Moment-based Uniform Deviation Bounds for k-means and Friends

Matus Telgarsky **Sanjoy Dasgupta**
Computer Science and Engineering, UC San Diego
{mtelgars,dasgupta}@cs.ucsd.edu

Abstract

Suppose k centers are fit to m points by heuristically minimizing the k-means cost; what is the corresponding fit over the source distribution? This question is resolved here for distributions with $p \geq 4$ bounded moments; in particular, the difference between the sample cost and distribution cost decays with m and p as $m^{\min\{-1/4,-1/2+2/p\}}$. The essential technical contribution is a mechanism to uniformly control deviations in the face of unbounded parameter sets, cost functions, and source distributions. To further demonstrate this mechanism, a soft clustering variant of k-means cost is also considered, namely the log likelihood of a Gaussian mixture, subject to the constraint that all covariance matrices have bounded spectrum. Lastly, a rate with refined constants is provided for k-means instances possessing some cluster structure.

1 Introduction

Suppose a set of k centers $\{p_i\}_{i=1}^k$ is selected by approximate minimization of k-means cost; how does the fit over the sample compare with the fit over the distribution? Concretely: given m points sampled from a source distribution ρ, what can be said about the quantities

$$\left|\frac{1}{m}\sum_{j=1}^{m}\min_i \|x_j - p_i\|_2^2 - \int \min_i \|x - p_i\|_2^2 d\rho(x)\right| \qquad (k\text{-means}), \tag{1.1}$$

$$\left|\frac{1}{m}\sum_{j=1}^{m}\ln\left(\sum_{i=1}^{k}\alpha_i p_{\theta_i}(x_j)\right) - \int \ln\left(\sum_{i=1}^{k}\alpha_i p_{\theta_i}(x)\right) d\rho(x)\right| \qquad (\text{soft } k\text{-means}), \tag{1.2}$$

where each p_{θ_i} denotes the density of a Gaussian with a covariance matrix whose eigenvalues lie in some closed positive interval.

The literature offers a wealth of information related to this question. For k-means, there is firstly a consistency result: under some identifiability conditions, the global minimizer over the sample will converge to the global minimizer over the distribution as the sample size m increases [1]. Furthermore, if the distribution is bounded, standard tools can provide deviation inequalities [2, 3, 4]. For the second problem, which is maximum likelihood of a Gaussian mixture (thus amenable to EM [5]), classical results regarding the consistency of maximum likelihood again provide that, under some identifiability conditions, the optimal solutions over the sample converge to the optimum over the distribution [6].

The task here is thus: to provide finite sample guarantees for these problems, but eschewing boundedness, subgaussianity, and similar assumptions in favor of moment assumptions.

1.1 Contribution

The results here are of the following form: given m examples from a distribution with a few bounded moments, and any set of parameters beating some fixed cost c, the corresponding deviations in cost (as in eq. (1.1) and eq. (1.2)) approach $\mathcal{O}(m^{-1/2})$ with the availability of higher moments.

- In the case of k-means (cf. Corollary 3.1), $p \geq 4$ moments suffice, and the rate is $\mathcal{O}(m^{\min\{-1/4,-1/2+2/p\}})$. For Gaussian mixtures (cf. Theorem 5.1), $p \geq 8$ moments suffice, and the rate is $\mathcal{O}(m^{-1/2+3/p})$.
- The parameter c allows these guarantees to hold for heuristics. For instance, suppose k centers are output by Lloyd's method. While Lloyd's method carries no optimality guarantees, the results here hold for the output of Lloyd's method simply by setting c to be the variance of the data, equivalently the k-means cost with a single center placed at the mean.
- The k-means and Gaussian mixture costs are only well-defined when the source distribution has $p \geq 2$ moments. The condition of $p \geq 4$ moments, meaning the variance has a variance, allows consideration of many heavy-tailed distributions, which are ruled out by boundedness and subgaussianity assumptions.

The main technical byproduct of the proof is a mechanism to deal with the unboundedness of the cost function; this technique will be detailed in Section 3, but the difficulty and its resolution can be easily sketched here.

For a single set of centers P, the deviations in eq. (1.1) may be controlled with an application of Chebyshev's inequality. But this does not immediately grant deviation bounds on another set of centers P', even if P and P' are very close: for instance, the difference between the two costs will grow as successively farther and farther away points are considered.

The resolution is to simply note that there is so little probability mass in those far reaches that the cost there is irrelevant. Consider a single center p (and assume $x \mapsto \|x-p\|_2^2$ is integrable); the dominated convergence theorem grants

$$\int_{B_i} \|x-p\|_2^2 d\rho(x) \quad \to \quad \int \|x-p\|_2^2 d\rho(x), \qquad\qquad \text{where } B_i := \{x \in \mathbb{R}^d : \|x-p\|_2 \leq i\}.$$

In other words, a ball B_i may be chosen so that $\int_{B_i^c} \|x-p\|_2^2 d\rho(x) \leq 1/1024$. Now consider some p' with $\|p-p'\|_2 \leq i$. Then

$$\int_{B_i^c} \|x-p'\|_2^2 d\rho(x) \leq \int_{B_i^c} (\|x-p\|_2 + \|p-p'\|_2)^2 d\rho(x) \leq 4\int_{B_i^c} \|x-p\|_2^2 d\rho(x) \leq \frac{1}{256}.$$

In this way, a single center may control the outer deviations of whole swaths of other centers. Indeed, those choices outperforming the reference score c will provide a suitable swath. Of course, it would be nice to get a sense of the size of B_i; this however is provided by the moment assumptions.

The general strategy is thus to split consideration into outer deviations, and local deviations. The local deviations may be controlled by standard techniques. To control outer deviations, a single pair of dominating costs — a lower bound and an upper bound — is controlled.

This technique can be found in the proof of the consistency of k-means due to Pollard [1]. The present work shows it can also provide finite sample guarantees, and moreover be applied outside hard clustering.

The content here is organized as follows. The remainder of the introduction surveys related work, and subsequently Section 2 establishes some basic notation. The core deviation technique, termed *outer bracketing* (to connect it to the bracketing technique from empirical process theory), is presented along with the deviations of k-means in Section 3. The technique is then applied in Section 5 to a soft clustering variant, namely log likelihood of Gaussian mixtures having bounded spectra. As a reprieve between these two heavier bracketing sections, Section 4 provides a simple refinement for k-means which can adapt to cluster structure.

All proofs are deferred to the appendices, however the construction and application of outer brackets is sketched in the text.

1.2 Related Work

As referenced earlier, Pollard's work deserves special mention, both since it can be seen as the origin of the outer bracketing technique, and since it handled k-means under similarly slight assumptions (just two moments, rather than the four here) [1, 7]. The present work hopes to be a spiritual successor, providing finite sample guarantees, and adapting technique to a soft clustering problem.

In the machine learning community, statistical guarantees for clustering have been extensively studied under the topic of *clustering stability* [4, 8, 9, 10]. One formulation of stability is: if parameters are learned over two samples, how close are they? The technical component of these works frequently involves finite sample guarantees, which in the works listed here make a boundedness assumption, or something similar (for instance, the work of Shamir and Tishby [9] requires the cost function to satisfy a bounded differences condition). Amongst these finite sample guarantees, the finite sample guarantees due to Rakhlin and Caponnetto [4] are similar to the development here *after* the invocation of the outer bracket: namely, a covering argument controls deviations over a bounded set. The results of Shamir and Tishby [10] do not make a boundedness assumption, but the main results are not finite sample guarantees; in particular, they rely on asymptotic results due to Pollard [7].

There are many standard tools which may be applied to the problems here, particularly if a boundedness assumption is made [11, 12]; for instance, Lugosi and Zeger [2] use tools from VC theory to handle k-means in the bounded case. Another interesting work, by Ben-david [3], develops specialized tools to measure the complexity of certain clustering problems; when applied to the problems of the type considered here, a boundedness assumption is made.

A few of the above works provide some negative results and related commentary on the topic of uniform deviations for distributions with unbounded support [10, Theorem 3 and subsequent discussion] [3, Page 5 above Definition 2]. The primary "loophole" here is to constrain consideration to those solutions beating some reference score c. It is reasonable to guess that such a condition entails that a few centers must lie near the bulk of the distribution's mass; making this guess rigorous is the first step here both for k-means and for Gaussian mixtures, and moreover the same consequence was used by Pollard for the consistency of k-means [1]. In Pollard's work, only optimal choices were considered, but the same argument relaxes to arbitrary c, which can thus encapsulate heuristic schemes, and not just nearly optimal ones. (The secondary loophole is to make moment assumptions; these sufficiently constrain the structure of the distribution to provide rates.)

In recent years, the empirical process theory community has produced a large body of work on the topic of maximum likelihood (see for instance the excellent overviews and recent work of Wellner [13], van der Vaart and Wellner [14], Gao and Wellner [15]). As stated previously, the choice of the term "bracket" is to connect to empirical process theory. Loosely stated, a bracket is simply a pair of functions which sandwich some set of functions; the *bracketing entropy* is then (the logarithm of) the number of brackets needed to control a particular set of functions. In the present work, brackets are paired with sets which identify the far away regions they are meant to control; furthermore, while there is potential for the use of many outer brackets, the approach here is able to make use of just a single outer bracket. The name bracket is suitable, as opposed to cover, since the bracketing elements need not be members of the function class being dominated. (By contrast, Pollard's use in the proof of the consistency of k-means was more akin to covering, in that remote fluctuations were compared to that of a a single center placed at the origin [1].)

2 Notation

The ambient space will always be the Euclidean space $\mathbb{R}^d$, though a few results will be stated for a general domain $\mathcal{X}$. The source probability measure will be ρ, and when a finite sample of size m is available, $\hat\rho$ is the corresponding empirical measure. Occasionally, the variable ν will refer to an arbitrary probability measure (where ρ and $\hat\rho$ will serve as relevant instantiations). Both integral and expectation notation will be used; for example, $\mathbb{E}(f(X)) = \mathbb{E}_\rho(f(X) = \int f(x)d\rho(x)$; for integrals, $\int_B f(x)d\rho(x) = \int f(x)\mathbb{1}[x \in B]d\rho(x)$, where $\mathbb{1}$ is the indicator function. The moments of ρ are defined as follows.

Definition 2.1. *Probability measure ρ has* order-p moment bound M with respect to norm $\|\cdot\|$ *when* $\mathbb{E}_\rho\|X - \mathbb{E}_\rho(X)\|^l \leq M$ *for* $1 \leq l \leq p$.

For example, the typical setting of k-means uses norm $\|\cdot\|_2$, and at least two moments are needed for the cost over ρ to be finite; the condition here of needing 4 moments can be seen as naturally arising via Chebyshev's inequality. Of course, the availability of higher moments is beneficial, dropping the rates here from $m^{-1/4}$ down to $m^{-1/2}$. Note that the basic controls derived from moments, which are primarily elaborations of Chebyshev's inequality, can be found in Appendix A.

The k-means analysis will generalize slightly beyond the single-center cost $x \mapsto \|x - p\|_2^2$ via *Bregman divergences* [16, 17].

Definition 2.2. *Given a convex differentiable function $f : \mathcal{X} \to \mathbb{R}$, the corresponding* Bregman divergence *is $\mathsf{B}_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y\rangle$.*

Not all Bregman divergences are handled; rather, the following regularity conditions will be placed on the convex function.

Definition 2.3. *A convex differentiable function f is* strongly convex *with modulus r_1 and has* Lipschitz gradients *with constant r_2, both respect to some norm $\|\cdot\|$, when f (respectively) satisfies*

$$\begin{aligned}
f(\alpha x + (1-\alpha)y) &\le \alpha f(x) + (1-\alpha) f(y) - \frac{r_1 \alpha(1-\alpha)}{2}\|x - y\|^2, \\
\|\nabla f(x) - \nabla f(y)\|_* &\le r_2 \|x - y\|,
\end{aligned}$$

where $x, y \in \mathcal{X}$, $\alpha \in [0, 1]$, and $\|\cdot\|_*$ is the dual of $\|\cdot\|$. (The Lipschitz gradient condition is sometimes called strong smoothness.*)*

These conditions are a fancy way of saying the corresponding Bregman divergence is sandwiched between two quadratics (cf. Lemma B.1).

Definition 2.4. *Given a convex differentiable function $f : \mathbb{R}^d \to R$ which is strongly convex and has Lipschitz gradients with respective constants r_1, r_2 with respect to norm $\|\cdot\|$, the* hard k-means cost *of a single point x according to a set of centers P is*

$$\phi_f(x; P) := \min_{p \in P} \mathsf{B}_f(x, p).$$

The corresponding k-means cost of a set of points (or distribution) is thus computed as $\mathbb{E}_\nu(\phi_f(X; P))$, and let $\mathcal{H}_f(\nu; c, k)$ denote all sets of at most k centers beating cost c, meaning

$$\mathcal{H}_f(\nu; c, k) := \{P : |P| \le k, \mathbb{E}_\nu(\phi_f(X; P)) \le c\}.$$

For example, choosing norm $\|\cdot\|_2$ and convex function $f(x) = \|x\|_2^2$ (which has $r_1 = r_2 = 2$), the corresponding Bregman divergence is $\mathsf{B}_f(x, y) = \|x - y\|_2^2$, and $\mathbb{E}_{\hat\rho}(\phi_f(X; P))$ denotes the vanilla k-means cost of some finite point set encoded in the empirical measure $\hat\rho$.

The hard clustering guarantees will work with $\mathcal{H}_f(\nu; c, k)$, where ν can be either the source distribution ρ, or its empirical counterpart $\hat\rho$. As discussed previously, it is reasonable to set c to simply the sample variance of the data, or a related estimate of the true variance (cf. Appendix A).

Lastly, the class of Gaussian mixture penalties is as follows.

Definition 2.5. *Given Gaussian parameters $\theta := (\mu, \Sigma)$, let p_θ denote Gaussian density*

$$p_\theta(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right).$$

Given Gaussian mixture parameters $(\alpha, \Theta) = (\{\alpha_i\}_{i=1}^k, \{\theta_i\}_{i=1}^k)$ with $\alpha \ge 0$ and $\sum_i \alpha_i = 1$ (written $\alpha \in \Delta$), the Gaussian mixture cost at a point x is

$$\phi_{\mathrm{g}}(x; (\alpha, \Theta)) := \phi_{\mathrm{g}}(x; \{(\alpha_i, \theta_i) = (\alpha_i, \mu_i, \Sigma_i)\}_{i=1}^k) := \ln\left(\sum_{i=1}^k \alpha_i p_{\theta_i}(x)\right),$$

Lastly, given a measure ν, bound k on the number of mixture parameters, and spectrum bounds $0 < \sigma_1 \le \sigma_2$, let $\mathcal{S}_{\mathrm{mog}}(\nu; c, k, \sigma_1, \sigma_2)$ denote those mixture parameters beating cost c, meaning

$$\mathcal{S}_{\mathrm{mog}}(\nu; c, k, \sigma_1, \sigma_2) := \{(\alpha, \Theta) : \sigma_1 I \preceq \Sigma_i \preceq \sigma_2 I, |\alpha| \le k, \alpha \in \Delta, \mathbb{E}_\nu(\phi_{\mathrm{g}}(X; (\alpha, \Theta))) \le c\}.$$

While a condition of the form $\Sigma \succeq \sigma_1 I$ is typically enforced in practice (say, with a Bayesian prior, or by ignoring updates which shrink the covariance beyond this point), the condition $\Sigma \preceq \sigma_2 I$ is potentially violated. These conditions will be discussed further in Section 5.

3 Controlling k-means with an Outer Bracket

First consider the special case of k-means cost.

Corollary 3.1. *Set $f(x) := \|x\|_2^2$, whereby ϕ_f is the k-means cost. Let real $c \geq 0$ and probability measure ρ be given with order-p moment bound M with respect to $\|\cdot\|_2$, where $p \geq 4$ is a positive multiple of 4. Define the quantities*

$$c_1 := (2M)^{1/p} + \sqrt{2c}, \quad M_1 := M^{1/(p-2)} + M^{2/p}, \quad N_1 := 2 + 576d(c_1 + c_1^2 + M_1 + M_1^2).$$

Then with probability at least $1 - 3\delta$ over the draw of a sample of size $m \geq \max\{(p/(2^{p/4+2}e))^2, 9\ln(1/\delta)\}$, every set of centers $P \in \mathcal{H}_f(\hat{\rho}; c, k) \cup \mathcal{H}_f(\rho; c, k)$ satisfies

$$\begin{aligned}&\left|\int \phi_f(x;P)d\rho(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right| \\ &\quad\leq m^{-1/2+\min\{1/4,2/p\}}\left(4 + (72c_1^2 + 32M_1^2)\sqrt{\frac{1}{2}\ln\left(\frac{(mN_1)^{dk}}{\delta}\right)} + \sqrt{\frac{2^{p/4}ep}{8m^{1/2}}}\left(\frac{2}{\delta}\right)^{4/p}\right).\end{aligned}$$

One artifact of the moment approach (cf. Appendix A), heretofore ignored, is the term $(2/\delta)^{4/p}$. While this may seem inferior to $\ln(2/\delta)$, note that the choice $p = 4\ln(2/\delta)/\ln(\ln(2/\delta))$ suffices to make the two equal.

Next consider a general bound for Bregman divergences. This bound has a few more parameters than Corollary 3.1. In particular, the term ϵ, which is instantiated to $m^{-1/2+1/p}$ in the proof of Corollary 3.1, catches the mass of points discarded due to the outer bracket, as well as the resolution of the (inner) cover. The parameter p', which controls the tradeoff between m and $1/\delta$, is set to $p/4$ in the proof of Corollary 3.1.

Theorem 3.2. *Fix a reference norm $\|\cdot\|$ throughout the following. Let probability measure ρ be given with order-p moment bound M where $p \geq 4$, a convex function f with corresponding constants r_1 and r_2, reals c and $\epsilon > 0$, and integer $1 \leq p' \leq p/2 - 1$ be given. Define the quantities*

$$\begin{aligned}R_B &:= \max\left\{(2M)^{1/p} + \sqrt{4c/r_1}, \max_{i\in[p']}(M/\epsilon)^{1/(p-2i)}\right\}, \\ R_C &:= \sqrt{r_2/r_1}\left((2M)^{1/p} + \sqrt{4c/r_1} + R_B\right) + R_B, \\ B &:= \left\{x \in \mathbb{R}^d : \|x - \mathbb{E}(X)\| \leq R_B\right\}, \\ C &:= \left\{x \in \mathbb{R}^d : \|x - \mathbb{E}(X)\| \leq R_C\right\}, \\ \tau &:= \min\left\{\sqrt{\frac{\epsilon}{2r_2}}, \frac{\epsilon}{2(R_B + R_C)r_2}\right\},\end{aligned}$$

and let $\mathcal{N}$ be a cover of C by $\|\cdot\|$-balls with radius τ; in the case that $\|\cdot\|$ is an l_p norm, the size of this cover has bound

$$|\mathcal{N}| \leq \left(1 + \frac{2R_C d}{\tau}\right)^d.$$

Then with probability at least $1 - 3\delta$ over the draw of a sample of size $m \geq \max\{p'/(e2^{p'}\epsilon), 9\ln(1/\delta)\}$, every set of centers $P \in \mathcal{H}_f(\rho; c, k) \cup \mathcal{H}_f(\hat{\rho}; c, k)$ satisfies

$$\left|\int \phi_f(x;P)d\rho(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right| \leq 4\epsilon + 4r_2R_C^2\sqrt{\frac{1}{2m}\ln\left(\frac{2|\mathcal{N}|^k}{\delta}\right)} + \sqrt{\frac{e2^{p'}\epsilon p'}{2m}}\left(\frac{2}{\delta}\right)^{1/p'}.$$

3.1 Compactification via Outer Brackets

The outer bracket is defined as follows.

Definition 3.3. *An outer bracket for probability measure ν at scale ϵ consists of two triples, one each for lower and upper bounds.*

1. *The function ℓ, function class Z_ℓ, and set B_ℓ satisfy two conditions: if $x \in B_\ell^c$ and $\phi \in Z_\ell$, then $\ell(x) \leq \phi(x)$, and secondly $|\int_{B_\ell^c} \ell(x) d\nu(x)| \leq \epsilon$.*
2. *Similarly, function u, function class Z_u, and set B_u satisfy: if $x \in B_u^c$ and $\phi \in Z_u$, then $u(x) \geq \phi(x)$, and secondly $|\int_{B_u^c} u(x) d\nu(x)| \leq \epsilon$.*

Direct from the definition, given bracketing functions (ℓ, u), a bracketed function $\phi_f(\cdot; P)$, and the bracketing set $B := B_u \cup B_\ell$,

$$-\epsilon \leq \int_{B^c} \ell(x) d\nu(x) \leq \int_{B^c} \phi_f(x; P) d\nu(x) \leq \int_{B^c} u(x) d\nu(x) \leq \epsilon; \tag{3.4}$$

in other words, as intended, this mechanism allows deviations on B^c to be discarded. Thus to uniformly control the deviations of the dominated functions $Z := Z_u \cup Z_\ell$ over the set B^c, it suffices to simply control the deviations of the pair (ℓ, u).

The following lemma shows that a bracket exists for $\{\phi_f(\cdot; P) : P \in \mathcal{H}_f(\nu; c, k)\}$ and compact B, and moreover that this allows sampled points and candidate centers in far reaches to be deleted.

Lemma 3.5. *Consider the setting and definitions in Theorem 3.2, but additionally define*

$$M' := 2^{p'} \epsilon, \qquad \ell(x) := 0, \qquad u(x) := 4r_2 \|x - \mathbb{E}(X)\|^2, \qquad \epsilon_{\hat{\rho}} := \epsilon + \sqrt{\frac{M' e p'}{2m}} \left(\frac{2}{\delta}\right)^{1/p'}.$$

The following statements hold with probability at least $1 - 2\delta$ over a draw of size $m \geq \max\{p'/(M'e), 9\ln(1/\delta)\}$.

1. *(u, ℓ) is an outer bracket for ρ at scale $\epsilon_\rho := \epsilon$ with sets $B_\ell = B_u = B$ and $Z_\ell = Z_u = \{\phi_f(\cdot; P) : P \in \mathcal{H}_f(\hat{\rho}; c, k) \cup \mathcal{H}_f(\rho; c, k)\}$, and furthermore the pair (u, ℓ) is also an outer bracket for $\hat{\rho}$ at scale $\epsilon_{\hat{\rho}}$ with the same sets.*
2. *For every $P \in \mathcal{H}_f(\hat{\rho}; c, k) \cup \mathcal{H}_f(\rho; c, k)$,*

$$\left| \int \phi_f(x; P) d\rho(x) - \int_B \phi_f(x; P \cap C) d\rho(x) \right| \leq \epsilon_\rho = \epsilon.$$

and

$$\left| \int \phi_f(x; P) d\hat{\rho}(x) - \int_B \phi_f(x; P \cap C) d\hat{\rho}(x) \right| \leq \epsilon_{\hat{\rho}}.$$

The proof of Lemma 3.5 has roughly the following outline.

1. Pick some ball B_0 which has probability mass at least $1/4$. It is not possible for an element of $\mathcal{H}_f(\hat{\rho}; c, k) \cup \mathcal{H}_f(\rho; c, k)$ to have all centers far from B_0, since otherwise the cost is larger than c. (Concretely, "far from" means at least $\sqrt{4c/r_1}$ away; note that this term appears in the definitions of B and C in Theorem 3.2.) Consequently, at least one center lies near to B_0; this reasoning was also the first step in the k-means consistency proof due to k-means Pollard [1].
2. It is now easy to dominate $P \in \mathcal{H}_f(\hat{\rho}; c, k) \cup \mathcal{H}_f(\rho; c, k)$ far away from B_0. In particular, choose any $p_0 \in B_0 \cap P$, which was guaranteed to exist in the preceding point; since $\min_{p \in P} \mathsf{B}_f(x, p) \leq \mathsf{B}_f(x, p_0)$ holds for all x, it suffices to dominate p_0. This domination proceeds exactly as discussed in the introduction; in fact, the factor 4 appeared there, and again appears in the u here, for exactly the same reason. Once again, similar reasoning can be found in the proof by Pollard [1].
3. Satisfying the integral conditions over ρ is easy: it suffices to make B huge. To control the size of B_0, as well as the size of B, and moreover the deviations of the bracket over B, the moment tools from Appendix A are used.

Now turning consideration back to the proof of Theorem 3.2, the above bracketing allows the removal of points and centers outside of a compact set (in particular, the pair of compact sets B and C, respectively). On the remaining truncated data and set of centers, any standard tool suffices; for mathematical convenience, and also to fit well with the use of norms in the definition of moments as well as the conditions on the convex function f providing the divergence B_f, norm structure used throughout the other properties, covering arguments are used here. (For details, please see Appendix B.)

4 Interlude: Refined Estimates via Clamping

So far, rates have been given that guarantee uniform convergence when the distribution has a few moments, and these rates improve with the availability of higher moments. These moment conditions, however, do not necessarily reflect any natural cluster structure in the source distribution. The purpose of this section is to propose and analyze another distributional property which is intended to capture cluster structure. To this end, consider the following definition.

Definition 4.1. *Real number R and compact set C are a* clamp *for probability measure ν and family of centers Z and cost ϕ_f at scale $\epsilon > 0$ if every $P \in Z$ satisfies*

$$|\mathbb{E}_\nu(\phi_f(X;P)) - \mathbb{E}_\nu\left(\min\left\{\phi_f(X;P\cap C)\ ,\ R\right\}\right)| \leq \epsilon.$$

Note that this definition is similar to the second part of the outer bracket guarantee in Lemma 3.5, and, predictably enough, will soon lead to another deviation bound.

Example 4.2. If the distribution has bounded support, then choosing a clamping value R and clamping set C respectively slightly larger than the support size and set is sufficient: as was reasoned in the construction of outer brackets, if no centers are close to the support, then the cost is bad. Correspondingly, the clamped set of functions Z should again be choices of centers whose cost is not too high.

For a more interesting example, suppose ρ is supported on k small balls of radius R_1, where the distance between their respective centers is some $R_2 \gg R_1$. Then by reasoning similar to the bounded case, all choices of centers achieving a good cost will place centers near to each ball, and thus the clamping value can be taken closer to R_1. ■

Of course, the above gave the existence of clamps under favorable conditions. The following shows that outer brackets can be used to show the existence of clamps in general. In fact, the proof is very short, and follows the scheme laid out in the bounded example above: outer bracketing allows the restriction of consideration to a bounded set, and some algebra from there gives a conservative upper bound for the clamping value.

Proposition 4.3. *Suppose the setting and definitions of Lemma 3.5, and additionally define*

$$R := 2((2M)^{2/p} + R_B^2).$$

Then (C, R) is a clamp for measure ρ and center $\mathcal{H}_f(\rho; c, k)$ at scale ϵ, and with probability at least $1 - 3\delta$ over a draw of size $m \geq \max\{p'/(M'e), 9\ln(1/\delta)\}$, it is also a clamp for $\hat{\rho}$ and centers $\mathcal{H}_f(\hat{\rho}; c, k)$ at scale $\epsilon_{\hat{\rho}}$.

The general guarantee using clamps is as follows. The proof is almost the same as for Theorem 3.2, but note that this statement is not used quite as readily, since it first requires the construction of clamps.

Theorem 4.4. *Fix a norm $\|\cdot\|$. Let (R, C) be a clamp for probability measure ρ and empirical counterpart $\hat{\rho}$ over some center class Z and cost ϕ_f at respective scales ϵ_ρ and $\epsilon_{\hat{\rho}}$, where f has corresponding convexity constants r_1 and r_2. Suppose C is contained within a ball of radius R_C, let $\epsilon > 0$ be given, define scale parameter*

$$\tau := \min\left\{\sqrt{\frac{\epsilon}{2r_2}}, \frac{r_1\epsilon}{2r_2R_3}\right\},$$

and let $\mathcal{N}$ be a cover of C by $\|\cdot\|$-balls of radius τ (as per lemma B.4, if $\|\cdot\|$ is an l_p norm, then $|\mathcal{N}| \leq (1 + (2R_Cd)/\tau)^d$ suffices). Then with probability at least $1 - \delta$ over the draw of a sample of size $m \geq p'/(M'e)$, every set of centers $P \in Z$ satisfies

$$\left|\int \phi_f(x;P)d\rho(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right| \leq 2\epsilon + \epsilon_\rho + \epsilon_{\hat{\rho}} + R^2\sqrt{\frac{1}{2m}\ln\left(\frac{2|\mathcal{N}|^k}{\delta}\right)}.$$

Before adjourning this section, note that clamps and outer brackets disagree on the treatment of the outer regions: the former replaces the cost there with the fixed value R, whereas the latter uses the value 0. On the technical side, this is necessitated by the covering argument used to produce the final theorem: if the clamping operation instead truncated beyond a ball of radius R centered at each $p \in P$, then the deviations would be wild as these balls moved and suddenly switched the value at a point from 0 to something large. This is not a problem with outer bracketing, since the same points (namely B^c) are ignored by every set of centers.

5 Mixtures of Gaussians

Before turning to the deviation bound, it is a good place to discuss the condition $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$, which must be met by every covariance matrix of every constituent Gaussian in a mixture.

The lower bound $\sigma_1 I \preceq \Sigma$, as discussed previously, is fairly common in practice, arising either via a Bayesian prior, or by implementing EM with an explicit condition that covariance updates are discarded when the eigenvalues fall below some threshold. In the analysis here, this lower bound is used to rule out two kinds of bad behavior.

1. Given a budget of at least 2 Gaussians, and a sample of at least 2 distinct points, arbitrarily large likelihood may be achieved by devoting one Gaussian to one point, and shrinking its covariance. This issue destroys convergence properties of maximum likelihood, since the likelihood score may be arbitrarily large over every sample, but is finite for well-behaved distributions. The condition $\sigma_1 I \preceq \Sigma$ rules this out.
2. Another phenomenon is a "flat" Gaussian, meaning a Gaussian whose density is high along a lower dimensional manifold, but small elsewhere. Concretely, consider a Gaussian over $\mathbb{R}^2$ with covariance $\Sigma = \text{diag}(\sigma, \sigma^{-1})$; as σ decreases, the Gaussian has large density on a line, but low density elsewhere. This phenomenon is distinct from the preceding in that it does not produce arbitrarily large likelihood scores over finite samples. The condition $\sigma_1 I \preceq \Sigma$ rules this situation out as well.

In both the hard and soft clustering analyses here, a crucial early step allows the assertion that good scores in some region mean the relevant parameter is nearby. For the case of Gaussians, the condition $\sigma_1 I \preceq \Sigma$ makes this problem manageable, but there is still the possibility that some far away, fairly uniform Gaussian has reasonable density. This case is ruled out here via $\sigma_2 I \succeq \Sigma$.

Theorem 5.1. *Let probability measure ρ be given with order-p moment bound M according to norm $\|\cdot\|_2$ where $p \geq 8$ is a positive multiple of 4, covariance bounds $0 < \sigma_1 \leq \sigma_2$ with $\sigma_1 \leq 1$ for simplicity, and real $c \leq 1/2$ be given. Then with probability at least $1 - 5\delta$ over the draw of a sample of size $m \geq \max\left\{(p/(2^{p/4+2}e))^2, 8\ln(1/\delta), d^2\ln(\pi\sigma_2)^2\ln(1/\delta)\right\}$, every set of Gaussian mixture parameters $(\alpha, \Theta) \in \mathcal{S}_{\text{mog}}(\hat{\rho}; c, k, \sigma_1, \sigma_2) \cup \mathcal{S}_{\text{mog}}(\rho; c, k, \sigma_1, \sigma_2)$ satisfies*

$$\left|\int \phi_{\text{g}}(x; (\alpha, \Theta))d\rho(x) - \int \phi_{\text{g}}(x; (\alpha, \Theta))d\hat{\rho}(x)\right|$$
$$= \mathcal{O}\left(m^{-1/2+3/p}\left(1 + \sqrt{\ln(m) + \ln(1/\delta)} + (1/\delta)^{4/p}\right)\right),$$

where the $\mathcal{O}(\cdot)$ drops numerical constants, polynomial terms depending on c, M, d, and k, σ_2/σ_1, and $\ln(\sigma_2/\sigma_1)$, but in particular has no sample-dependent quantities.

The proof follows the scheme of the hard clustering analysis. One distinction is that the outer bracket now uses both components; the upper component is the log of the largest possible density — indeed, it is $\ln((2\pi\sigma_1)^{-d/2})$ — whereas the lower component is a function mimicking the log density of the steepest possible Gaussian — concretely, the lower bracket's definition contains the expression $\ln((2\pi\sigma_2)^{-d/2}) - 2\|x - \mathbb{E}_\rho(X)\|_2^2/\sigma_1$, which lacks the normalization of a proper Gaussian, highlighting the fact that bracketing elements need not be elements of the class. Superficially, a second distinction with the hard clustering case is that far away Gaussians can not be entirely ignored on local regions; the influence is limited, however, and the analysis proceeds similarly in each case.

Acknowledgments

The authors thank the NSF for supporting this work under grant IIS-1162581.

References

[1] David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981.

[2] Gbor Lugosi and Kenneth Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40: 1728–1740, 1994.

[3] Shai Ben-david. A framework for statistical clustering with a constant time approximation algorithms for k-median clustering. In *COLT*, pages 415–426. Springer, 2004.

[4] Alexander Rakhlin and Andrea Caponnetto. Stability of k-means clustering. In *NIPS*, pages 1121–1128, 2006.

[5] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2 edition, 2001.

[6] Thomas S. Ferguson. *A course in large sample theory*. Chapman & Hall, 1996.

[7] David Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10 (4):919–926, 1982.

[8] Shai Ben-david, Ulrike Von Luxburg, and Dávid Pál. A sober look at clustering stability. In *In COLT*, pages 5–19. Springer, 2006.

[9] Ohad Shamir and Naftali Tishby. Cluster stability for finite samples. In *Annals of Probability, 10(4)*, pages 919–926, 1982.

[10] Ohad Shamir and Naftali Tishby. Model selection and stability in k-means clustering. In *COLT*, 2008.

[11] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[12] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, 2013.

[13] Jon Wellner. Consistency and rates of convergence for maximum likelihood estimators via empirical process theory. 2005.

[14] Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[15] FuChang Gao and Jon A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52(7):1525–1538, 2009.

[16] Yair Al Censor and Stavros A. Zenios. *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, 1997.

[17] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[18] Terence Tao. 254a notes 1: Concentration of measure, January 2010. URL `http://terrytao.wordpress.com/2010/01/03/254a-notes-1-concentration-of-measure/`.

[19] I. F. Pinelis and S. A. Utev. Estimates of the moments of sums of independent random variables. *Teor. Veroyatnost. i Primenen.*, 29(3):554–557, 1984. Translation to English by Bernard Seckler.

[20] Shai Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, July 2007.

[21] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Publishing Company, Incorporated, 2001.

A Moment Bounds

This section provides the basic probability controls resulting from moments. The material deals with the following slight generalization of the bounded moment definition from Section 2.

Definition A.1. *A function $\tau : \mathcal{X} \to \mathbb{R}^d$ has* order-p moment bound M for probability measure ρ with respect to norm $\|\cdot\|$ *if $\mathbb{E}_\rho(\|\tau(X)\|^l) \le M$ for all $1 \le l \le p$. (For convenience, measure ρ and norm $\|\cdot\|$ will be often be implicit.)*

To connect this to the earlier definition, simply choose the map $\tau(x) := x - \mathbb{E}_\rho(X)$. As was the case in Section 2, this definition requires a uniform bound across all l^{th} moments for $1 \le l \le p$. Of course, working with a probability measure implies these moments are all finite when just the p^{th} moment is finite. The significance of working with a bound across all moments will be discussed again in the context of Lemma A.3 below.

The first result controls the measures of balls thanks to moments. This result is only stated for the source distribution ρ, but Hoeffding's inequality suffices to control $\hat{\rho}$.

Lemma A.2. *Suppose τ has order-p moment bound M. Then for any $\epsilon > 0$,*

$$\Pr\left[\|\tau(X)\| \le (M/\epsilon)^{1/p}\right] \ge 1 - \epsilon.$$

Proof. If $M = 0$, the result is immediate. Otherwise, when $M > 0$, for any $R > 0$, by Chebyshev's inequality,

$$\Pr\left[\|\tau(X)\| < R\right] = 1 - \Pr\left[\|\tau(X)\| \ge R\right] \ge 1 - \frac{\mathbb{E}\|\tau(X)\|^p}{R^p} \ge 1 - \frac{M}{R^p};$$

the result follows by choosing $R := (M/\epsilon)^{1/p}$. □

The following fact will be the basic tool for controlling empirical averages via moments. Both the statement and proof are close to one by Tao [18, Equation 7], which rather than bounded moments uses boundedness (almost surely). As discussed previously, the term $1/\delta^{1/l}$ overtakes $\ln(1/\delta)$ when $l = \ln(1/\delta)/\ln(\ln(1/\delta))$.

For simplicity, this result is stated in terms of univariate random variables; to connect with the earlier development, the random variable X will be substituted with the map $x \mapsto \|\tau(x)\|$.

Lemma A.3. *(Cf. Tao [18, Equation 7].) Let m i.i.d. copies $\{X_i\}_{i=1}^m$ of a random variable X, even integer $p \ge 2$, real $M > 0$ with $\mathbb{E}(|X - \mathbb{E}(X)|^l) \le M$ for $2 \le l \le p$, and $\epsilon > 0$ be given. If $m \ge p/(Me)$, then*

$$\Pr\left(\left|\frac{1}{n}\sum_i X_i - \mathbb{E}(X)\right| \ge \epsilon\right) \le \frac{2}{(\epsilon\sqrt{m})^p}\left(\frac{Mpe}{2}\right)^{p/2}.$$

In other words, with probability at least $1 - \delta$ over a draw of size $m \ge p/(Me)$,

$$\left|\frac{1}{n}\sum_i X_i - \mathbb{E}(X)\right| \le \sqrt{\frac{Mpe}{2m}}\left(\frac{2}{\delta}\right)^{1/p}.$$

Proof. Without loss of generality, suppose $\mathbb{E}(X_1) = 0$ (i.e., given Y_1 with $\mathbb{E}(Y_1) \ne 0$, work with $X_i := Y_i - \mathbb{E}(Y_1)$). By Chebyshev's inequality,

$$\Pr\left(\left|\frac{1}{m}\sum_i X_i\right| \ge \epsilon\right) \le \frac{\mathbb{E}\left|\frac{1}{m}\sum_i X_i\right|^p}{\epsilon^p} = \frac{\mathbb{E}\left|\sum_i X_i\right|^p}{(m\epsilon)^p}. \tag{A.4}$$

Recalling p is even, consider the term

$$\mathbb{E}\left|\sum_i X_i\right|^p = \mathbb{E}\left(\sum_i X_i\right)^p = \sum_{i_1, i_2, \ldots, i_p \in [m]} \mathbb{E}\left(\prod_{j=1}^p X_{i_j}\right).$$

If some i_j is equal to none of the others, then, by independence, a term $\mathbb{E}(X_{i_j}) = 0$ is introduced and the product vanishes; thus the product is nonzero when each i_j has some copy $i_j = i_{j'}$, and thus there are at most $p/2$ distinct values amongst $\{i_j\}_{j=1}^p$. Each distinct value contributes a term $\mathbb{E}(X^l) \le \mathbb{E}(|X|^l) \le M$ for some $2 \le l \le p$, and thus

$$\mathbb{E}\left|\sum_i X_i\right|^p \le \sum_{r=1}^{p/2} M^r N_r, \tag{A.5}$$

where N_r is the number of ways to choose a multiset of size p from $[m]$, subject to the constraint that each number appears at least twice, and at most r distinct numbers appear. One way to over-count this is to first choose a subset of size r from m, and then draw from it (with repetition) p times:

$$N_r \le \binom{m}{r} r^p \le \frac{m^r r^p}{r!} \le \frac{m^r r^p}{(r/e)^r} = (me)^r r^{p-r}.$$

Plugging this into eq. (A.5), and thereafter re-indexing with $r := p/2 - j$,

$$\begin{aligned}
\mathbb{E}\left|\sum_i X_i\right|^p &\le \sum_{r=1}^{p/2} (Mme)^r r^{p-r} \le \sum_{r=1}^{p/2} (Mme)^r (p/2)^{p-r} \\
&\le \sum_{j=0}^{p/2} (Mme)^{p/2-j} (p/2)^{p/2+j} \le \left(\frac{Mmpe}{2}\right)^{p/2} \sum_{j=0}^{p/2} \left(\frac{p}{2Mme}\right)^j.
\end{aligned}$$

Since $p \le Mme$,

$$\mathbb{E}\left|\sum_i X_i\right|^p \le 2\left(\frac{Mmpe}{2}\right)^{p/2},$$

and the result follows by plugging this into eq. (A.4). □

Thanks to Chebyshev's inequality, proving Lemma A.3 boils down to controlling $\mathbb{E}|\sum_i X_i|^p$, which here relied on a combinatorial scheme by Tao [18, Equation 7]. There is, however, another approach to controlling this quantity, namely Rosenthal inequalities, which write this p^{th} moment of the sum in terms of the 2^{nd} and p^{th} moments of individual random variables (general material on these bounds can be found in the book of Boucheron et al. [12, Section 15.4], however the specific form provided here is most easily presented by Pinelis and Utev [19]). While Rosenthal inequalities may seem a more elegant approach, they involve different constants, and thus the approach and bound here are followed instead to suggest further work on how to best control $\mathbb{E}|\sum_i X_i|^p$.

Returning to task, as was stated in the introduction, the dominated convergence theorem provides that $\int_{B_i} \|x\|_2^2 d\rho(x) \to \int \|x\|_2^2 d\rho(x)$ (assuming integrability of $x \mapsto \|x\|_2^2$), where the sequence of balls $\{B_i\}_{i=1}^\infty$ grow in radius without bound; moment bounds allow the rate of this process to be quantified as follows.

Lemma A.6. *Suppose τ has order-p moment bound M, and let $0 < k < p$ be given. Then for any $\epsilon > 0$, the ball*

$$B := \left\{x \in \mathcal{X} : \|\tau(X)\| \le (M/\epsilon)^{1/(p-k)}\right\}$$

satisfies

$$\int_{B^c} \|\tau(x)\|^k d\rho(x) \le \epsilon.$$

Proof. Let the B be given as specified; an application of Lemma A.2 with $\epsilon' := (\epsilon^p/M^k)^{1/(p-k)}$ yields

$$\int \mathbb{1}[x \in B^c] d\rho(x) = \Pr[\|\tau(x)\| > (M/\epsilon)^{1/(p-k)}] = \Pr[\|\tau(x)\| > (M/\epsilon')^{1/p}] \le \epsilon'.$$

By Hölder's inequality with conjugate exponents p/k and $p/(p-k)$ (where the condition $0 < k < p$ means each lies within $(1, \infty)$),

$$\begin{aligned}\int_{B^c} \|\tau(x)\|^k d\rho(x) &= \int \|\tau(x)\|^k \mathbb{1}[x \in B^c] d\rho(x) \\ &\le \left(\int \|\tau(x)\|^{k(p/k)} d\rho(x)\right)^{k/p} \left(\int \mathbb{1}[x \in B^c]^{p/(p-k)} d\rho(x)\right)^{(p-k)/p} \\ &\le (M)^{k/p} \left(\frac{\epsilon^{p/(p-k)}}{M^{k/(p-k)}}\right)^{(p-k)/p} \\ &= \epsilon\end{aligned}$$

as desired. □

Lastly, thanks to the moment-based deviation inequality in Lemma A.3, the deviations on this outer region may be controlled. Note that in order to control the k-means cost (i.e., an exponent $k = 2$), at least 4 moments are necessary ($p \ge 4$).

Lemma A.7. *Let integers $k \ge 1$ and $p' \ge 1$ be given, and set $\tilde{p} := k(p'+1)$. Suppose τ has order-$\tilde{p}$ moment bound M, and let $\epsilon > 0$ be arbitrary. Define the radius R and ball B as*

$$R := \max\{(M/\epsilon)^{1/(\tilde{p}-ik)} : 1 \le i < \tilde{p}/k\} \qquad \text{and} \qquad B := \{x \in \mathcal{X} : \|\tau(x)\| \le R\},$$

and set $M' := 2^{p'}\epsilon$. With probability at least $1-\delta$ over the draw of a sample of size $m \ge p'/(M'e)$,

$$\left|\int_{B^c} \|\tau(x)\|^k d\hat{\rho}(x) - \int_{B^c} \|\tau(x)\|^k d\rho(x)\right| \le \sqrt{\frac{M'ep'}{2m}} \left(\frac{2}{\delta}\right)^{1/p'}.$$

Proof. Consider a fixed $1 \le i < \tilde{p}/k = p' + 1$, and set $l = ik$. Let B_l be the ball provided by Lemma A.6 for exponent l. Since $B \supseteq B_l$,

$$\int_{B^c} \|\tau(x)\|^l d\rho(x) \le \int_{B_l^c} \|\tau(x)\|^l d\rho(x) \le \epsilon.$$

As such, by Minkowski's inequality, since $z \mapsto z^l$ is convex for $l \ge 1$,

$$\begin{aligned}&\left(\int \left| \|\tau(x)\| \mathbb{1}[x \in B^c] - \int_{B^c} \|\tau(x)\| d\rho(x)\right|^l d\rho(x)\right)^{1/l} \\ &\quad \le \left(\int_{B^c} \|\tau(x)\|^l d\rho(x)\right)^{1/l} + \left(\int_{B^c} \|\tau(x)\| d\rho(x)\right)^{l/l} \\ &\quad \le 2\left(\int_{B^c} \|\tau(x)\|^l d\rho(x)\right)^{1/l},\end{aligned}$$

meaning

$$\int \left| \|\tau(x)\| \mathbb{1}[x \in B^c] - \int_{B^c} \|\tau(x)\| d\rho(x)\right|^l d\rho(x) \le 2^l \int_{B^c} \|\tau(x)\|^l \le 2^l \int_{B_l^c} \|\tau(x)\|^l \le 2^l \epsilon.$$

Since $l = ik$ had $1 \le i < \tilde{p}/k = p' + 1$ arbitrary, it follows that the map $x \mapsto \|\tau(x)\|^k \mathbb{1}[x \in B^c]$ has its first p' moments bounded by $2^{p'}\epsilon$.

The finite sample bounds will now proceed with an application of Lemma A.3, where the random variable X will be the map $x \mapsto \|\tau(x)\|^k \mathbb{1}[x \in B^c]$. Plugging the above moment bounds for this random variable into Lemma A.3, the result follows. □

B Deferred Material from Section 3

Before proceeding with the main proofs, note that Bregman divergences in the setting here are sandwiched between quadratics.

Lemma B.1. *If differentiable f is r_1 strongly convex with respect to $\|\cdot\|$, then $\mathsf{B}_f(x,y) \geq r_1\|x-y\|^2$. If differentiable f has Lipschitz gradients with parameter r_2 with respect to $\|\cdot\|$, then $\mathsf{B}_f(x,y) \leq r_2\|x-y\|^2$.*

Proof. The first part (strong convexity) is standard (see for instance the proof by Shalev-Shwartz [20, Lemma 13], or a similar proof by Hiriart-Urruty and Lemaréchal [21, Theorem B.4.1.4]). For the second part, by the fundamental theorem of calculus, properties of norm duality, and the Lipschitz gradient property,

$$\begin{aligned} f(x) &= f(y) + \langle \nabla f(y), x-y\rangle + \int_0^1 \langle \nabla f(y+t(x-y)) - \nabla f(y), x-y\rangle\, dt \\ &\leq f(y) + \langle \nabla f(y), x-y\rangle + \int_0^1 \|\nabla f(y+t(x-y)) - \nabla f(y)\|_* \|x-y\| dt \\ &\leq f(y) + \langle \nabla f(y), x-y\rangle + \frac{r_2}{2}\|x-y\|^2. \end{aligned}$$

(The preceding is also standard; see for instance the beginning of a proof by Hiriart-Urruty and Lemaréchal [21, Theorem E.4.2.2], which only differs by fixing the norm $\|\cdot\|_2$.) □

B.1 Proof of Lemma 3.5

The first step is the following characterization of $\mathcal{H}_f(\nu; c, k)$: at least one center must fall within some compact set. (The lemma works more naturally with the contrapositive.) The proof by Pollard [1] also started by controlling a single center.

Lemma B.2. *Consider the setting of Lemma 3.5, and additionally define the two balls*

$$\begin{aligned} B_0 &:= \left\{ x \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\| \leq (2M)^{1/p} \right\}, \\ C_0 &:= \left\{ x \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\| \leq (2M)^{1/p} + \sqrt{4c/r_1} \right\}, \end{aligned}$$

Then $\rho(B_0) \geq 1/2$, and for any center set P, if $P \cap C_0 = \emptyset$ then $\mathbb{E}_\rho(\phi_f(X;P)) \geq 2c$. Furthermore, with probability at least $1-\delta$ over a draw from ρ of size at least

$$m \geq 9 \ln\left(\frac{1}{\delta}\right).$$

then $\hat\rho(B_0) > 1/4$ and $P \cap C_0 = \emptyset$ implies $\mathbb{E}_{\hat\rho}(\phi_f(X;P)) > c$.

Proof. The guarantee $\rho(B_0) \geq 1/2$ is direct from Lemma A.2 with moment map $\tau(x) := x - \mathbb{E}_\rho(X)$. By Hoeffding's inequality and the lower bound on m, with probability at least $1-\delta$,

$$\hat\rho(B_0) \geq \rho(B_0) - \sqrt{\frac{1}{2m}\ln\left(\frac{1}{\delta}\right)} > \frac{1}{4}.$$

By the definition of C_0, every $p \in C_0^c$ and $x \in B_0$ satisfies

$$\mathsf{B}_f(x,p) \geq r_1\|x-p\|^2 \geq 4c.$$

Now let ν denote either ρ or $\hat\rho$; then for any set of centers P with $P \cap C_0 = \emptyset$ (meaning $P \subseteq C_0^c$),

$$\begin{aligned} \int \phi_f(x;P) d\nu(x) &= \int \min_{p\in P} \mathsf{B}_f(x,p) d\nu(x) \\ &\geq \int_{B_0} \min_{p\in P} \mathsf{B}_f(x,p) d\nu(x) \\ &\geq \int_{B_0} \min_{p\in P} 4c d\nu(x) \\ &= 4c\nu(B_0). \end{aligned}$$

Instantiating ν with ρ or $\hat\rho$, the results follow. □

With this tiny handle on the structure of a set of centers P satisfying $\phi_f(x;P) \le c$, the proof of Lemma 3.5 follows.

Proof of Lemma 3.5. Throughout both sections, let B_0 and C_0 be as defined in Lemma B.2; it follows by Lemma B.2, with probability at least $1-\delta$, that $P \in \mathcal{H}_f(\rho;c,k) \cup \mathcal{H}_f(\hat\rho;c,k)$ implies $P \cap C_0 \neq \emptyset$. Henceforth discard this failure event, and fix any $P \in \mathcal{H}_f(\rho;c,k) \cup \mathcal{H}_f(\hat\rho;c,k)$.

1. Since $P \cap C_0 \neq \emptyset$, fix some $p_0 \in P \cap C_0$. Since $B \supseteq C_0$ by definition, it follows, for every $x \in B^c$ that

$$\begin{aligned}\phi_f(x;P) &= \min_{p\in P} \mathsf{B}_f(x,p) \le r_2\|x-p_0\|^2 \le r_2(\|x-\mathbb{E}_\rho(X)\| + \|p_0 - \mathbb{E}_\rho(X)\|)^2 \\ &\le 4r_2\|x-\mathbb{E}_\rho(X)\|^2 = u(x).\end{aligned}$$

Additionally,

$$\ell(x) = 0 \le \min_{p\in P} r_1\|x-p\|^2 \le \phi_f(x;P),$$

meaning u and ℓ properly bracket $Z_\ell = Z_u$ over B^c; what remains is to control their mass over B^c.

Since $\ell = 0$,

$$\left|\int_{B^c} \ell(x) d\hat\rho(x)\right| = \left|\int_{B^c} \ell(x) d\rho(x)\right| = 0 < \epsilon.$$

Next, for u with respect to ρ, the result follows from the definition of u together with Lemma A.6 (using the map $\tau(x) = x - \mathbb{E}_\rho(X)$ together with exponent 2).

Lastly, to control u with respect to $\hat\rho$, note that $p' \le p/2 - 1$ means $\tilde p := 2(p'+1) \le p$, and thus the map $\tau(x) := \|x - \mathbb{E}_\rho(X)\|^2$ has order-$\tilde p$ moment bound M. Thus, by Lemma A.7 and the triangle inequality,

$$\left|\int_{B^c} u(x) d\hat\rho(x)\right| \le \epsilon + \sqrt{\frac{M' e p'}{2m}} \left(\frac{2}{\delta}\right)^{1/p'} = \epsilon_{\hat\rho}.$$

2. Throughout this proof, let ν denote either ρ or $\hat\rho$; the above established

$$\left|\int_{B^c} u(x) d\nu(x)\right| \le \epsilon_\nu,$$

where in the case of $\nu = \hat\rho$, this statement holds with probability $1-\delta$; henceforth discard this failure event, and thus the statement holds for both cases.

By definition of C, for any $p \in C^c$ and $x \in B$,

$$\begin{aligned}\mathsf{B}_f(x,p) &\ge r_1\|x-p\|^2 \ge r_1\left(\sqrt{r_2/r_1}\left((2M)^{1/p} + \sqrt{4c/r_1} + R_B\right)\right)^2 \\ &= r_2\left((2M)^{1/p} + \sqrt{4c/r_1} + R_B\right)^2.\end{aligned}$$

On the other hand, fixing any $p_0 \in P \cap C_0$ (which was guaranteed to exist at the start of this proof), since $C_0 \subseteq C$,

$$\sup_{x\in B} \phi_f(x; P\cap C) \le \sup_{x\in B} r_2\|x-p_0\|^2 \le r_2\left((2M)^{1/p} + \sqrt{4c/r_1} + R_B\right)^2.$$

Consequently, no element of B is closer to an element of $P \cap C$ than to any element of $P \setminus C$. As such,

$$\int \phi_f(x;P) d\nu(x) \ge \int_B \phi_f(x;P) d\nu(x) + \int_{B^c} \ell(x) d\nu(x) = \int_B \phi_f(x; P\cap C) d\nu(x).$$

(Note here that $\ell(x) = 0$ was used directly, rather than the ϵ provided by outer covering; in the case of Gaussian mixtures, both bracket elements are nonzero, and ϵ will be used.) This establishes one direction of the bound.

For the other direction, note that adding centers back in only decreases cost (because $\min_{p\in P\cap C}$ is replaced with $\min_{p\in P}$), and thus recalling the properties of the outer bracket element u established above,

$$\begin{aligned}\int_B \phi_f(x;P\cap C)d\nu(x) &= \int \phi_f(x;P\cap C)d\nu(x) - \int_{B^c}\phi_f(x;P\cap C)d\nu(x)\\ &\geq \int \phi_f(x;P\cap C)d\nu(x) - \int_{B^c} u(x)d\nu(x)\\ &\geq \int \phi_f(x;P)d\nu(x) - \epsilon_\nu,\end{aligned}$$

which gives the result(s).

□

B.2 Covering Properties

The next step is to control the deviations over the bounded portion; this is achieved via uniform covers, as developed in this subsection.

First, another basic fact about Bregman divergences.

Lemma B.3. *Let differentiable convex function f be given with Lipschitz gradient constant r_2 with respect to norm $\|\cdot\|$, and let B_f be the corresponding Bregman divergence. For any $\{x,y,z\}\subseteq\mathcal{X}$,*

$$\mathsf{B}_f(x,z) \leq \mathsf{B}_f(x,y) + \mathsf{B}_f(y,z) + r_2\|x-y\|\|y-z\|.$$

Similarly, given finite sets $Y\subseteq\mathcal{X}$ and $Z\subseteq\mathcal{X}$, and letting $Y(p)$ and $Z(p)$ respectively select (any) closest point in Y and Z to p according to B_f, meaning

$$Y(p) := \arg\min_{y\in Y}\mathsf{B}_f(y,p) \qquad \text{and} \qquad Z(p) := \arg\min_{z\in Z}\mathsf{B}_f(z,p),$$

then

$$\min_{z\in Z}\mathsf{B}_f(x,z) \leq \min_{y\in Y}\mathsf{B}_f(x,y) + \mathsf{B}_f(Y(x),Z(Y(x))) + r_2\|x-Y(x)\|\|Y(x)-Z(Y(x))\|.$$

Proof. By definition of B_f, properties of dual norms, and the Lipschitz gradient property,

$$\begin{aligned}\mathsf{B}_f(x,z) - \mathsf{B}_f(x,y) - \mathsf{B}(y,z) &= f(x) - f(z) - f(x) + f(y) - f(y) + f(z)\\ &\qquad - \langle\nabla f(z), x-z\rangle + \langle\nabla f(y), x-y\rangle + \langle\nabla f(z), y-z\rangle\\ &= \langle\nabla f(y) - \nabla f(z), x-y\rangle\\ &\leq \|\nabla f(y) - \nabla f(z)\|_*\|x-y\|\\ &\leq r_2\|y-z\|\|x-y\|;\end{aligned}$$

rearranging this inequality gives the first statement.

The second statement follows the first instantiated with $y = Y(x)$ and $z = Z(Y(x))$, since

$$\begin{aligned}\min_{z\in Z}\mathsf{B}_f(x,z) &\leq \mathsf{B}_f(x,Z(Y(x)))\\ &\leq \mathsf{B}_f(x,Y(x)) + \mathsf{B}_f(Y(x),Z(Y(x))) + r_2\|x-Y(x)\|\|Y(x)-Z(Y(x))\|,\end{aligned}$$

and using $\mathsf{B}_f(x,Y(x)) = \min_{y\in Y}\mathsf{B}_f(x,y)$. □

The covers will be based on norm balls; the following estimate is useful.

Lemma B.4. *If $\|\cdot\|$ is an l_p norm over $\mathbb{R}^d$, then the ball of radius R admits a cover $\mathcal{N}$ with size*

$$|\mathcal{N}| \leq \left(1 + \frac{2Rd}{\tau}\right)^d.$$

Proof. It suffices to grid the B with l_∞ balls centered at grid points at scale τ/d; the result follows since the l_∞ balls of radius τ/d are contained in l_p balls of radius τ for all $p\geq 1$. □

The uniform covering result is as follows.

Lemma B.5. *Let scale $\epsilon > 0$, ball $B := \{x \in \mathbb{R}^d : \|x - \mathbb{E}(X)\| \le R\}$, parameter set $Z := \{x \in \mathbb{R}^d : \|x - \mathbb{E}(X)\| \le R_2\}$, and differentiable convex function f with Lipschitz gradient parameter r_2 with respect to norm $\|\cdot\|$ be given. Define resolution parameter*

$$\tau := \min\left\{\sqrt{\frac{\epsilon}{2r_2}}, \frac{\epsilon}{2(R_2 + R)r_2}\right\},$$

and let $\mathcal{N}$ be set of centers for a cover of Z by $\|\cdot\|$-balls of radius τ (see Lemma B.4 for an estimate when $\|\cdot\|$ is an l_p norm). It follows that there exists a uniform cover $\mathcal{F}$ at scale ϵ with cardinality $|\mathcal{N}|^k$, meaning for any collection $P = \{p_i\}_{i=1}^l$ with $p_i \in Z$ and $l \le k$, there is a cover element Q with

$$\sup_{x \in B} \left| \min_{p \in P} \mathsf{B}_f(x, p) - \min_{q \in Q} \mathsf{B}_f(x, q) \right| \le \epsilon.$$

Proof. Given a collection P as specified, choose Q so that for every $p \in P$, there is $q \in Q$ with $\|p - q\| \le \tau$, and vice versa. By Lemma B.3 (and using the notation therein), for any $x \in B^c$,

$$\begin{aligned}
\min_{p \in P} \mathsf{B}_f(x, p) &\le \min_{q \in Q} \mathsf{B}_f(x, q) + \mathsf{B}_f(Q(x), P(Q(x))) + r_2\|x - Q(x)\|\|Q(x) - P(Q(x))\| \\
&\le \min_{q \in Q} \mathsf{B}_f(x, q) + r_2\tau^2 + r_2\tau(R + R_2) \\
&\le \min_{q \in Q} \mathsf{B}_f(x, q) + \epsilon;
\end{aligned}$$

the reverse inequality holds for the same reason, and the result follows. □

B.3 Proof of Theorem 3.2 and Corollary 3.1

First, the proof of the general rate for $\mathcal{H}_f(\nu; c, k)$.

Proof of Theorem 3.2. For convenience, define $M' = 2^{p'}\epsilon$. By Lemma B.5, let $\mathcal{N}$ be a cover of the set C, whereby every set of centers $P \subseteq C$ with $|P| \le k$ has a cover element $Q \in \mathcal{N}^k$ with

$$\sup_{x \in B} \left| \min_{p \in P} \mathsf{B}_f(x, p) - \min_{q \in Q} \mathsf{B}_f(x, q) \right| \le \epsilon; \tag{B.6}$$

when $\|\cdot\|$ is an l_p norm, Lemma B.4 provides the stated estimate of its size. Since $B \subseteq C$ and

$$\sup_{x \in B} \sup_{p \in C} \mathsf{B}_f(x, p) \le r_2 \sup_{x \in B} \sup_{p \in C} \|x - p\|^2 \le 4r_2 R_C^2,$$

it follows by Hoeffding's inequality and a union bound over $\mathcal{N}^k$ that with probability at least $1 - \delta$,

$$\sup_{Q \in \mathcal{N}^k} \left| \int_B \phi(x; Q) d\hat{\rho}(x) - \int_B \phi(x; Q) d\rho(x) \right| \le 4r_2 R_C^2 \sqrt{\frac{1}{2m} \ln\left(\frac{2|\mathcal{N}|^k}{\delta}\right)}. \tag{B.7}$$

For the remainder of this proof, discard the corresponding failure event.

Now let any $P \in \mathcal{H}_f(\rho; c, k) \cup \mathcal{H}_f(\hat{\rho}; c, k)$ be given, and let $Q \in \mathcal{N}^k$ be a cover element satisfying eq. (B.6) for $P \cap C$. By eq. (B.6), eq. (B.7), and Lemma 3.5 (and thus discarding an additional failure event having probability 2δ),

$$\begin{aligned}\left|\int \phi_f(x;P)d\rho(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right| &\leq \left|\int \phi_f(x;P)d\rho(x) - \int_B \phi_f(x;P\cap C)d\rho(x)\right| \\ &+ \left|\int_B \phi_f(x;P\cap C)d\rho(x) - \int_B \phi_f(x;Q)d\rho(x)\right| \\ &+ \left|\int_B \phi_f(x;Q)d\rho(x) - \int_B \phi_f(x;Q)d\hat{\rho}(x)\right| \\ &+ \left|\int_B \phi_f(x;Q)d\hat{\rho}(x) - \int_B \phi_f(x;P\cap C)d\hat{\rho}(x)\right| \\ &+ \left|\int_B \phi_f(x;P\cap C)d\hat{\rho}(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right| \\ &\leq 2\epsilon + 4r_2R_C^2\sqrt{\frac{1}{2m}\ln\left(\frac{2|\mathcal{N}|^k}{\delta}\right)} + \epsilon_\rho + \epsilon_{\hat{\rho}},\end{aligned}$$

and the result follows by unwrapping the definitions of ϵ_ρ and $\epsilon_{\hat{\rho}}$ from Lemma 3.5, and $M' = 2^{p'}\epsilon$ as above. □

The more concrete bound for the k-means cost is proved as follows.

Proof of Corollary 3.1. Set

$$\epsilon := m^{-1/2+1/p}, \qquad p' := p/4, \qquad M' := 2^{p'}\epsilon = 2^{p/4}m^{-1/2+1/p},$$

and recall $f(x) := \|x\|_2^2$ has convexity constants $r_1 = r_2 = 2$. Since

$$m = \sqrt{m}\sqrt{m} \geq \frac{p\sqrt{m}}{2^{p/4+2}e} \geq \frac{p'm^{1/2-1/p}}{2^{p'}e} = \frac{p'}{M'e}$$

and $p' = p/2 - p/4 \leq p/2 - 1$, the conditions for Theorem 3.2 are met, and thus with probability at least $1-\delta$,

$$\left|\int \phi_f(x;P)d\rho(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right| \leq 4\epsilon + 4R_C^2\sqrt{\frac{1}{2m}\ln\left(\frac{2|\mathcal{N}|^k}{\delta}\right)} + \sqrt{\frac{2^{p/4}ep\epsilon}{8m}}\left(\frac{2}{\delta}\right)^{4/p},$$

where

$$\begin{aligned} R_C &:= (2M)^{1/p} + \sqrt{2c} + 2R_B, \\ R_B &:= \max\left\{(2M)^{1/p} + \sqrt{2c}, \max_{i\in[p']}(M/\epsilon)^{1/(p-2i)}\right\}, \\ |\mathcal{N}| &\leq \left(1 + \frac{2R_Cd}{\tau}\right)^d, \\ \tau &:= \min\left\{\sqrt{\frac{\epsilon}{4}}, \frac{\epsilon}{4(R_B+R_C)}\right\}. \end{aligned}$$

To simplify these quantities, since $\epsilon \leq 1$, the term $1/\epsilon^{1/(p-2i)}$, as i ranges between 1 and $p-2p'$, is maximized at $i = 1/(p-2p') = 2/p$. Therefore, by choice of M_1 and ϵ,

$$\begin{aligned} R_B &\leq c_1 + (M/\epsilon)^{1/(p-2)} + (M/\epsilon)^{1/(p-2p')} \leq c_1 + (M^{1/(p-2)} + M^{1/(p-2p')})/\epsilon^{2/p} \\ &= c_1 + M_1m^{1/p-2/p^2}. \end{aligned}$$

Consequently,

$$R_C = c_1 + 2R_B \leq 3c_1 + 2M_1m^{1/p-2/p^2} \qquad \text{and} \qquad R_C^2 \leq 18c_1^2 + 8M_1^2m^{2/p-4/p^2}.$$

This entails

$$\begin{aligned}\frac{2R_C d}{\tau} &\le 2R_C d\left(2m^{1/4-1/(2p)} + 4(R_B + R_C)m^{1/2-1/p}\right)\\ &\le 8d\left((3c_1 + 2M_1 m^{1/p-2/p^2})m^{1/4-1/(2p)} + (36c_1^2 + 16M_1^2 m^{2/p-4/p^2})m^{1/2-1/p}\right)\\ &\le 288dm(c_1 + c_1^2 + M_1 + M_1^2).\end{aligned}$$

Secondly,

$$\frac{R_C^2}{\sqrt{m}} \le (18c_1^2 + 8M_1^2 m^{2/p-4/p^2})m^{-1/2} \le m^{\min\{1/4,-1/2+2/p\}}(18c_1^2 + 8M_1^2).$$

The last term is direct, since

$$\sqrt{\epsilon/m} = m^{-1/4+1/(2p)-1/2} = m^{-1/2+1/(2p)}m^{-1/4}.$$

Combining these pieces, the result follows. ☐

C Deferred Material from Section 4

First, the deferred proof that outer brackets give rise to clamps.

Proof of Proposition 4.3. Throughout this proof, let ν refer to either ρ or $\hat{\rho}$, with ϵ_ν similarly referring to either ϵ_ρ or $\epsilon_{\hat{\rho}}$. Let $P \in \mathcal{H}_f(\rho; c, k) \cup \mathcal{H}_f(\hat{\rho}; c, k)$ be given.

One direction is direct:

$$\begin{aligned}\int \phi_f(x; P)d\nu(x) &\ge \int \phi_f(x; P \cap C)d\nu(x)\\ &\ge \int \min\{\phi_f(x; P \cap C), R\}d\nu(x).\end{aligned}$$

For the second direction, with probability at least $1 - \delta$, Lemma B.2 grants the existence of $p' \in P \cap C_0 \subseteq P \cap C$. Consequently, for any $x \in B$,

$$\begin{aligned}\min_{p\in P} \mathsf{B}_f(x, p) &\le \min_{p\in P\cap C} \mathsf{B}_f(x, p) \le \mathsf{B}_f(x, p')\\ &\le r_2\|x - p'\|^2 \le 2r_2\left(\|x - \mathbb{E}_\rho(X)\|^2 + \|p' - \mathbb{E}_\rho(X)\|^2\right)\\ &\le R;\end{aligned}$$

in other words, if $x \in B$, then $\min\{\phi_f(x; P \cap C), R\} = \phi_f(x; P \cap C)$. Combining this with the last part of Lemma 3.5.

$$\begin{aligned}\int \min\{\phi_f(x; P \cap C), R\}d\nu(x) &\ge \int_B \min\{\phi_f(x; P \cap C), R\}d\nu(x)\\ &\ge \int_B \phi_f(x; P \cap C)d\nu(x)\\ &\ge \int \phi_f(x; P)d\nu(x) - \epsilon_\nu.\end{aligned}$$

☐

The proof of Theorem 4.4 will depend on the following uniform covering property of the clamped cost (which mirrors Lemma B.5 for the unclamped cost).

Lemma C.1. *Let scale $\epsilon > 0$, clamping value R_3, parameter set C contained within a $\|\cdot\|$-ball of some radius R_2, and differentiable convex function f with Lipschitz gradient parameter r_2 and strong convexity modulus r_1 with respect to norm $\|\cdot\|$ be given. Define resolution parameter*

$$\tau := \min\left\{\sqrt{\frac{\epsilon}{2r_2}}, \frac{r_1\epsilon}{2r_2R_3}\right\},$$

and let $\mathcal{N}$ be set of centers for a cover of C by $\|\cdot\|$-balls of radius τ (see Lemma B.4 for an estimate when $\|\cdot\|$ is an l_p norm). It follows that there exists a uniform cover $\mathcal{F}$ at scale ϵ with cardinality $|\mathcal{N}|^k$, meaning for any collection $P = \{p_i\}_{i=1}^l$ with $p_i \in C$ and $l \le k$, there is a cover element Q with

$$\sup_x \left| \min\left\{R_3, \min_{p\in P} \mathsf{B}_f(x,p)\right\} - \min\left\{R_3, \min_{q\in Q} \mathsf{B}_f(x,q)\right\}\right| \le \epsilon.$$

Proof. Given a collection P as specified, choose Q so that for every $p \in P$, there is $q \in Q$ with $\|p - q\| \le \tau$, and vice versa.

First suppose $\min_{q\in Q} \mathsf{B}_f(x,q) \ge R_3$; then

$$\min\left\{R_3, \min_{p\in P} \mathsf{B}_f(x,p)\right\} \le R_3 = \min\left\{R_3, \min_{q\in Q} \mathsf{B}_f(x,q)\right\}$$

as desired.

Otherwise, $\min_{q\in Q} \mathsf{B}_f(x,q) < R_3$, which by the sandwiching property (cf. Lemma B.1) means

$$r_1\|x - Q(x)\| \le \mathsf{B}_f(x, Q(x)) < R_3.$$

By Lemma B.3,

$$\begin{aligned}
\min\left\{R_3, \min_{p\in P} \mathsf{B}_f(x,p)\right\} &\le \min\left\{R_3, \min_{q\in Q} \mathsf{B}_f(x,q) + \mathsf{B}_f(Q(x), P(Q(x))) + r_2\|x - Q(x)\|\|Q(x) - P(Q(x))\|\right\} \\
&\le \min\left\{R_3, \min_{q\in Q} \mathsf{B}_f(x,q) + r_2\tau^2 + r_2\tau\|x - Q(x)\|\right\} \\
&\le \min\left\{R_3, \min_{q\in Q} \mathsf{B}_f(x,q) + r_2\tau^2 + \frac{r_2 R_3}{r_1}\tau\right\} \\
&\le \min\left\{R_3, \min_{q\in Q} \mathsf{B}_f(x,q)\right\} + \epsilon.
\end{aligned}$$

The reverse inequality is analogous. □

The proof of Theorem 4.4 follows.

Proof of Theorem 4.4. This proof is a minor alteration of the proof of Theorem 3.2.

By Lemma C.1, let $\mathcal{N}$ be a cover of the set C, whereby every set of centers $P \subseteq C$ with $|P| \le k$ has a cover element $Q \in \mathcal{N}^k$ with

$$\sup_x |\min\{\phi_f(x;P), R\} - \min\{\phi_f(x;Q), R\}| \le \epsilon; \tag{C.2}$$

when $\|\cdot\|$ is an l_p norm, Lemma B.4 provides the stated estimate of its size. Since $\min\{\phi_f(x;Q), R\} \in [0, R]$, it follows by Hoeffding's inequality and a union bound over $\mathcal{N}^k$ that with probability at least $1-\delta$,

$$\sup_{Q\in\mathcal{N}^k} \left| \int_B \phi_f(x;Q)d\hat\rho(x) - \int_B \phi_f(x;Q)d\rho(x)\right| \le R\sqrt{\frac{1}{2m}\ln\left(\frac{2|\mathcal{N}|^k}{\delta}\right)}. \tag{C.3}$$

For the remainder of this proof, discard the corresponding failure event.

Now let any $P \in Z$ be given, and let $Q \in \mathcal{N}^k$ be a cover element satisfying eq. (C.2) for $P \cap C$. By eq. (C.2), eq. (C.3), and lastly by the definition of clamp,

$$\begin{aligned}\left|\int \phi_f(x;P)d\rho(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right| &\leq \left|\int \phi_f(x;P)d\rho(x) - \int \min\{\phi_f(x;P\cap C),R\}d\rho(x)\right| \\ &\quad + \left|\int \min\{\phi_f(x;P\cap C),R\}d\rho(x) - \int \min\{\phi_f(x;Q),R\}d\rho(x)\right| \\ &\quad + \left|\int \min\{\phi_f(x;Q),R\}d\rho(x) - \int \min\{\phi_f(x;Q),R\}d\hat{\rho}(x)\right| \\ &\quad + \left|\int \min\{\phi_f(x;Q),R\}d\hat{\rho}(x) - \int \min\{\phi_f(x;P\cap C),R\}d\hat{\rho}(x)\right| \\ &\quad + \left|\int \min\{\phi_f(x;P\cap C),R\}d\hat{\rho}(x) - \int \phi_f(x;P)d\hat{\rho}(x)\right| \\ &\leq 2\epsilon + \epsilon_\rho + \epsilon_{\hat{\rho}} + R^2\sqrt{\frac{1}{2m}\ln\left(\frac{2|\mathcal{N}|^k}{\delta}\right)}.\end{aligned}$$

□

D Deferred Material from Section 5

The following notation for restricting a Gaussian mixture to a certain set of means will be convenient throughout this section.

Definition D.1. *Given a Gaussian mixture with parameters (α, Θ) (where $\alpha = \{\alpha_i\}_{i=1}^k$ and $\Theta = \{\theta_i\}_{i=1}^k = \{(\mu_i, \Sigma_i)\}_{i=1}^k$), and a set of means $B \subseteq \mathbb{R}^d$, define*

$$(\alpha,\Theta) \sqcap B := \{(\{\alpha_i\}_{i\in I}, \{(\mu_i,\Sigma_i)\}_{i\in I}) : I = \{1 \leq i \leq k : \mu_i \in B\}\}.$$

(Note that potentially $\sum_{i\in I}\alpha_i < 1$, and thus the terminology partial Gaussian mixture *is sometimes employed.)*

D.1 Constructing an Outer Bracket

The first step is to show that pushing a mean far away from a region will rapidly decrease its density there, which is immediate from the condition $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$.

Lemma D.2. *Let probability measure ρ, accuracy $\epsilon > 0$, covariance lower bound $0 < \sigma_1 \leq \sigma_2$, and radius R with corresponding l_2 ball $B := \{x \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\|_2 \leq R\}$ be given. Define*

$$\begin{aligned}R_1 &:= \sqrt{2\sigma_2 \ln\left(\frac{1}{(2\pi\sigma_1)^{d/2}\epsilon^2}\right)} \\ R_2 &:= R + R_1, \\ B_2 &:= \{\mu \in \mathbb{R}^d : \|\mu - \mathbb{E}_\rho(X)\|_2 \leq R_2\}.\end{aligned}$$

If $\theta = (\mu, \Sigma)$ is the parameterization of a Gaussian density p_θ with $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$ but $\mu \notin B_2$, then $p_\theta(x) < \epsilon$ for every $x \in B$.

Proof. Let Gaussian parameters $\theta = (\mu, \Sigma)$ be given with $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$, but $\mu \notin B_2$. By the definition of B_2, for any $x \in B_1$,

$$p_\theta(x) < (2\pi\sigma_1)^{-d/2}\exp(-R_1^2/(2\sigma_2)) = \epsilon.$$

□

The upper component of the outer bracket will be constructed first (and indeed used in the construction of the lower component).

Lemma D.3. *Let probability measure ρ with order-p moment bound with respect to $\|\cdot\|_2$, target accuracy $\epsilon > 0$, and covariance lower bound $0 < \sigma_1$ be given. Define*

$$\begin{aligned} p_{\max} &:= (2\pi\sigma_1)^{-d/2}, \\ u(x) &:= \ln(p_{\max}), \\ R_u &:= (M|\ln(p_{\max})|/\epsilon)^{1/p}, \\ B_u &:= \left\{x \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\|_2 \le R_u\right\}. \end{aligned}$$

If p_θ denotes a Gaussian density with parameters $\theta = (\mu, \Sigma)$ satisfying $\Sigma \succeq \sigma_1 I$, then $p_\theta \le u$ everywhere. Additionally,

$$\left|\int_{B_u^c} u(x) d\rho(x)\right| \le \int_{B_u^c} |u(x)| d\rho(x) \le \epsilon,$$

and with probability at least $1 - \delta$ over the draw of m points from ρ,

$$\left|\int_{B_u^c} u(x) d\hat\rho(x)\right| \le \int_{B_u^c} |u(x)| d\hat\rho(x) \le \epsilon + |\ln(p_{\max})| \sqrt{\frac{1}{2m} \ln\left(\frac{1}{\delta}\right)}.$$

(That is to say, u is the upper part of an outer bracket for all Gaussians (and mixtures thereof) where each covariance Σ satisfies $\Sigma \succeq \sigma_1 I$.)

Proof. Let p_θ with $\theta = (\mu, \Sigma)$ satisfying $\Sigma \succeq \sigma_1 I$ be given. Then

$$p_\theta(x) \le \frac{1}{\sqrt{(2\pi)^d \sigma_1^d}} \exp(\ 0\ ) = p_{\max}.$$

Next, given the form of B_u, if $\ln(p_{\max}) = 0$, the result is immediate, thus suppose $\ln(p_{\max}) \neq 0$; Lemma A.2 provides that $\rho(B_u) \ge 1 - \epsilon/|\ln(p_{\max})|$, whereby

$$\left|\int_{B_u^c} u(x) d\rho(x)\right| \le \int_{B_u^c} |u(x)| d\rho(x) = |\ln(p_{\max})| \rho(B_u^c) \le \epsilon.$$

For the finite sample guarantee, by Hoeffding's inequality,

$$\hat\rho(B_u^c) \le \rho(B_u^c) + \sqrt{\frac{1}{2m} \ln\left(\frac{1}{\delta}\right)} \le \frac{\epsilon}{|\ln(p_{\max})|} + \sqrt{\frac{1}{2m} \ln\left(\frac{1}{\delta}\right)},$$

which gives the result similarly to the case for ρ. ∎

From, here, a tiny control on $\mathcal{S}_{\text{mog}}(\nu; c, k, \sigma_1, \sigma_2)$ emerges, analogous to Lemma B.2 for $\mathcal{H}_f(\nu; c, k)$.

Lemma D.4. *Let covariance bounds $0 < \sigma_1 \le \sigma_2$, cost $c \le 1/2$, and probability measure ρ with order-p moment bound M with respect to $\|\cdot\|_2$ be given. Define*

$$\begin{aligned} p_{\max} &:= (2\pi\sigma_1)^{-d/2}, \\ R_3 &:= (2M|\ln(p_{\max})|)^{1/p}, \\ R_4 &:= (2M)^{1/p}, \\ R_5 &:= \sqrt{2\sigma_2 \left(\ln\left(\frac{8e}{(2\pi\sigma_1)^{d/2}}\right) - 4c\right)}. \\ R_6 &:= \max\{R_3, R_4\} + R_5. \\ B_6 &:= \{x \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\|_2 \le R_6\}. \end{aligned}$$

Suppose

$$m \ge 2\ln(1/\delta) \max\{4, |\ln(p_{\max})|^2\}.$$

With probability at least $1 - 2\delta$, given any $(\alpha, \Theta) \in \mathcal{S}_{\text{mog}}(\rho; c, k, \sigma_1, \sigma_2) \cup \mathcal{S}_{\text{mog}}(\hat\rho; c, k, \sigma_1, \sigma_2)$, the restriction $(\alpha', \Theta') = (\alpha, \Theta) \sqcap B_6$ is nonempty, and moreover satisfies $\sum_{\alpha_i \in \alpha'} \alpha_i \ge$ $\exp(4c)/(8ep_{\max})$.

Proof. Define

$$B_3 := \left\{x \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\|_2 \le \max\{R_3, R_4\}\right\}.$$

Since B_3 has radius at least R_4, Lemma A.2 provides

$$\rho(B_3) \ge 1/2,$$

and Hoeffding's inequality and the lower bound on m provide (with probability at least $1 - \delta$)

$$\hat{\rho}(B_3) \ge \frac{1}{2} - \sqrt{\frac{2}{m}\ln\left(\frac{1}{\delta}\right)} > \frac{1}{4}.$$

Additionally, since B_3 also has radius at least R_3, by Lemma D.3, the choice of B_3, and the lower bound on m, and letting B_4 denote the ball of radius R_3,

$$\left|\int_{B_3^c} u d\rho\right| \le \int_{B_4^c} |u| d\rho \le \int_{B_4^c} |u| d\rho \le 1/2 \qquad \text{and} \qquad \left|\int_{B_3^c} u d\hat{\rho}\right| < 1,$$

where the statement for $\hat{\rho}$ is with probability at least $1-\delta$. For the remainder of the proof, let ν refer to either ρ or $\hat{\rho}$, and discard the 2δ failure probability of either of the above two events.

For convenience, define $p_0 := \exp(4c)/(8e)$, whereby

$$R_5 = \sqrt{2\sigma_2 \ln\left(\frac{1}{p_0(2\pi\sigma_1)^{d/2}}\right)}.$$

By Lemma D.2, any Gaussian parameters $\theta = (\mu, \Sigma)$ with $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$ and $\mu \notin B_6$ have $p_\theta(x) < p_0$ everywhere on B_3. As such, a mixture (α, Θ) where each $\theta_i \in \Theta$ satisfies these conditions also satisfies

$$\begin{aligned}\int \ln\left(\sum_i \alpha_i p_{\theta_i}\right) d\nu &\le \int_{B_3} \ln\left(\sum_{(\alpha_i,\theta_i)\in(\alpha,\Theta)\sqcap B_6} \alpha_i p_{\theta_i} + \sum_{(\alpha_i,\theta_i)\notin(\alpha,\Theta)\sqcap B_6} \alpha_i p_{\theta_i}\right) d\nu + \int_{B_3^c} u d\nu \\ &< \ln\left(\sum_{(\alpha_i,\theta_i)\in(\alpha,\Theta)\sqcap B_6} \alpha_i p_{\max} + \sum_{\alpha_i,\theta_i)\notin(\alpha,\Theta)\sqcap B_6} \alpha_i p_0\right) \nu(B_3) + 1\end{aligned}$$

Suppose contradictorily that $(\alpha, \Theta) \sqcap B_6 = \emptyset$ or $\sum_{(\alpha_i,\theta_i)\in(\alpha,\Theta)\sqcap B_6} \alpha_i < p_0/p_{\max}$. But $c \le 1/2$ implies $p_0 \le 1/2$ and so $\ln(2p_0) \le 0$, thus $\ln(2p_0)\nu(B_3) \le \ln(2p_0)/4$ which together with $p_0 \le \exp(4c)/(8e)$ and the above display gives

$$\int \ln\left(\sum_i \alpha_i p_{\theta_i}\right) d\nu < \ln(2p_0)/4 + 1 \le c,$$

which contradicts $\mathbb{E}_\nu(\phi_{\mathrm{g}}(X; (\alpha, \Theta))) \ge c$. $\square$

Now that significant weight can be shown to reside in some restricted region, the outer bracket and its basic properties follow (i.e., the analog to Lemma 3.5).

Lemma D.5. *Let target accuracy $0 < \epsilon \le 1$, covariance bounds $0 < \sigma_1 \le \sigma_2$ with $\sigma_1 \le 1$, target cost c, confidence parameter $\delta \in (0, 1]$, probability measure ρ with order-p moment bound M with respect to $\|\cdot\|_2$ with $p \ge 4$, and integer $1 \le p' \le p/2 - 1$. Define first the basic quantities*

$$\begin{aligned}M' &:= 2^{p'}\epsilon, \\ p_{\max} &:= (2\pi\sigma_1)^{-d/2}, \\ R_6 &:= (2M|\ln(p_{\max})|)^{1/p} + (2M)^{1/p} + \sqrt{2\sigma_2\left(\ln\left(\frac{8e}{(2\pi\sigma_1)^{d/2}}\right) - 4c\right)}, \\ B_6 &:= \{x \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\|_2 \le R_6\|\}.\end{aligned}$$

Additionally define the outer bracket elements

$$
\begin{aligned}
Z_\ell &:= \left\{ (\alpha,\Theta) : \forall(\alpha_i,(\mu_i,\theta_i)) \in (\alpha,\Theta) \centerdot \mu_i \in B_6, \sigma_1 I \preceq \Sigma \preceq \sigma_2 I, \sum_i \alpha_i \geq \exp(4c)/(8ep_{\max}) \right\},\\
c_\ell &:= 4c - \ln(8ep_{\max}) - \frac{d}{2}\ln(2\pi\sigma_2),\\
\ell(x) &:= c_\ell - \frac{2}{\sigma_1}\|x - \mathbb{E}_\rho(X)\|_2^2,\\
u(x) &:= \ln(p_{\max}),\\
\epsilon_{\hat\rho} &:= \epsilon + (|c_\ell| + |\ln(p_{\max})|)\sqrt{\frac{1}{2m}\ln\left(\frac{1}{\delta}\right)} + \sqrt{\frac{M'ep'}{2m}}\left(\frac{2}{\delta}\right)^{1/p'},\\
M_1 &:= (2M|c_\ell|)^{1/p} + (4M\sigma_1)^{1/(p-2)} + \max_{1\leq i\leq p'} M^{1/(p-2i)} + (M|\ln(p_{\max}))^{1/p},\\
R_B &= R_6 + M_1/\epsilon^{1/(p-2p')},\\
B &:= \{x \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\|_2 \leq R_B\}.
\end{aligned}
$$

The following statements hold with probability at least $1-4\delta$ *over a draw of size*

$$m \geq \max\left\{p'/(M'e), 8\ln(1/\delta), 2|\ln(p_{\max})|^2\ln(1/\delta)\right\}.$$

1. (u,ℓ) *is an outer bracket for* ρ *at scale* $\epsilon_\rho := \epsilon$ *with sets* $B_\ell := B_u := B$*, center set class* Z_ℓ *as above, and* $Z_u = \mathcal{S}_{\text{mog}}(\rho;\infty,k,\sigma_1,\sigma_2)$*. Additionally,* (u,ℓ) *is also an outer bracket for* $\hat\rho$ *at scale* $\epsilon_{\hat\rho}$ *with the same sets.*
2. *Define*

$$
\begin{aligned}
R_C &:= 1 + R_B(1+\sqrt{8\sigma_2/\sigma_1}) + \sqrt{4\sigma_2\ln(1/\epsilon)} + \sqrt{2\sigma_2\left(\ln\left(\frac{64e^2(2\pi\sigma_2)^d}{(2\pi)^d p_{\max}^4}\right) - 8c\right)},\\
C &:= \{\mu \in \mathbb{R}^d : \|x - \mathbb{E}_\rho(X)\|_2 \leq R_C\}.
\end{aligned}
$$

Every $(\alpha,\Theta) \in \mathcal{S}_{\text{mog}}(\rho;c,k,\sigma_1,\sigma_2) \cup \mathcal{S}_{\text{mog}}(\hat\rho;c,k,\sigma_1,\sigma_2)$ *satisfies* $\sum_{(\alpha_i,\theta_i)\in(\alpha,\Theta)\sqcap C}\alpha_i \geq \exp(4c)/(8ep_{\max})$*, and*

$$\left|\int \phi_{\text{g}}(x;(\alpha,\Theta))d\rho(x) - \int_B \phi_{\text{g}}(x;(\alpha,\Theta)\sqcap C)d\rho(x)\right| \leq \epsilon_\rho = 2\epsilon$$

and

$$\left|\int \phi_{\text{g}}(x;(\alpha,\Theta))d\hat\rho(x) - \int_B \phi_{\text{g}}(x;(\alpha,\Theta)\sqcap C)d\hat\rho(x)\right| \leq \epsilon + \epsilon_{\hat\rho}.$$

Proof of Lemma D.5. It is useful to first expand the choice of R_B, which was chosen large enough to carry a collection of other radii. In particular, since $\epsilon \leq 1$, then $1/\epsilon \geq 1$, and therefore $1/\epsilon^a \leq 1/\epsilon^b$ when $a \leq b$. As such, since $p' \leq p/2 - 1$,

$$
\begin{aligned}
R_B &= R_6 + M_1/\epsilon^{1/(p-2p')}\\
&= R_6 + \left((2M|c_\ell|)^{1/p} + (4M\sigma_1)^{1/(p-2)} + \max_{1\leq i\leq p'} M^{1/(p-2i)} + (M|\ln(p_{\max}))^{1/p}\right)/\epsilon^{1/(p-2p')}\\
&\geq R_6 + \left((2M|c_\ell|/\epsilon)^{1/p} + (4M\sigma_1/\epsilon)^{1/(p-2)} + \max_{1\leq i\leq p'}(M/\epsilon)^{1/(p-2i)} + (M|\ln(p_{\max})|/\epsilon)^{1/p}\right).
\end{aligned}
$$

Since every term is nonnegative, R_B dominates each individual term.

1. The upper bracket and its guarantees were provided by Lemma D.3; note that $\epsilon_{\hat\rho}$ is defined large enough to include the deviations there, and similarly $R_B \geq (M|\ln(p_{\max})|/\epsilon)^{1/p}$ means the B here is defined large enough to contain the B_u there; correspondingly, discard a failure event with probability mass at most δ.

Let the lower bracket be defined as in the statement; note that its properties are much more conservative as compared with the upper bracket. Let $(\alpha, \Theta) \in Z_\ell$ be given. For every $\theta_i = (\mu_i, \Sigma_i)$, $\|\mu_i - \mathbb{E}_\rho(X)\|_2 \le R_6$, whereas $R_B \ge R_6$ meaning $x \in B^c$ implies $\|x - \mathbb{E}_\rho(X)\|_2 \ge R_6$, so

$$\|x - \mu_i\|_2 \le \|x - \mathbb{E}_\rho(X)\|_2 + \|\mu_i - \mathbb{E}_\rho(X)\|_2 \le 2\|x - \mathbb{E}_\rho(X)\|_2,$$

which combined with $\sigma_1 I \preceq \Sigma_i \preceq \sigma_2 I$ gives

$$\begin{aligned}\ln\left(\sum_i \alpha_i p_{\theta_i}(x)\right) &\ge \ln\left(\sum_i \alpha_i \frac{1}{(2\pi\sigma_2)^{d/2}} \exp\left(-\frac{1}{2\sigma_1}\|x - \mu_i\|_2^2\right)\right)\\ &\ge \ln(p_0/p_{\max}) - \frac{d}{2}\ln(2\pi\sigma_2) - \frac{2}{\sigma_1}\|x - \mathbb{E}_\rho(X)\|_2^2\\ &= \ell(x),\end{aligned}$$

which is the dominance property.

Next come the integral properties of ℓ. By Lemma A.2 and since $R_B \ge (2M|c_\ell|/\epsilon)^{1/p}$,

$$\left|\int_{B^c} c_\ell d\rho\right| \le \int_{B^c} |c_\ell| d\rho \le \int_{B^c} |c_\ell| d\rho = \rho(B^c)|c_\ell| \le \epsilon/2.$$

Similarly, by Hoeffding's inequality, with probability at least $1 - \delta$,

$$\left|\int_{B_\ell^c} c_\ell d\hat{\rho}\right| \le \epsilon/2 + |c_\ell|\sqrt{\frac{1}{2m}\ln\left(\frac{1}{\delta}\right)}.$$

Now define

$$\ell_1(x) := -\frac{2}{\sigma_1}\|x - \mathbb{E}_\rho(X)\|_2^2 = \ell(x) - c_\ell.$$

By Lemma A.6 and since $R_B \ge (4\sigma_1 M/\epsilon)^{1/(p-2)}$,

$$\left|\int_{B^c} \ell_1 d\rho\right| \le \int_{B^c} |\ell_1| d\rho = \frac{2}{\sigma_1}\int_{B^c} \|x - \mathbb{E}_\rho(X)\|_2^2 d\rho(x) \le \epsilon/2.$$

Furthermore by Lemma A.7 and the above estimate, and since $R_B \ge \max_{1 \le i \le p'}(M/\epsilon)^{1/(p-2i)}$ (where the maximum is attained at one of the endpoints), then with probability at least $1 - \delta$

$$\left|\int_{B^c} \ell_1 d\hat{\rho}\right| \le \frac{\epsilon}{2} + \sqrt{\frac{M'ep'}{2m}}\left(\frac{2}{\delta}\right)^{1/p'}.$$

Unioning together the above failure probabilities, the general controls for $\ell = c_\ell + \ell_1$ follow by the triangle inequality and definition of $\epsilon_{\hat{\rho}}$.

2. Throughout the following, let ν denote either ρ or $\hat{\rho}$, and correspondingly let ϵ_ν respectively refer to ϵ_ρ or $\epsilon_{\hat{\rho}}$; let the above bracketing properties hold throughout (with events appropriately discarded for $\hat{\rho}$). Furthermore, for convenience, define

$$p_0 := \exp(4c)/(8e).$$

Let any (α, Θ) be given with $(\alpha, \Theta) \in \mathcal{S}_{\text{mog}}(\rho; c, k, \sigma_1, \sigma_2) \cup \mathcal{S}_{\text{mog}}(\hat{\rho}; c, k, \sigma_1, \sigma_2)$. Define the two index sets

$$\begin{aligned}I_C &:= \{i \in [k] : (\alpha_i, \theta_i) \in (\alpha, \Theta) \sqcap C\},\\ I_6 &:= \{i \in [k] : (\alpha_i, \theta_i) \in (\alpha, \Theta) \sqcap B_6\}.\end{aligned}$$

By Lemma D.4, with probability at least $1 - \delta$, $\sum_{i \in I_6} \alpha_i \ge p_0/p_{\max}$; henceforth discard the corresponding failure event, bringing the total discarded probability mass to 4δ.

To start, since $\ln(\cdot)$ is concave and thus $\ln(a+b) \le \ln(a) + b/a$ for any positive a, b,

$$\begin{aligned}\int \ln\left(\sum_i \alpha_i p_{\theta_i}(x)\right) d\nu(x) &\le \int_B \ln\left(\sum_i \alpha_i p_{\theta_i}(x)\right) d\nu(x) + \int_{B^c} u(x) d\nu(x) \\ &\le \int_B \ln\left(\sum_{i\in I_C} \alpha_i p_{\theta_i}(x)\right) d\nu(x) + \int_B \frac{\sum_{i\notin I_C} \alpha_i p_{\theta_i}(x)}{\sum_{i\in I_C} \alpha_i p_{\theta_i}(x)} d\nu(x) + \epsilon_\nu.\end{aligned}$$

In order to control the fraction, both the numerator and denominator will be uniformly controlled for every $x \in B$, whereby the result follows since ν is a probability measure (i.e., the integral is upper bounded with an upper bound on the numerator times $\nu(B) \le 1$ divided by a lower bound on the denominator).

For the purposes of controlling this fraction, define

$$\begin{aligned}p_1 &:= \frac{1}{(2\pi\sigma_2)^{d/2}} \exp\left(-\frac{R_B^2 + R_6^2}{\sigma}\right), \\ p_2 &:= \epsilon p_1 p_0 / p_{\max},\end{aligned}$$

Observe, by choice of R_C and since $\sigma_1 \le 1$, that

$$\begin{aligned}R_B + \sqrt{2\sigma_2 \ln\left(\frac{1}{p_2^2 (2\pi)^d \sigma_1^{d-1}}\right)} &\le R_B + \sqrt{2\sigma_2 \ln\left(\frac{64 e^2 p_{\max}^2 (2\pi\sigma_2)^d \exp(2(R_B^2 + R_6^2))}{\epsilon^2 \exp(8c)(2\pi)^d \sigma_1^d}\right)} \\ &\le R_B + \sqrt{2\sigma_2 \left(\ln\left(\frac{64 e^2 (2\pi\sigma_2)^d}{\epsilon^2 (2\pi)^d p_{\max}^4}\right) - 8c - 4R_B^2/\sigma\right)} \\ &\le R_B + \sqrt{2\sigma_2 \left(\ln\left(\frac{64 e^2 (2\pi\sigma_2)^d}{(2\pi)^d p_{\max}^4}\right) - 8c\right)} \\ &\quad + \sqrt{4\sigma_2 \ln(1/\epsilon)} + R_B\sqrt{8\sigma_2/\sigma_1} \\ &\le R_C.\end{aligned}$$

For the denominator, first note for every $x \in B$ and parameters $\theta = (\mu, \Sigma)$ with $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$ and $\mu \in B_6$ that

$$\begin{aligned}p_\theta(x) &\ge \frac{1}{(2\pi\sigma_2)^{d/2}} \exp\left(-\frac{1}{2\sigma_1}\|x-\mu\|_2^2\right) \\ &\ge \frac{1}{(2\pi\sigma_2)^{d/2}} \exp\left(-\frac{1}{2\sigma_1}\left(\|x - \mathbb{E}_\rho(X)\|_2 + \|\mathbb{E}_\rho(X) - \mu_i\|_2\right)^2\right) \\ &\ge p_1.\end{aligned}$$

Consequently, for $x \in B$,

$$\sum_{i\in I_C} \alpha_i p_i(x) \ge \sum_{i\in I_6} \alpha_i p_i(x) \ge p_1 \sum_{i\in I_6} \alpha_i \ge p_1 p_0 / p_{\max}.$$

For the numerator, by choice of C (as developed above with the definitions of p_1 and p_2) and an application of Lemma D.2, for p_i corresponding to $i \notin I_C$,

$$p_i(x) \le \epsilon p_1 p_0 / p_{\max} = p_2.$$

It follows that the fractional term is at most ϵ, which gives the first direction of the desired inequality.

To get the other direction, since $\sum_{i\in I_6} \alpha_i \ge p_0/p_{\max}$ due to Lemma D.4 as discussed above, it follows that $(\alpha, \Theta) \sqcap B_6 \in Z_\ell$, meaning the corresponding partial Gaussian mixture can be controlled by ℓ. As such, since $R_6 \le R_B$ thus $I_6 \subseteq I_C$, and since $\ln$ is nondecreasing,

$$\begin{aligned}\int_B \ln\left(\sum_{i\in I_C}\alpha_i p_i\right)d\nu &= \int \ln\left(\sum_{i\in I_C}\alpha_i p_i\right)d\nu - \int_{B^c} \ln\left(\sum_{i\in I_C}\alpha_i p_i\right)d\nu \\ &\le \int \ln\left(\sum_{i\in I_C}\alpha_i p_i\right)d\nu - \int_{B^c} \ln\left(\sum_{i\in I_6}\alpha_i p_i\right)d\nu \\ &\le \int \ln\left(\sum_{i\in I_C}\alpha_i p_i\right)d\nu - \int_{B^c} \ell d\nu \\ &\le \int \ln\left(\sum_{i\in I_C}\alpha_i p_i\right)d\nu + \epsilon_\nu \\ &\le \int \ln\left(\sum_{i}\alpha_i p_i\right)d\nu + \epsilon_\nu .\end{aligned}$$

□

D.2 Uniform Covering of Gaussian Mixtures

First, a helper lemma for covering covariance matrices.

Lemma D.6. *Let scale $\epsilon > 0$ and eigenvalue bounds $0 < \sigma_1 \le \sigma_2$ be given. There exists a subset $\mathcal{M}$ of the positive definite matrices satisfying $\sigma_1 I \preceq M \preceq \sigma_2 I$ so that*

$$|\mathcal{M}| \le (1 + 32\sigma_2/\epsilon)^{d^2}\left(\left(1 + \frac{\sigma_2 - \sigma_1}{\epsilon/2}\right)^d + \left(\frac{\ln(\sigma_2/\sigma_1)}{\epsilon/d}\right)^d\right),$$

and for any A with $\sigma_1 I \preceq A \preceq \sigma_2 I$, there exists $B \in \mathcal{M}$ with

$$\exp(-\epsilon) \le \frac{|A|}{|B|} \le \exp(\epsilon) \qquad \text{and} \qquad \|A - B\|_2 \le \epsilon.$$

Proof. The mechanism of the proof is to separately cover the set of orthogonal matrices and the set of possible eigenvalues; this directly leads to the determinant control, and after some algebra, the spectral norm control follows as well.

With foresight, set the scales

$$\begin{aligned}\tau &:= \epsilon/(8\sigma_2), \\ \tau' &:= \epsilon/2, \\ \tau'' &:= \exp(\epsilon/d).\end{aligned}$$

First, a cover of the orthogonal $d \times d$ matrices at scale τ is constructed as follows. The entries of these orthogonal matrices are within $[-1, +1]$, thus first construct a cover Q' of all matrices $[-1, +1]^{d\times d}$ at scale $\tau/2$ according to the maximum-norm, which simply measures the max among entrywise differences; this cover can be constructed by gridding each coordinate at scale $\tau/2$, and thus

$$|Q'| \le (1 + 4/\tau)^{d^2}.$$

Now, to produce a cover of the orthogonal matrices, for each $M' \in Q'$, if it is within max-norm distance $\tau/2$ of some orthogonal matrix M, include M in the new cover Q; otherwise, ignore M'. Since Q' was a max-norm cover of $[-1, +1]^{d\times d}$ at scale $\tau/2$, then Q must be a max-norm cover of the orthogonal matrices at scale τ (by the triangle inequality), and it still holds that

$$|Q| \le (1 + 4/\tau)^{d^2}.$$

Since the max-norm is dominated by the spectral norm, for any orthogonal matrix O, there exists $Q \in M$ with $\|O - Q\|_2 \le \tau$.

Second, a cover of the set of possible eigenvalues is constructed as follows; since both a multiplicative and an additive guarantee are needed for the eigenvalues, two covers will be unioned together. First, produce a cover L_1 of the set $[\sigma_1, \sigma_2]^d$ at scale τ' entrywise as usual, which means $|L_1| \leq (1 + (\sigma_2 - \sigma_1)/\tau')^d$. Second, the cover L_2 will cover each coordinate multiplicatively, meaning each coordinate cover consists of $\sigma_1, \sigma_1\tau'', \sigma_1(\tau'')^2$, and so on; consequently, this cover has size $|L_2| \leq \ln(\sigma_2/\sigma_1)/\ln(\tau'')$. Together, the cover $L := L_1 \cup L_2$ has size

$$|L| \leq \left(1 + \frac{\sigma_2 - \sigma_1}{\tau'}\right)^d + \left(\frac{\ln(\sigma_2/\sigma_1)}{\ln(\tau'')}\right)^d,$$

and for any vector $\Lambda \in [\sigma_1, \sigma_2]^d$, there exists $\Lambda' \in L$ with

$$\frac{1}{\tau''} \leq \max_i \Lambda'_i/\Lambda_i \leq \tau'' \qquad \text{and} \qquad \max_i |\Lambda'_i - \Lambda_i| \leq \tau.$$

Note there was redundancy in this construction: L need only contain nondecreasing sequences.

The final cover $\mathcal{M}$ is thus the cross product of Q and L, and correspondingly its size is the product of their sizes. Given any A with $\sigma_1 I \preceq A \preceq \sigma_2 I$ with spectral decomposition $O_1^\top \Lambda_1 O_1$, pick a corresponding $O_2 \in Q$ which is closest to O_1 in spectral norm, and $\Lambda_2 \in L$ which is closest to Λ_1 in max-norm, and set $B = O_2^\top \Lambda_2 O_2$. By the multiplicative guarantee on L, it follows that

$$\left(\frac{1}{\tau''}\right)^d \leq \frac{|\Lambda_2|}{|\Lambda_1|} = \frac{|B|}{|A|} \leq (\tau'')^d;$$

by the choice of τ'', the determinant guarantee follows. Secondly, relying on a few properties of spectral norms ($\|XY\|_2 \leq \|X\|_2\|Y\|_2$ for square matrices, and $\|Z\|_2 = 1$ for orthogonal matrices, and of course the triangle inequality),

$$\begin{aligned}
\|A - B\|_2 &= \left\|(O_1 - O_2 + O_2)^\top \Lambda_1 (O_1 - O_2 + O_2)^\top - O_2^\top \Lambda_2 O_2\right\|_2 \\
&\leq \|O_2^\top \Lambda_1 O_2 - O_2^\top \Lambda_2 O_2\|_2 + 2\|O_2^\top \Lambda_1 (O_1 - O_2)\|_2 + \|(O_1 - O_2)^\top \Lambda_1 (O_1 - O_2)\|_2 \\
&\leq \|\Lambda_1 - \Lambda_2\|_2 + 2\|O_1 - O_2\|_2\|\Lambda_1\|_2 + \|O_1 - O_2\|_2\|\Lambda_1\|_2(\|O_1\|_2 + \|O_2\|_2) \\
&\leq \tau' + 4\tau\sigma_2,
\end{aligned}$$

and the second guarantee follows by choice of τ and τ'. □

The covering lemma is as follows.

Lemma D.7. *Let scale $\epsilon > 0$, ball $B := \{x \in \mathbb{R}^d : \|x - \mathbb{E}(X)\| \leq R\}$, mean set $X := \{x \in \mathbb{R}^d : \|x - \mathbb{E}(X)\| \leq R_2\}$, covariance eigenvalue bounds $0 < \sigma_1 \leq \sigma_2$, mass lower bound $c_1 > 0$, and number of mixtures $k > 0$ be given. Then there exists a cover set $\mathcal{N}$ (where $(\mu, \Sigma) \in \mathcal{N}$ has $\mu \in X$ and $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$) of size*

$$|\mathcal{N}| \leq \left(\left(\frac{\ln(1/\alpha_0)}{\ln(\tau_0)} + \frac{1-\alpha_0}{\tau_4}\right) \cdot \left(1 + \frac{2R_2 d}{\tau_1}\right)^d \cdot (1 + 32/(\sigma_1\tau_2))^{d^2} \left(\left(1 + \frac{\sigma_1^{-1} - \sigma_2^{-1}}{\tau_2/2}\right)^d + \left(\frac{\ln(\sigma_2/\sigma_1)}{\tau_2/d}\right)^d\right)\right)^k$$

where

$$\begin{aligned}
\tau_0 &:= \exp(\epsilon/4), \\
\tau_1 &:= \min\left\{\frac{\epsilon\sigma_1}{16(R + R_2)}, \sqrt{\frac{\epsilon\sigma_1}{8}}\right\}, \\
\tau_2 &:= \frac{\epsilon}{4\max\{1, (R + R_2)^2\}}, \\
p_{\min} &:= \frac{1}{(2\pi\sigma_2)^{d/2}} \exp(-(R + R_2)^2/(2\sigma_1)), \\
p_{\max} &:= (2\pi\sigma_1)^{-d/2}, \\
\alpha_0 &:= \frac{\epsilon c_1 p_{\min}}{4k(p_{\max} + \epsilon p_{\min}/2)}, \\
\tau_4 &:= \alpha_0,
\end{aligned}$$

(whereby $p_{\min} \le p_\theta(x) \le p_{\max}$ for $x \in B$ and $\theta = (\mu, \Sigma)$ satisfies $\mu \in X$ and $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$,) so that for every partial Gaussian mixture $(\alpha, \Theta) = \{(\alpha_i, \mu_i, \Sigma_i)\}$ with $\alpha_i \ge 0$, $c_1 \le \sum_i \alpha_i \le 1$, $\mu_i \in X$, and $\sigma_1 I \preceq \Sigma_i \preceq \sigma_2 I$ there is an element $(\alpha', \Theta') \in \mathcal{N}$ with weights $c_1 - k\alpha_0 \le \sum_i \alpha'_i \le 1$ so that, for every $x \in B$,

$$|\ln(p_{\alpha,\Theta}(x)) - \ln(p_{\alpha',\Theta'}(x))| \le \epsilon.$$

Proof. The proof controls components in two different ways. For those where the weight α_i is not too small, both α_i and p_{θ_i} are closely (multiplicatively) approximated. When α_i is small, its contribution can be discarded. Between these two cases, the bound follows.

Note briefly that for any $\theta = (\mu, \Sigma)$ with $\mu \in X$ and $\sigma_1 I \preceq \Sigma \preceq \sigma_2 I$,

$$\begin{aligned}
p_\theta(x) &\le \frac{1}{(2\pi\sigma_1)^{d/2}} \exp(\ 0\ ) = p_{\max}, \\
p_\theta(x) &\ge \frac{1}{(2\pi\sigma_2)^{d/2}} \exp(-\|x-\mu\|_2^2/(2\sigma_1)) \\
&\ge \frac{1}{(2\pi\sigma_2)^{d/2}} \exp(-(\|x-\mathbb{E}_\rho(X)\|_2 + \|\mu - \mathbb{E}_\rho(X)\|)^2/(2\sigma_1)) \\
&= p_{\min}.
\end{aligned}$$

Next, the covers of each element of the Gaussian mixture are as follows.

1. Union together a multiplicatively grid of $[\alpha_0, 1]$ at scale τ_0 (meaning produce a sequence of the form $\alpha_0, \alpha_0\tau_0, \alpha_0\tau_0^2$, and so on), and an additive grid of $[\alpha_0, 1]$ at scale τ_4; together, the grid has a size of at most
$$\frac{\ln(1/\alpha_0)}{\ln(\tau_0)} + \frac{1-\alpha_0}{\tau_4}.$$
2. Grid the candidate center set X at scale τ_1, which by Lemma B.5 can be done with size at most
$$\left(1 + \frac{2R_2 d}{\tau_1}\right)^d.$$
3. Lastly, grid the *inverse* of covariance matrices (sometimes called precision matrices), meaning $\sigma_2^{-1} I \preceq \Sigma^{-1} \preceq \sigma_1^{-1}$, whereby Lemma D.6 grants that a cover of size
$$(1 + 32/(\sigma_1\tau_2))^{d^2} \left(\left(\frac{\sigma_1^{-1} - \sigma_2^{-1}}{\tau_2/2}\right)^d + \left(\frac{\ln(\sigma_1/\sigma_2)}{\tau_2/d}\right)^d\right)$$
suffices to provide that for any permissible Σ^{-1}, there exists a cover element A with
$$\exp(-\tau_2) \le \frac{|\Sigma^{-1}|}{|A|} \le \exp(\tau_2) \qquad \text{and} \qquad \|\Sigma^{-1} - A\|_2 \le \tau_2.$$

Producting the size of these various covers and raising to the power k (to handle at most k components), the cover size in the statement is met.

Now consider a component $(\alpha_i, \mu_i, \Sigma_i)$ with $\alpha_i \ge \alpha_0$; a relevant cover element (a_i, c_i, B_i) is chosen as follows.

1. Choose the largest $a_i \le \alpha_i$ in the gridding of $[\alpha_0, 1]$, whereby it follows that $\sum_i a_i \le \sum_i \alpha_i \le 1$, and also
$$\tau_0^{-1} \le a_i/\alpha_i \le \tau_0 \qquad \text{and} \qquad a_i \ge \alpha_i - \tau_4.$$
Thanks to the second property,
$$\sum_{\alpha_i \ge \alpha_0} a_i \ge \left(\sum_{\alpha_i \ge \alpha_0} \alpha_i\right) - k\tau_4.$$

2. Choose c_i in the grid on X so that $\|\mu_i - c_i\| \le \tau_1$.

3. Choose covariance B_i so that

$$\exp(-\tau_2) \le \frac{|B_i|}{|\Sigma_i|} \le \exp(\tau_2) \qquad \text{and} \qquad \|\Sigma^{-1} - B_i^{-1}\|_2 \le \tau_2.$$

The first property directly controls for the determinant term in the Gaussian density. To control the Mahalanobis term, note that the above display, combined with $\|\mu_i - c_i\| \le \tau_1$, gives, for every $x \in B$,

$$\begin{aligned}
&\left|(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i) - (x-c_i)^\top B_i^{-1}(x-c_i)\right| \\
&= \left|(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i) - (x-c_i)^\top (B_i^{-1} - \Sigma_i^{-1} + \Sigma_i^{-1})(x-c_i)\right| \\
&\le \left|(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i) - (x-c_i)^\top \Sigma_i^{-1}(x-c_i)\right| + \|x-c_i\|_2^2 \|B_i^{-1} - \Sigma_i^{-1}\|_2 \\
&\le \left|(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i) - (x-c_i)^\top \Sigma_i^{-1}(x-c_i)\right| + (R+R_2)^2\tau_2 \\
&\le \left|(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i) - (x-c_i)^\top \Sigma_i^{-1}(x-c_i)\right| + \epsilon/4.
\end{aligned}$$

Continuing with the (still uncontrolled) first term,

$$\begin{aligned}
&\left|(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i) - (x-c_i)^\top \Sigma_i^{-1}(x-c_i)\right| \\
&= \left|(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i) - (x-\mu_i+\mu_i-c_i)^\top \Sigma_i^{-1}(x-\mu_i+\mu_i-c_i)\right| \\
&\le 2\|x-\mu_i\|_2\|\mu_i-c_i\|_2\|\Sigma^{-1}\|_2 + \|\mu_i-c_i\|_2^2\|\Sigma^{-1}\|_2 \\
&\le 2(R+R_2)\tau_1/\sigma_1 + \tau_1^2/\sigma_1 \\
&\le \epsilon/4.
\end{aligned}$$

Combining these various controls with the choices of scale parameters, for some provided probability $\alpha_i p_i$ and cover element probability $a_i p_i'$, it follows for $x \in B$ that

$$\exp(-3\epsilon/4) \le \frac{\alpha_i p_i(x)}{a_i p_i'(x)} \le \exp(3\epsilon/4).$$

Lastly, when $\alpha_i < \alpha_0$, simply do not bother to exhibit a cover element.

To show $|\ln(p_{\alpha,\Theta}(x)) - \ln(p_{\alpha',\Theta'}(x))| \le \epsilon$, consider the two directions separately as follows.

1. Given the various constructions above, since $\ln$ is nondecreasing,

$$\begin{aligned}
\ln\left(\sum_i a_i p_{\theta_i'}(x)\right) &\le \ln\left(\sum_{\alpha_i \ge \alpha_0} \alpha_i p_{\theta_i}(x)\exp(3\epsilon/4) + \sum_{\alpha_i < \alpha_0} \alpha_i p_{\theta_i}(x)\right) \\
&\le \ln\left(\sum_i \alpha_i p_{\theta_i}(x)\right) + \frac{3\epsilon}{4}.
\end{aligned}$$

2. On the other hand,

$$\begin{aligned}
\ln\left(\sum_i \alpha_i p_{\theta_i}(x)\right) &= \ln\left(\sum_{\alpha_i \ge \alpha_0} \alpha_i p_{\theta_i}(x) + \sum_{\alpha_i < \alpha_0} \alpha_i p_{\theta_i}(x)\right) \\
&\le \ln\left(\sum_{\alpha_i \ge \alpha_0} a_i p_{\theta_i'}(x)\exp(3\epsilon/4) + k\alpha_0 p_{\max}\right) \\
&= \ln\Bigg((1+\epsilon/4)\sum_{\alpha_i \ge \alpha_0} a_i p_{\theta_i'}(x)\exp(3\epsilon/4) \\
&\qquad - \epsilon/4 \sum_{\alpha_i \ge \alpha_0} a_i p_{\theta_i'}(x)\exp(3\epsilon/4) + k\alpha_0 p_{\max}\Bigg).
\end{aligned}$$

But since $\sum_i a_i \geq c_1 - k(\tau_4 + \alpha_0)$,

$$-\epsilon/4 \sum_{\alpha_i \geq \alpha_0} a_i p_{\theta_i'}(x) \exp(3\epsilon/4) \leq -(\epsilon/4)(c_1 - k(\tau_4 + \alpha_0))p_{\min} + k\alpha_0 p_{\max} \leq 0.$$

As such, since $(1 + \epsilon/4) \leq \exp(\epsilon/4)$, the result follows in this case as well.

□

D.3 Proof of Theorem 5.1

Proof of Theorem 5.1. This proof is based on the proof of Theorem 3.2. Let the various quantities in Lemma D.5 be given; in particular, let balls B, C and their radii R_B, R_C be as provided there. Additionally, define $p_0 := \exp(4c)/8e$ for convenience. Near the end of the proof, the choices $p' = p/4$ and $\epsilon := m^{-1/2+1/p}$ will be made.

By Lemma D.7, let $\mathcal{N}$ be a cover of the set C, with all parameters having the same names as those here, except the R there is the radius R_B here, and R_2 there is radius R_C here, the lower bound c_1 is $p_0/p_{\max}$. By the construction of the cover there, every set of partial Gaussian parameters $(\alpha, \Theta) \in C$ with $\sum_{\alpha_i} \alpha_i \geq c_1 = p_0/p_{\max}$ and cardinality at most k has a cover element $Q \in \mathcal{N}$ with

$$\sup_{x \in B} |\phi_{\mathrm{g}}(x; (\alpha, \Theta)) - \phi_{\mathrm{g}}(x; Q)| \leq \epsilon; \tag{D.8}$$

note that Lemma D.7 also provides the stated estimate of the size. Next, note for $x \in B$ and every cover element $Q \in \mathcal{N}$ that Lemma D.7 provides

$$\ln((c_1 - k\alpha_0)p_{\min}) \leq p_Q(x) \leq \ln(p_{\max})$$

where $c_1 = p_0/p_{\max}$ as above and

$$\alpha_0 = \frac{\epsilon c_1 p_{\min}}{4k(p_{\max} + \epsilon p_{\min}/2)} \leq \frac{\epsilon c_1 p_{\min}}{4kp_{\max}},$$

which combined with $\epsilon \leq 2$ and $p_{\min} \leq p_{\max}$ gives

$$c_1 - k\alpha_0 \geq c_1 \left(1 - \frac{\epsilon p_{\min}}{4p_{\max}}\right) \geq \frac{c_1}{2}.$$

Thus, by Hoeffding's inequality,

$$\begin{aligned}\sup_{Q \in \mathcal{N}} \left| \int_B \phi_{\mathrm{g}}(x; Q) d\hat{\rho}(x) - \int_B \phi_{\mathrm{g}}(x; Q) d\rho(x) \right| &\leq \ln\left(\frac{p_{\max}}{p_{\min}(c_1 - k\alpha_0)}\right) \sqrt{\frac{1}{2m} \ln\left(\frac{2|\mathcal{N}|}{\delta}\right)} \\ &\leq \ln\left(\frac{2p_{\max}^2}{p_{\min} p_0}\right) \sqrt{\frac{1}{2m} \ln\left(\frac{2|\mathcal{N}|}{\delta}\right)}.\end{aligned} \tag{D.9}$$

For the remainder of this proof, discard the corresponding failure event.

To further simplify eq. (D.9), note firstly that

$$\begin{aligned}\ln\left(\frac{1}{p_{\min}}\right) &= \ln\left((2\pi\sigma_2)^{d/2} \exp((R_B + R_C)^2/(2\sigma_1))\right) \\ &= \ln((2\pi\sigma_2)^{d/2}) + 2R_C^2/\sigma_1,\end{aligned}$$

where

$$R_C^2 \leq 3R_B^2(1 + \sqrt{8\sigma_2/\sigma_1})^2 + 12\sigma_2 \ln(1/\epsilon) + 6\sigma_2 \left(\ln\left(\frac{64e^2(2\pi\sigma_2)^d}{(2\pi)^d p_{\max}^4}\right) - 8c\right)$$

and

$$R_B^2 \leq 2R_6^2 + M_1^2/\epsilon^{2/(p-2p')}.$$

Next, to control $|\mathcal{N}|$, the scale term $\tau = \min\{\tau_1, \tau_2\}$ must first be controlled. Since $\epsilon \le 1$ and $\sigma_1 \le 1$ and $R_C \ge 1$,

$$\tau \ge \frac{\epsilon\sigma_1}{16(R_B + R_C)^2} \ge \frac{\epsilon\sigma_1}{64R_C^2},$$

and thus

$$\ln\left(\frac{\epsilon}{\tau}\right) \le \ln(64R_C^2/\sigma_1).$$

Together with $\tau_0 = \exp(\epsilon/4)$ and $\alpha_0 \ge \epsilon c_1 p_{\min}/(8kp_{\max}) = p_0 p_{\min}/(8kp_{\max}^2)$, and letting $\mathcal{O}(\cdot)$ swallow terms depending only on numerical constants, c, σ_1, and σ_2, but in particular not touching terms depending on ϵ, d, k or m or δ,

$$\begin{aligned}
\ln(|\mathcal{N}|) &\le \ln\left(\left(\left(\frac{5}{\epsilon}\left(\frac{8kp_{\max}^2}{\epsilon p_0 p_{\min}}\right)\right)\left(\frac{3R_C d}{\tau}\right)^d\left(\frac{33}{\sigma_1\tau}\right)^{d^2}\left(\left(\frac{\sigma_1^{-1}}{\tau/2}\right)^d + \left(\frac{\ln(\sigma_2/\sigma_1)}{\tau/2}\right)^d\right)\right)^k\right) \\
&\le 3d^2k\left(5\ln(1/\epsilon) + \ln(1/p_{\min}) + 3\ln(\epsilon/\tau) + \ln(R_C) + \mathcal{O}(1)\right) \\
&\le 3d^2k\left(5\ln(1/\epsilon) + 2R_C^2/\sigma_1 + 3\ln(\epsilon/\tau) + 4\ln(R_C) + \mathcal{O}(1)\right) \\
&= \mathcal{O}\left(d^2k(\ln(1/\epsilon) + \epsilon^{-2/(p-2p')}\right).
\end{aligned}$$

Together, the full expression in eq. (D.9) may be simplified down to

$$\begin{aligned}
&\sup_{Q\in\mathcal{N}}\left|\int_B \phi_{\mathrm{g}}(x;Q)d\hat{\rho}(x) - \int_B \phi_{\mathrm{g}}(x;Q)d\rho(x)\right| \\
&\quad\le \mathcal{O}\left(\mathrm{poly}(d,k)\left(\frac{1}{\epsilon}\right)^{2/(p-2p')}\sqrt{\frac{(\ln(1/\epsilon) + (1/\epsilon)^{2/(p-2p')} + \ln(1/\delta))}{m}}\right) \\
&\quad\le \mathcal{O}\left(\mathrm{poly}(d,k)\left(\frac{\epsilon^{-3/(p-2p')}}{\sqrt{m}} + \sqrt{\frac{(\ln(1/\epsilon) + \ln(1/\delta))}{m}}\right)\right) \\
&\quad\le \mathcal{O}\left(\mathrm{poly}(d,k)\left(m^{-1/2+3/p} + \sqrt{\frac{(\ln(m) + \ln(1/\delta))}{m}}\right)\right)
\end{aligned} \tag{D.10}$$

where the final step used the choice $p' = p/4$ and $\epsilon := m^{-1/2+1/p}$.

Now let any $(\alpha,\Theta) \in \mathcal{S}_{\mathrm{mog}}(\rho; c, k, \sigma_1, \sigma_2) \cup \mathcal{S}_{\mathrm{mog}}(\hat{\rho}; c, k, \sigma_1, \sigma_2)$ be given, and let $Q \in \mathcal{N}$ be a cover element satisfying eq. (D.8) for $(\alpha,\theta) \sqcap C$. By eq. (D.8), eq. (D.9), and Lemma D.5 (and thus discarding an additional failure event having probability 4δ),

$$\begin{aligned}
\left|\int \phi_{\mathrm{g}}(x;(\alpha,\Theta))d\rho(x) - \int \phi_{\mathrm{g}}(x;(\alpha,\Theta))d\hat{\rho}(x)\right| &\le \left|\int \phi_{\mathrm{g}}(x;(\alpha,\Theta))d\rho(x) - \int_B \phi_{\mathrm{g}}(x;(\alpha,\Theta)\sqcap C)d\rho(x)\right| \\
&\quad + \left|\int_B \phi_{\mathrm{g}}(x;(\alpha,\Theta)\sqcap C)d\rho(x) - \int_B \phi_{\mathrm{g}}(x;Q)d\rho(x)\right| \\
&\quad + \left|\int_B \phi_{\mathrm{g}}(x;Q)d\rho(x) - \int_B \phi_{\mathrm{g}}(x;Q)d\hat{\rho}(x)\right| \\
&\quad + \left|\int_B \phi_{\mathrm{g}}(x;Q)d\hat{\rho}(x) - \int_B \phi_{\mathrm{g}}(x;(\alpha,\Theta)\sqcap C)d\hat{\rho}(x)\right| \\
&\quad + \left|\int_B \phi_{\mathrm{g}}(x;(\alpha,\Theta)\sqcap C)d\hat{\rho}(x) - \int \phi_{\mathrm{g}}(x;(\alpha,\Theta))d\hat{\rho}(x)\right| \\
&\le 4\epsilon + \ln\left(\frac{2p_{\max}^2}{p_{\min}p_0}\right)\sqrt{\frac{1}{2m}\ln\left(\frac{2|\mathcal{N}|}{\delta}\right)} + \epsilon_\rho + \epsilon_{\hat{\rho}}, \\
&= \text{poly(d,k)}\mathcal{O}(m^{-1/2+3/p}\left(1 + \sqrt{\ln(m) + \ln(1/\delta)} + (1/\delta)^{4/p}\right),
\end{aligned}$$

where the final step uses the above simplification of the cover term, the choices $\epsilon = m^{-1/2+1/p}$ and $p' = p/4$, and additionally unwrapping the forms of ϵ_ρ and $\epsilon_{\hat{\rho}}$ from Lemma D.5. □