leq \langle \nabla A(\boldsymbol{\theta}') - \nabla A(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \\ &\leq \frac{3L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2. \end{aligned}$$

□

Compare this to the Gaussian case (e.g., see H.8 of [38]) where this was $p_{\boldsymbol{\theta}}(S) \geq \frac{\alpha}{2}\exp\left(-r \cdot \sqrt{2\log 1/\alpha} - \frac{1}{2}r^2\right)$ for $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < r$.

C.4 Proof of Lemma 3.7

Let $\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T(\mathbf{x}_i)$ be the empirical mean sufficient statistics given our samples $\{\mathbf{x}_i\}_{i=1}^{n}$ each $\mathbf{x}_i \sim p_{\boldsymbol{\theta}^*}^S$.

Let $\epsilon_S > 0$. For $n \geq \Omega\left(\frac{2\beta}{\epsilon_S}\log\left(\frac{1}{\delta}\right)\right)$,

$$\|\overline{T} - \mathbb{E}_{p_{\boldsymbol{\theta}^*}}[T(\mathbf{x})]\| \leq \epsilon_S + \mathcal{O}(\log 1/\alpha)$$

with probability at least $1 - \delta$.

Proof. Let $\boldsymbol{\nu}^* = \mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}^*}}[T(\mathbf{x})]$.

For any event A, we have that

$$\mathbb{P}_{p_{\boldsymbol{\theta}^*}^S}[A] = \int \mathbb{1}\{\omega \in A\} dp_{\boldsymbol{\theta}^*}^S(\omega) = \frac{1}{\alpha}\int \mathbb{1}\{\omega \in A\}\mathbb{1}\{\omega \in S\} dp_{\boldsymbol{\theta}^*}(\omega) \leq \frac{1}{\alpha}\mathbb{P}_{p_{\boldsymbol{\theta}^*}}[A]$$

and for the product measure with n independent components $\mathbb{P}_{\Pi_{i \in [n]} p_{\boldsymbol{\theta}^*}^S}[A] \leq \left(\frac{1}{\alpha}\right)^n \mathbb{P}_{\Pi_{i \in [n]} p_{\boldsymbol{\theta}^*}}[A]$. So we can bound the probability of events on $p_{\boldsymbol{\theta}^*}^S$ with those on $p_{\boldsymbol{\theta}^*}$. In particular, by Claim 1 and by the composition property of independent sub-exponential random variables, we have that

$$\begin{aligned} &\mathbb{P}_{p_{\boldsymbol{\theta}^*}}\left(\frac{1}{n}\left|\mathbf{u}^\top\left(\sum_i T(\mathbf{x}_i) - \boldsymbol{\nu}^*\right)\right| \geq t\right) \leq \exp\left(-\frac{nt}{2\beta}\right) \qquad \text{for any unit vector } \mathbf{u} \\ \Rightarrow &\mathbb{P}_{p_{\boldsymbol{\theta}^*}}\left(\left\|\frac{1}{n}\sum_i T(\mathbf{x}_i) - \boldsymbol{\nu}^*\right\| \geq t\right) \leq \exp\left(-\frac{nt}{2\beta}\right). \end{aligned}$$

To translate this to the probability of the same event on $p^S_{\boldsymbol{\theta}^*}$, note that

$$\left(\frac{1}{\alpha}\right)^n \exp\left(-\frac{nt}{2\beta}\right) \leq \delta \iff \exp\left(n \cdot \left(\log 1/\alpha - \frac{t}{2\beta}\right)\right) \leq \delta$$

which holds when $t = 2\beta\left(\log 1/\alpha + \frac{1}{n}\log 1/\delta\right)$. Thus for $n > \frac{2\beta}{\epsilon_S}\log 1/\delta$ samples from the truncated $p^S_{\boldsymbol{\theta}^*}$ we have that with probability at least $1-\delta$, the quantity $\|\overline{T} - \boldsymbol{\nu}^*\| \leq 2\beta(\log 1/\alpha) + \epsilon_S$.

□

D Additional Proofs for Algorithm Analysis

Algorithm 3 Stochastic Gradient Descent

```
Initialize some $\boldsymbol{\theta}_0 \in K$.
for iteration $t = 1, 2, \ldots, T$ do
    Compute $\mathbf{v}_t$ such that $\mathbb{E}[\mathbf{v}_t \mid \boldsymbol{\theta}_t] = \nabla f(\boldsymbol{\theta}_t)$
    $\widetilde{\boldsymbol{\theta}}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta\mathbf{v}_t$
    $\widetilde{\boldsymbol{\theta}}_{t+1} = \Pi_K(\widetilde{\boldsymbol{\theta}}_{t+1})$        (Project onto $K$)
end for
Return $\boldsymbol{\theta}_T$
```

D.1 SGD Algorithm and its Analysis

Although the setting of Theorem 5.7 of [21] is when the objective is a sum of many functions, the proof and its result can be easily adapted to our setting.

Theorem. Let f be a λ-strongly convex function. Let $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in K} f(\boldsymbol{\theta})$. Consider the sequence $\{\boldsymbol{\theta}_t\}_{t=1}^N$ generated by SGD (Algorithm 3) and $\{\mathbf{v}_t\}_{t=1}^N$ the sequence random vectors satisfying $\mathbb{E}[\mathbf{v}_t \mid \boldsymbol{\theta}_t] = \nabla f(\boldsymbol{\theta}_t)$ and $\mathbb{E}[\|\mathbf{v}_t\|^2 \mid \boldsymbol{\theta}_t] < \rho^2$ for all t, with a constant step size η satisfying $0 < \eta < \frac{1}{\lambda}$. It follows that for $t \geq 0$,

$$\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 \leq (1-2\eta\lambda)^t\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \frac{\eta}{\lambda}\rho^2.$$

Proof. At any iteration i,

$$\begin{aligned}\widetilde{\boldsymbol{\theta}}_{i+1} &= \boldsymbol{\theta}_i - \eta\mathbf{v}_i \\ \widetilde{\boldsymbol{\theta}}_{i+1} - \boldsymbol{\theta}^* &= \boldsymbol{\theta}_i - \boldsymbol{\theta}^* - \eta\mathbf{v}_i \qquad (1) \\ \|\widetilde{\boldsymbol{\theta}}_{i+1} - \boldsymbol{\theta}^*\|^2 &= \|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|^2 - 2\eta\langle\mathbf{v}_i, \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\rangle + \eta^2\|\mathbf{v}_i\|^2\end{aligned}$$

where the last line comes from multiplying the line (1) with the transpose of the same equation on either side. After projecting to the set K to obtain $\boldsymbol{\theta}_{i+1} = \arg\min_{\boldsymbol{\theta}\in K}\|\widetilde{\boldsymbol{\theta}}_{i+1} - \boldsymbol{\theta}\|^2$ and given that $\boldsymbol{\theta}^* \in K$, we have that $\|\widetilde{\boldsymbol{\theta}}_{i+1} - \boldsymbol{\theta}^*\|^2 \geq \|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}^*\|^2$, so

$$\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}^*\|^2 \leq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|^2 - 2\eta\langle\mathbf{v}_i, \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\rangle + \eta^2\|\mathbf{v}_i\|^2 \qquad (2)$$

Fixing $\boldsymbol{\theta}_i$ in the i^{th} iteration and taking the conditional expectation in (2) gives

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}^*\|^2 \mid \boldsymbol{\theta}_i] &\leq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|^2 - 2\eta\langle\nabla f(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\rangle + \eta^2\mathbb{E}[\|\mathbf{v}_t\|^2 \mid \boldsymbol{\theta}_i] \\ &\leq (1-2\eta\lambda)\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|^2 + \eta^2\mathbb{E}[\|\mathbf{v}_t\|^2 \mid \boldsymbol{\theta}_i]\end{aligned}$$

where the last line is due to strong convexity, $\langle\nabla f(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\rangle \geq \lambda\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|^2$. By taking iterated expectations and recursively applying the above, we get that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2] &\leq (1-2\eta\lambda)^T\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \eta^2\rho^2\sum_{i=0}^{T-1}(1-2\eta\lambda)^i \\ &\leq (1-2\eta\lambda)^T\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \eta^2\frac{1}{\eta\lambda}\rho^2 \\ &= (1-2\eta\lambda)^T\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \frac{\eta}{\lambda}\rho^2\end{aligned}$$

where in the second line we used that $\sum_{i=0}^{T-1}(1-2\eta\lambda)^i = \frac{1-(1-2\eta\lambda)^T}{1-(1-2\eta\lambda)} < \frac{2}{2\eta\lambda}$ provided $\eta < 1/\lambda$. □

We can derive the complexity (number of iterations) to get $\mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}\|^2] < \epsilon$ using the following Lemma from [21].

Lemma D.1 (Lemma A.2 of [21])**.** Consider the recurrence given by

$$\alpha_k \leq (1-\eta\mu)^t \alpha_0 + A\eta,$$

where $\mu > 0$, and $A, C \geq 0$ are given constants and $\eta < 1/C$. If

$$\eta = \min\left\{\frac{\epsilon}{2A}, \frac{1}{C}\right\}$$

then

$$t \geq \max\left\{\frac{1}{\epsilon}\frac{2A}{\mu}, \frac{C}{\mu}\right\} \log\left(\frac{2\alpha_0}{\epsilon}\right) \Rightarrow \alpha_k \leq \epsilon.$$

Note that to get bounds on $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|$ rather than $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2$, we can solve the number of iterations we need to get ϵ^2 on the right hand side, and we will get the number of iterations for $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\| < \epsilon$. Then the resulting complexity bounds will replace $1/\epsilon$ with $1/\epsilon^2$.

D.2 Approximate Sampling of Non-Truncated Distribution

Many analyses of stochastic gradient descent assume unbiased directions at every iteration of the algorithm, but since we need to be able to sample from $p_{\boldsymbol{\theta}_i}$ multiple times at each iteration i until we get a sample in S, our directions are only unbiased if we can indeed sample exactly from $p_{\boldsymbol{\theta}_i}$ each time for all i.

However if $p_{\boldsymbol{\theta}_i}$ is complicated, exact sampling can be difficult or take too long. Since we assume that $p_{\boldsymbol{\theta}}$ is log-concave in $\mathbf{x}$ for all $\boldsymbol{\theta} \in \Theta$, we can at least approximately sample from it efficiently via Langevin Monte Carlo, MALA, or other algorithms with convergence guarantees for log-concave densities.

Lemma D.2 (SGD Analysis with Biased Directions)**.** Let f be a λ-strongly convex function. Let $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in K \subseteq B(\boldsymbol{\theta}^*, D)} f(\boldsymbol{\theta})$. Consider the sequence $\{\boldsymbol{\theta}_t\}_{t=1}^N$ generated by SGD but with $\{\widetilde{\mathbf{v}}_t\}_{t=1}^N$ the sequence random vectors satisfying $\mathbb{E}[\widetilde{\mathbf{v}}_t \mid \boldsymbol{\theta}_t] = \mathbf{b}_t - \nabla f(\boldsymbol{\theta}_t)$ with $\|\mathbf{b}_t\| < B$ and $\mathbb{E}[\|\widetilde{\mathbf{v}}_t\|^2 \mid \boldsymbol{\theta}_t] < \rho'^2$ for all t, with a constant step size η satisfying $0 < \eta < \frac{1}{\lambda}$. It follows that for $t \geq 0$,

$$\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 \leq (1-2\eta\lambda)^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \frac{\eta}{\lambda}\rho^2 + \frac{2BD}{\lambda}.$$

Proof. Define $\mathbf{b}_t \coloneqq \mathbb{E}[\mathbf{b}_t \mid \boldsymbol{\theta}_t] - \nabla f(\boldsymbol{\theta}_t)$ the bias for each $t \geq 0$. The analysis of Theorem 3.10 can be applied generically to get Eq. (2):

$$\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}^*\|^2 \leq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|^2 - 2\eta\langle \widetilde{\mathbf{v}}_i, \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\rangle + \eta^2\|\widetilde{\mathbf{v}}_i\|^2$$

Now when taking the conditional expectation, we get

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}^*\|^2 \mid \boldsymbol{\theta}_t] &\leq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|^2 - 2\eta\langle \mathbf{b}_t, \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\rangle - 2\eta\langle \nabla f(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\rangle - \eta^2\mathbb{E}[\|\widetilde{\mathbf{v}}_i\|^2 \mid \boldsymbol{\theta}_i] \\ &\leq (1-2\eta\lambda)\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|^2 + \eta^2\rho'^2 + 2\eta\|\mathbf{b}_i\|\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|\end{aligned}$$

Taking iterated expectations and recursively applying this gives now

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2] &\leq (1-2\eta\lambda)^T\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \sum_{i=0}^{T-1}(1-2\eta\lambda)^i \cdot \left(\eta^2\rho'^2 + 2\eta\|\mathbf{b}_i\|\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\|\right) \\ &\leq (1-2\eta\lambda)^T\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \frac{\eta^2\rho'^2 + 2\eta BD}{\eta\lambda} \\ &= (1-2\eta\lambda)^T\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \frac{\eta\rho'^2}{\lambda} + \frac{2BD}{\lambda}\end{aligned}$$

where the second line holds if $\|\mathbf{b}_i\| \leq B, \|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\| < D, \forall i$ (which holds under the assumptions). □

Note that in our Algorithm 1, D here is simply $\frac{2}{\lambda}d(\alpha)$ by construction. We can also control B through the following.

Bounding the bias. Fix some t. Let $\widetilde{\mathbf{v}}_t \coloneqq T(\mathbf{z}) - T(\mathbf{x})$ where $\mathbf{z} \sim \widetilde{p}^S_{\boldsymbol{\theta}_i}$ and $\mathbf{x} \sim p^S_{\boldsymbol{\theta}^*}$. Then we can write

$$\begin{aligned}
\mathbf{b}_t &= \mathbb{E}_{\mathbf{z}\sim\widetilde{p}^S_{\boldsymbol{\theta}_i}}[T(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}^*}}[T(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p^S_{\boldsymbol{\theta}_i}}[T(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p^S_{\boldsymbol{\theta}^*}}[T(\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{z}\sim\widetilde{p}^S_{\boldsymbol{\theta}_i}}[T(\mathbf{z})] - \mathbb{E}_{\mathbf{z}\sim p^S_{\boldsymbol{\theta}_i}}[T(\mathbf{z})] \\
&= \int_{\mathcal{X}} T(\mathbf{x}) \cdot (\widetilde{p}^S_{\boldsymbol{\theta}_i}(\mathbf{x}) - p^S_{\boldsymbol{\theta}_i}(\mathbf{x}))d\mathbf{x} \qquad (3)
\end{aligned}$$

If we know that $T(\mathbf{x})$ is bounded over S, we can upper-bound this given the TV distance between $\widetilde{p}^S_{\boldsymbol{\theta}_i}$ and $p^S_{\boldsymbol{\theta}_i}$:

$$\|\mathbf{b}_t\| \leq \sup_{\mathbf{x}\in S} \|T(\mathbf{x})\| \left\|\widetilde{p}^S_{\boldsymbol{\theta}_i} - p^S_{\boldsymbol{\theta}_i}\right\|_{TV}.$$

Since we assume that $\mathbb{E}_{p_{\boldsymbol{\theta}_i}}[T(\mathbf{x})]$ is finite, it should be the case that $T(\mathbf{x})$ is bounded over its support except potentially on some negligible sets. In that case, we can replace $T(\mathbf{x})$ with $\widetilde{T}(\mathbf{x})$ which replaces those potentially infinite values on negligible sets with 0 and the integral expression in (3) would be equal to one which uses $\widetilde{T}(\mathbf{x})$ instead of $T(\mathbf{x})$, and the bound on its norm holds given that $\widetilde{T}(\mathbf{x})$ is bounded.

Otherwise we can use bounds from Lemma C.1 to bound this as

$$\|\mathbf{b}_t\| \leq \sqrt{\chi^2(\widetilde{p}^S_{\boldsymbol{\theta}_i} || p^S_{\boldsymbol{\theta}_i})} \cdot \sqrt{\mathbf{Var}_{p^S_{\boldsymbol{\theta}_i}}(T(\mathbf{x}))}$$

if we have control over the chi-square divergence (see Section C.1 for definitions).

If we know that $T(x)$ is a 1-Lipschitz, real-valued function (e.g., when $T(x) = x$), we can use the dual representation of W_1 distance to bound this as

$$\mathbf{b}_t = \int_S T(x) d\widetilde{p}^S_{\boldsymbol{\theta}_i} - \int_S T(x) dp^S_{\boldsymbol{\theta}_i} \leq \sup_{f\in\mathcal{F}_{1\text{Lip}}} \int f(x) d\widetilde{p}^S_{\boldsymbol{\theta}_i} - \int f(x) dp^S_{\boldsymbol{\theta}_i} = W_1(\widetilde{p}^S_{\boldsymbol{\theta}_i}, p^S_{\boldsymbol{\theta}_i}).$$

Proposition D.3 (Bounds on truncated total variation, given bounds on non-truncated)**.** Suppose $\|\widetilde{p}_{\boldsymbol{\theta}_i} - p_{\boldsymbol{\theta}_i}\|_{TV} \leq \epsilon_{TV}$ for some $\epsilon_{TV} > 0$. Then

$$\|\widetilde{p}^S_{\boldsymbol{\theta}_i} - p^S_{\boldsymbol{\theta}_i}\|_{TV} \leq \frac{\epsilon^2_{TV}}{p_{\boldsymbol{\theta}_i}(S) - \epsilon_{TV}}.$$

Proof. First, given that $\|\widetilde{p}_{\boldsymbol{\theta}_i} - p_{\boldsymbol{\theta}_i}\|_{TV} \leq \epsilon_{TV}$, we have by one characterization of the total variation distance (the supremum of the difference in mass over all measurable sets)

$$\widetilde{p}_{\boldsymbol{\theta}_i}(S) \geq p_{\boldsymbol{\theta}_i}(S) - \epsilon_{TV}.$$

Now for the truncated densities,

$$\begin{aligned}
\|\widetilde{p}^S_{\boldsymbol{\theta}_i} - p^S_{\boldsymbol{\theta}_i}\|_{TV} &= \frac{1}{2}\int |\widetilde{p}^S_{\boldsymbol{\theta}_i}(\mathbf{x}) - p^s_{\boldsymbol{\theta}_i}(\mathbf{x})| d\mathbf{x} \\
&= \frac{1}{2}\int \mathbb{1}\{\mathbf{x}\in S\} \cdot \left|\frac{\widetilde{p}_{\boldsymbol{\theta}_i}(\mathbf{x})}{\widetilde{p}_{\boldsymbol{\theta}_i}(S)} - \frac{p_{\boldsymbol{\theta}_i}(\mathbf{x})}{p_{\boldsymbol{\theta}_i}(S)}\right| d\mathbf{x} \\
&\leq \frac{\epsilon_{TV}}{p_{\boldsymbol{\theta}_i}(S) - \epsilon_{TV}} \cdot \frac{1}{2}\int |\widetilde{p}_{\boldsymbol{\theta}_i}(\mathbf{x}) - p_{\boldsymbol{\theta}_i}(\mathbf{x})| d\mathbf{x} \\
&\leq \frac{\epsilon^2_{TV}}{p_{\boldsymbol{\theta}_i}(S) - \epsilon_{TV}}
\end{aligned}$$

□

Efficient, Approximate Sampling. There exist several results in sampling which give bounds in TV distance in polynomial time (mixing time bounds) for log-concave distributions, e.g., [16], [3], [34] but which also usually require that the log density is also smooth (in $\mathbf{x}$, not $\boldsymbol{\theta}$). There are also proofs for Langevin Monte Carlo when the log density is convex and Lipschtiz, not necessarily smooth (e.g., see Chapter 4 of [7]), or under LSI (which is implied by strong log-concavity) with convergence in Renyi divergence (e.g., Chapter 5 of [7]). We can also use the proximal sampler to achieve convergence in KL divergence under log-concavity (e.g., Chapter 8.4 of [7]), which by Pinsker's inequality can bound the TV distance.

Proposition D.4 (Bounded variance step with bias)**.** If $\mathbb{E}[\|\mathbf{v}_i\| \mid \boldsymbol{\theta}_i] \leq \rho^2$ where $\mathbf{v}_i = T(\mathbf{z}) - T(\mathbf{x})$ with $\mathbf{z} \sim p_{p^S_{\boldsymbol{\theta}_i}}$ and $\mathbf{x} \sim p^S_{\boldsymbol{\theta}^*}$,then $\mathbb{E}[\|\widetilde{\mathbf{v}}_i\| \mid \boldsymbol{\theta}_i] \leq \rho'^2$ for $\widetilde{\mathbf{v}}_i = T(\widetilde{\mathbf{z}}) - T(\mathbf{x})$ with $\widetilde{\mathbf{z}} \sim \widetilde{p}_{p^S_{\boldsymbol{\theta}_i}}$ and $\mathbf{x} \sim p^S_{\boldsymbol{\theta}^*}$, where

$$\rho'^2 = \mathbf{Var}_{\widetilde{p}^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) - \mathbf{Var}_{p^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) + \rho^2 + B^2,$$

where $\|\mathbf{b}_i\| = \|\mathbb{E}_{\widetilde{\mathbf{z}} \sim \widetilde{p}^S_{\boldsymbol{\theta}_i}}[T(\widetilde{\mathbf{z}})] - \mathbb{E}_{\mathbf{z} \sim p^S_{\boldsymbol{\theta}_i}}[T(\mathbf{z})]\| < B$.

Proof. As in the proof of the exact sampling version, we can write

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\mathbf{v}}_i\| \mid \boldsymbol{\theta}_i] &= \mathbf{Var}_{\widetilde{p}^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) + \mathbf{Var}_{p^S_{\boldsymbol{\theta}^*}}(T(\mathbf{x})) + \|\mathbb{E}_{\widetilde{p}^S_{\boldsymbol{\theta}_i}}[T(\mathbf{z})] - \mathbb{E}_{p^S_{\boldsymbol{\theta}^*}}[T(\mathbf{x})]\|^2 \\
&\leq \mathbf{Var}_{\widetilde{p}^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) + \mathbf{Var}_{p^S_{\boldsymbol{\theta}^*}}(T(\mathbf{x})) + \|\mathbb{E}_{p^S_{\boldsymbol{\theta}_i}}[T(\mathbf{z})] - \mathbb{E}_{p^S_{\boldsymbol{\theta}^*}}[T(\mathbf{x})]\|^2 + \|\mathbf{b}_i\|^2 \\
&= \mathbf{Var}_{\widetilde{p}^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) - \mathbf{Var}_{p^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) \\
&\qquad + \underbrace{\mathbf{Var}_{p^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) + \mathbf{Var}_{p^S_{\boldsymbol{\theta}^*}}(T(\mathbf{x})) + \|\mathbb{E}_{p^S_{\boldsymbol{\theta}_i}}[T(\mathbf{z})] - \mathbb{E}_{p^S_{\boldsymbol{\theta}^*}}[T(\mathbf{x})]\|^2}_{\leq \rho^2} + \underbrace{\|\mathbf{b}_i\|^2}_{\leq B^2} \\
&\leq \mathbf{Var}_{\widetilde{p}^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) - \mathbf{Var}_{p^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) + \rho^2 + B^2
\end{aligned}$$

□

We can bound the difference $\mathbf{Var}_{\widetilde{p}^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z})) - \mathbf{Var}_{p^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z}))$ if have bounds on $\mathbf{Var}_{\widetilde{p}^S_{\boldsymbol{\theta}_i}}(T(\mathbf{z}))$, through bounds like in Lemma C.1 if we can say something about the chi-square divergence, or through similar arguments to the bias bound if we assume some bounds on $\|T(\mathbf{x})\|^2$ over its support.

E Numerical Example

To illustrate how the algorithm performs in different dimensions, we implemented our algorithm for 2-, 5-, 10-, and 20-dimensional exponential distributions. In all cases, the truncation set is the (hyper-)cube $[0, 2]^d$. We chose true parameters in all cases which resulted in an initial error at most 2.5. In all cases, we use 1500 iterations and step size 0.01, each repeated 10 times. In the end, all have (average) L2 error at most 0.15. For stability (and to bypass repeating the algorithm multiple times as stated in the analysis), we instead calculated gradients using the average of 10 samples which was sufficient to have stable training results. See Figure 2.

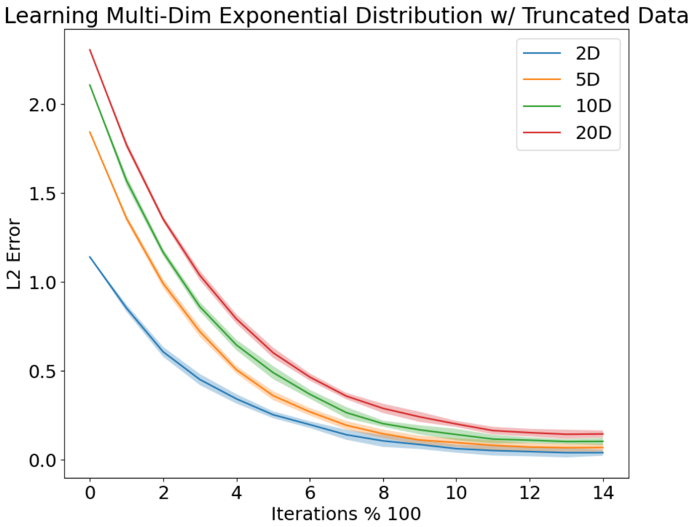

Figure 2: Learning 2-, 5-, 10-, and 20-dimensional truncated exponential distributions. In all cases, the truncation set is the (hyper-)cube $[0, 2]^d$.

The wall clock time to finish all 1500 iterations of training for 5-, 10-, and 20-dimensions was 42.9 ± 2.2, 49.1 ± 6.5, and 61.2 ± 3.0 seconds, respectively. We can see that the running time is not doubling with the doubling of dimensions.